Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Nonlinear estimation for linear inverse problems with error in the operator

Marc Hoffmann
*Laboratoire d'Analyse et
de Mathématiques Appliquées
CNRS UMR 8050 et
Université de Marne-la-Vallée*
hoffmann@univ-mlv.fr

Markus Reiß
*Weierstraß Institute for
Applied Analysis and Stochastics
WIAS Berlin*
mreiss@wias-berlin.de

submitted: December 9, 2004

No. 990
Berlin 2004

2000 *Mathematics Subject Classification.* 65J20, 62G07.

Key words and phrases. statistical inverse problem, Galerkin projection method, wavelet thresholding, minimax rate, degree of ill-posedness, matrix compression.

The project has been supported in part by the European network Dynstoch.

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

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

We consider nonlinear estimation methods for statistical inverse problems in the case where the operator is not exactly known. For a canonical formulation a Gaussian operator white noise framework is developed. Two different nonlinear estimators are constructed, which correspond to the different order of the linear inversion and nonlinear smoothing step. We show that both estimators are rate-optimal over a wide range of Besov smoothness classes. The construction is based on the Galerkin projection method and wavelet thresholding schemes for the data and the operator.

1 Introduction

1.1 Linear inverse problems and ill-posedness

We consider the usual statistical formulation of a linear inverse problem: given a domain $\mathcal{D} \subset \mathbb{R}^d$, an unknown object of interest $f \in L^2(\mathcal{D})$ is to be recovered from

$$g_\varepsilon = Kf + \varepsilon \dot{W}, \tag{1.1}$$

where g_ε is the data, K is a known linear operator

$$K : L^2(\mathcal{D}) \to L^2(\mathcal{Q}),$$

$\mathcal{Q}$ is a domain in $\mathbb{R}^q$ and $\dot{W}$ is a Gaussian white noise on $L^2(\mathcal{Q})$. We assess the quality of recovery of an estimator $\hat{f}$ by

$$\mathcal{R}(\hat{f}, f) := \mathbb{E}\left[\|\hat{f} - f\|^2_{L^2(\mathcal{D})}\right], \tag{1.2}$$

$\mathbb{E}[\bullet]$ denoting expectation. In most interesting cases K^{-1} does not exist as a bounded linear operator and the crude estimate $\hat{f}_\varepsilon := K^{-1}(g_\varepsilon)$ is not feasible without further regularisation: the estimation problem (1.1) is ill-posed. For a review of the concept in statistics and numerics see e.g. Nußbaum and Pereverzev [27], Whaba [32], Engl *et al.* [17].

Among the most popular regularization methods, we mention the *singular value decomposition* or SVD (Johnstone and Silverman [23], Mair and Ruymgaart [26], and *projection methods* (Dicken and Maass [13], Mathé and Perverzev [25]) together with their nonlinear counterparts (Cavalier and Tsybakov [5], Cavalier *et al.* [7], Tsybakov [31], Goldenshluger and Pereverzev [20], Efromovich and Kolchinskii [16]), including wavelet approaches (Donoho [14], Abramovich and Silverman [1], Cohen *et. al.* [10], Johnstone *et al.* [22], Antoniadis and Bigot [2]). The main difficulty for the statistician is whether the chosen representation should be optimally adapted to K (by using the eigenfunctions in the SVD) or rather to f and g_ε (by using projection on approximation spaces). Within a chosen framework, classical smoothing and regularisation techniques can then be applied, *e.g.* penalisation (Tychonov regularisation, ridge regression) or filtering (Pinsker or block constant weights, series truncation).

This paper addresses the problem (1.1) when, in addition, the acting operator K is not known exactly. In this context, we do not have access to the exact singular value decomposition of K. Moreover, we want to provide a spatially adaptive estimator $\hat{f}$ of f. To circumvent both difficulties, we propose the use of projection methods based on *nonlinear* wavelet decompositions. The scientific interest of this generalization ranges from technical applications over numerical discretisations to statistical inference, more in Section 2 below.

1.2 Linear inverse problems with error in the operator

The statistical model. We do not have access to K exactly, but rather to

$$K_\delta = K + \delta \dot{B}. \tag{1.3}$$

The process K_δ is a blurred version of K, polluted by a Gaussian operator white noise, at level $\delta > 0$. There are basically two ways of interpreting (1.3):

- The operator K acting on f is unknown and treated as a nuisance parameter. However, preliminary statistical inference about K is possible, with an accuracy governed by δ. Thus (1.3) is viewed as a limiting experiment for K and can be incorporated to our dataset. More in the examples of Section 2.
- For experimental reasons (numerical errors or systematic deficiency of measuring devices) we do not have access to K exactly, but rather to K_δ. In this context, the error level δ can be linked to the accuracy of supplementary experiments or training data. See Efromovich and Kolchinskii [16] for an elaboration of this approach and the examples in Section 2.

Finally, our statistical model is given by the observation of (1.1) together with the supplementary data (1.3): we observe $(g_\varepsilon, K_\delta)$ with

$$g_\varepsilon = Kf + \varepsilon \dot{W}, \quad K_\delta = K + \delta \dot{B}. \tag{1.4}$$

Asymptotics are taken as δ, $\varepsilon \to 0$ simultaneously. The interplay between δ and ε is crucial to understand Model (1.4). At first glance, if $\delta \ll \varepsilon$ one expects to approximately recover Model (1.1). On the other hand, we will see that recovering f is a completely different issue if $\varepsilon \ll \delta$. Even when the error in the signal dominates, the fact that $\delta \neq 0$ has to be handled carefully.

Gaussian operator white noise. In rigorous probabilistic terms, observable quantities take the form

$$\langle g_\varepsilon, k\rangle := \langle Kf, k\rangle_{L^2(\mathcal{Q})} + \varepsilon \langle \dot{W}, k\rangle \ \ \forall\, k \in L^2(\mathcal{Q})$$

and

$$\langle K_\delta h, k\rangle := \langle Kh, k\rangle_{L^2(\mathcal{Q})} + \delta \langle \dot{B}h, k\rangle, \ \ \forall\, (h,k) \in L^2(\mathcal{D}) \times L^2(\mathcal{Q}).$$

The mapping $k \in L^2(\mathcal{Q}) \mapsto \langle \dot{W}, k\rangle$ defines a centred Gaussian linear form, with covariance

$$\mathbb{E}[\langle \dot{W}, k_1\rangle\langle \dot{W}, k_2\rangle] = \langle k_1, k_2\rangle_{L^2(\mathcal{Q})}, \quad k_1,\, k_2 \in L^2(\mathcal{Q}).$$

Likewise, $(h, k) \in L^2(\mathcal{D}) \times L^2(\mathcal{Q}) \mapsto \langle \dot{B}h, k\rangle$ defines a centred Gaussian bilinear form with covariance

$$\mathbb{E}[\langle \dot{B}h_1, k_1\rangle\langle \dot{B}h_2, k_2\rangle] = \langle h_1, h_2\rangle_{L^2(\mathcal{D})}\langle k_1, k_2\rangle_{L^2(\mathcal{Q})}.$$

In particular, if $(h_i)_{i\geqslant 1}$ and $(k_i)_{i\geqslant 1}$ form orthonormal bases of $L^2(\mathcal{D})$ and $L^2(\mathcal{Q})$, respectively, the infinite vector $(\langle \dot{W}, k_j\rangle)_{j\geqslant 1}$ and the infinite matrix $(\langle \dot{B}h_i, k_j\rangle)_{i,j\geq 1}$ have i.i.d. standard Gaussian entries. Alternatively, the law of $\dot{B}$ is centred Gaussian and characterised by its reproducing kernel Hilbert space $HS(L^2(\mathcal{D}), L^2(\mathcal{Q}))$, the space of Hilbert-Schmidt operators. Yet another description of the operator white noise is given by stochastic integration using a Brownian sheet $\tilde{B}$ on $\mathcal{D} \times \mathcal{Q}$ as kernel:

$$\langle \dot{B}h, k\rangle = \int_{\mathcal{Q}}\int_{\mathcal{D}} h(x)k(y)\, d\tilde{B}(x,y),$$

which gives a natural interpretation as white noise perturbation of the operator kernel.

1.3 Main results

We restrict our attention to selfadjoint operators K on $L^2(\mathcal{D})$, denote by d the dimension of $\mathcal{D}$ and loosely speak about a degree t of ill-posedness of the operator K, see Section 3.2 for details.

As a starting point we consider in Section 4 the linear Galerkin projection method without taking care of the error in the operator. For functions in the L^2-Sobolev space H^s and suitable approximation spaces, the linear estimator converges with the minimax rate $\max\{\delta, \varepsilon\}^{2s/(2s+2t+d)}$. The standard nonparametric rate for ill-posed problems is obtained, with an accuracy dominated by the largest of the two noise levels ε and δ.

The linear Galerkin method performs regularisation and inversion at the same time. For a spatially adaptive procedure, we have to separate the two steps of *Galerkin inversion* and *adaptive regularisation* or smoothing. On a rough methodological level we can adopt one of the following two strategies:

$$\text{Method I:} \qquad (g_\varepsilon,\, K_\delta) \xrightarrow{\text{inversion}} \hat{f}_{\delta,\varepsilon}^{\text{linear}} \xrightarrow{\text{smoothing}} \text{estimator } \hat{f}_{\delta,\varepsilon}^{I},$$

$$\text{Method II:} \qquad (g_\varepsilon,\, K_\delta) \xrightarrow{\text{smoothing}} (\hat{g}_\varepsilon,\, \hat{K}_\delta) \xrightarrow{\text{inversion}} \text{estimator } \hat{f}_{\delta,\varepsilon}^{II}.$$

In this context, $\hat{f}_{\delta,\varepsilon}^{\text{linear}}$ is considered as a preliminary undersmoothed estimator. We investigate Method I and Method II, with the Galerkin scheme on a high-dimensional space as inversion procedure and wavelet thresholding as adaptive smoothing technique. Technical details are a level-dependent thresholding rule in method I and a noise reduction in the operator by entrywise thresholding of the wavelet matrix representation in method II. To our knowledge, thresholding for the operator has not yet been studied in the statistical literature, while advocated in numerical analysis (*a-posteriori compression*, see Dahmen *et al.* [11]), and may be thus of independent interest.

As it turns out, the inversion step is critical in both methods and we cannot choose an arbitrarily large approximation space for the inversion, even in method II. Nevertheless, both methods are provably rate-optimal (up to a log factor in some cases) over a wide range of sparse nonparametric classes, expressed in terms of Besov spaces $B^s_{p,p}$ with $p \leqslant 2$. In fact, the rate $\max\{\delta, \varepsilon\}^{2s/(2s+2t+d)}$ is essentially extended to the Besov spaces $B^s_{p,p}$ for all p with $\frac{1}{p} \leq \frac{1}{2} + \frac{s}{2t+d}$. Both methods can be classified as reasonable general purpose procedures for linear inverse problems with errors in the operator. In the case of small regularity, though, there exist certain limitations for both methods, which at a closer look give some hints which method is preferable depending on the sparsity of the data.

In the next Section 2 related approaches and examples are discussed. After introducing precise model assumptions in Section 3 the construction of the linear estimator and the two nonlinear estimators $\hat{f}^{II}_{\delta,\varepsilon}$ and $\hat{f}^{I}_{\delta,\varepsilon}$ together with their asymptotic properties are presented in Sections 4 to 7. In Section 8 we further discuss and compare both nonlinear methods. The proofs of the main theorems are deferred to Sections 9 and 10. The appendix contains short proofs for the linear method and the lower bound, as well as some essential tools from approximation theory.

2 Related approaches with error in the operator

Perturbed singular values. In the context of the SVD, Cavalier and Hengartner [6] assume that the singular functions of K are known, but that its singular values are perturbed by noise. Typical examples include convolution operators. By an oracle-inequality approach, they show how to reconstruct f efficiently when $\delta \leqslant \varepsilon$. A similar problem is encountered as *blind deconvolution* in image analysis, which shares the main ingredients of our model except that the knowledge about the convolution kernel is reduced to a strong localisation (e.g. Pruessner and O'Leary [28]).

Physical devices. In many physical situations we are given an integral equation $Kf = g$ on a closed boundary surface Γ of a domain $\Omega \subset \mathbb{R}^{d+1}$, where the boundary integral operator

$$Kh(x) = \int_\Gamma k(x,y)h(y)\sigma_\Gamma(dy)$$

is of order $t > 0$, that is $K : H^{-t/2}(\Gamma) \to H^{t/2}(\Gamma)$ is given by a smooth kernel $k(x,y)$ as a function of x and y off the diagonal, but that is typically singular on the diagonal. Such kernels arise, for instance, by applying a boundary integral formulation to second-order elliptic problems, e.g. in potential theory. Typical examples include Abel-type operators, with $k(x,y) = b(x,y)/(x-y)^\beta$ for some $\beta > 0$, with $\Gamma = [0,1]$, and b at least Lipschitz-continuous and bounded below (see e.g. Dahmen *et al.* [11], Mathé and Pereverzev [25]). Assuming that k is tractable only up to some experimental error, due to unknown exact parameter values or variability in the device itself, we postulate the knowledge of $k_\delta = k + \delta\dot{\Lambda}$, where $\dot{\Lambda}$ is Gaussian white noise on $L^2(\Gamma \times \Gamma)$. Assuming moreover that our data g is perturbed by measurement noise as in (1.1), we recover the framework of (1.4) with $\mathcal{D} = \mathcal{Q} = \Gamma$.

Statistical inference. Let us first focus on the widespread structural model of instrumental variables, e.g. Florens [18], Hall and Horowitz [21]. We observe i.i.d. (X_i, Y_i, W_i) for $i = 1, \dots, n$, where (X_i, Y_i) follow a regression model

$$Y_i = g(X_i) + U_i$$

with the exception that $\mathbb{E}[U_i \,|\, X_i] \neq 0$, but under the additional information given by the instrumental variables W_i that satisfy $\mathbb{E}[U_i \,|\, W_i] = 0$. Denoting by f_{XW} the joint density of X and W, we define

$$k(x,z) := \int f_{XW}(x,w) f_{XW}(z,w)\, dw, \quad Kh(z) := \int k(x,z)h(z)\, dz.$$

To draw inference on g, we use the identity

$$Kg(x) = \mathbb{E}[\mathbb{E}[Y \,|\, W] f_{XW}(x, W)].$$

The data easily allows estimation of the right-hand side and of the kernel function k. We face exactly an ill-posed inverse problem with errors in the operator as in model (1.4), except for certain correlations between the two noise sources and for the fact that the noise is caused by a density estimation problem. Note that K has a symmetric non-negative kernel and is therefore self-adjoint and non-negative on L^2. Under technical conditions, Hall and Horowitz [21] obtain in their Theorem 4.2 the linear rate of Section 4 when replacing their terms as follows: $\varepsilon = \delta = n^{-1/2}$, $t = \alpha$, $s = \beta + 1/2$, $d = 1$.

In other statistical problems random matrices or operators are of key importance or even the main subject of interest, for instance the linear response function in functional data analysis (Ramsay and Silverman [29]), the Markov transition operator in discretely observed diffusion processes (Gobet *et al.* [19]) or the empirical covariance operator for stochastic processes. A typical instance of the latter is given by Reiß [30] who considers the problem of estimating the weight function f in stochastic delay differential equations of the form

$$dX_t = \Big(\int_{-r}^{0} X_{t+u}\, f(u)\, du \Big) dt + dW_t, \quad t \in [0, T].$$

Maximum-likelihood theory suggests an estimator $\hat{f}_T$ which is a solution of the equation $Q_T \hat{f}_T = b_T$, where

$$Q_T h(v) = \int_{-r}^{0} \frac{1}{T} \int_{0}^{T} X_{t+u} X_{t+v}\, dt\, h(u)\, du, \; b_T(v) = \frac{1}{T} \int_0^T X_{t+v}\, dX_t.$$

Under stationarity the empirical covariance operator Q_T can be shown to approximate the true covariance operator Q like $Q + \delta Q \dot{B}$ and the noise in the data is coloured in the sense that $b_T \cong Qf + \varepsilon Q^{1/2}\dot{W}$ with noise levels $\delta = \varepsilon = T^{-1/2}$. Along a scale of Besov spaces the covariance operator Q is always ill-posed of degree 2. Decorrelating the noise in the data by applying formally $Q^{-1/2}$, we obtain the model (1.4) with $K = Q^{1/2}$, $K_\delta = K + \delta K\dot{B}$ and $g_\varepsilon = Q^{1/2}f + \varepsilon\dot{W}$. Here as well as in functional data analysis applications the coloured noise in the covariance operator makes life generally easier, compare also the good bounds for the principal component analysis (or SVD) derived by estimates of the Hilbert-Schmidt norm instead of the tighter operator norm in Cai and Hall [4].

Numerical discretisation. Even if the operator is precisely known in theory, it must be implemented and hence discretised. For the projection methods this means usually that integrals over products of the kernel with the basis functions have to be calculated. The numerical analyst is thus confronted with the same question of error in the operator under a different angle: up to which accuracy should the operator be discretised? Even more importantly, by not using all available information on the operator the objects typically have a sparse data structure and thus require far less memory capacity and the algorithms are much faster, see Dahmen *et al.* [11].

The deterministic nature of the noise in the operator leads to essentially different estimates because the analysis cannot profit from the very strong concentration properties of random matrices, cf. Lemma 9.1 below. When the entries of the Galerkin stiffness matrix are calculated independently and without a systematic bias, a stochastic error modelisation like in model (1.4) seems appropriate and yields a large reduction in the theoretical error bounds.

3 Model assumptions

We write $a \lesssim b$ when $a \leq cb$ for some constant $c > 0$, i.e. $a = \mathcal{O}(b)$ in the Landau notation, and $a \sim b$ when $a \lesssim b$ and $b \lesssim a$ simultaneously. The uniformity in c will be obvious from the context.

3.1 Smoothness and sparsity of the signal

The function spaces we consider are defined on domains $\mathcal{D} \subset \mathbb{R}^d$, appended with boundary conditions. We measure the smoothness of f in L^p-norm with $1 \leq p \leq 2$, cf. Appendix 11.3 for the notion of Besov and Sobolev spaces. For $p = 2$ we introduce the nonparametric class of L^2-Sobolev balls of regularity $\alpha \geq 0$ and radius M:

$$W^\alpha(M) := \{f \in H^\alpha \ ; \ \|f\|_{H^\alpha} \leq M\}.$$

For $1 \leq p < 2$ we model spatial inhomogeneity and introduce the Besov balls of regularity (s,p) and radius M:

$$V_p^s(M) := \{f \in B^s_{p,p} \ ; \ \|f\|_{B^s_{p,p}} \leq M\}, \ \ s > 0.$$

The restriction on (s,p) is given in (3.2) and discussed in detail in Sections 7 and 8.

Finally, we use regular orthogonal wavelet bases $(\psi_\lambda)_\lambda$ adapted to the domain $\mathcal{D}$ that provide unconditional bases for $B^s_{p,p}$. The multi-index $\lambda = (j,k)$ concatenates the spatial index k and the resolution level $j = |\lambda|$. Thus, for $f \in L^2$ we have

$$f = \sum_{j \geq -1} \sum_{|\lambda|=j} f_\lambda \psi_\lambda, \quad f_\lambda := \langle f, \psi_\lambda \rangle,$$

where we use the level $j = -1$ to incorporate the low frequency part of the decomposition.

3.2 Smoothness and sparsity of the operator

The ill-posedness comes from the fact that K^{-1} is not L^2-continuous. We quantify this by a smoothing action with *ill-posedness* degree $t > 0$. If $\mathcal{D} = \mathcal{Q}$ and K is selfadjoint, this smoothing action becomes an ellipticity property:

Assumption 3.1. *$\mathcal{D} = \mathcal{Q}$ and K is selfadjoint. Moreover*

$$\langle Kf, f \rangle \sim \|f\|^2_{H^{-t/2}}. \tag{3.1}$$

This means that $K^{1/2}$ is well defined and maps $H^{-t/2}$ to L^2 isomorphically. By duality, this implies that $K^{1/2} : L^2 \to H^{t/2}$ is isomorphic, and this extends to $K : H^{-t/2} \to H^{t/2}$ isomorphically.

Remark 3.2. *The restriction that K is selfadjoint can be removed by transferring Assumption 3.1 from K to K^*K, with K^* denoting the adjoint of K. Likewise for Assumptions 3.3 and 3.4 below. The subsequently used Galerkin method then becomes the usual least squares method, see Cohen et al. [10].*

The choice of the loss function (1.2) in assessing the error and the range of smoothness for f require further mapping properties. Let us first introduce the following restriction on (s,p), considering t and d as fixed by the problem (more in Sections 7 and 8):

$$\frac{1}{p} \leq \frac{1}{2} + \frac{s}{2t+d}. \tag{3.2}$$

Assumption 3.3. *The parameters $s \geqslant 0$ and $p \geqslant 1$ satisfy* (3.2). *Moreover, $K : B^s_{p,p} \to B^{s+t}_{p,p}$ isomorphically.*

Finally, we state a hypothesis on the sparsity of K expressed in its wavelet discretisation, specified by parameters $(\bar{s}, \bar{p})$ related to K. Here we allow for values $p < 1$ since only approximation properties (and not the Banach space structure) are needed.

Assumption 3.4. *The parameters $\bar{s} \geq 0$, and $\bar{p} \geq 0$ satisfy* (3.2). *Moreover, uniformly over all multi-indices λ we have*

$$\|K\psi_\lambda\|_{B^{\bar{s}+t}_{\bar{p},\bar{p}}} \lesssim 2^{|\lambda|(\bar{s}+d/2-d/\bar{p})}$$

for the specified wavelet basis.

Remark 3.5. *Assumption 3.4 is implied by Assumption 3.3 for $\bar{p} \geqslant 1$ with $s = \bar{s}$ and $p = \bar{p}$ due to $\|\psi_\lambda\|_{B^s_{p,p}} \sim 2^{|\lambda|(s+d/2-d/p)}$.*

The case $\bar{p} < 1$ expresses high sparsity for K. For instance, if K is diagonal in the wavelet basis with eigenvalues of order $2^{-|\lambda|t}$, then Assumption 3.4 holds for all $\bar{s}$, $\bar{p} \geqslant 0$.

4 A preliminary linear estimation method

We briefly study the linear Galerkin projection estimator, as a particular case of the approach in Efromovich and Kolchinskii [16]. For $j \geq 0$ we search for $\hat{f}_{\delta,\varepsilon} \in V_j$ such that

$$\langle K_\delta \hat{f}_{\delta,\varepsilon}, v \rangle = \langle g_\varepsilon, v \rangle \quad \forall v \in V_j, \tag{4.1}$$

where $V_j = \operatorname{span}\{\psi_\lambda,\ |\lambda| \leq j\}$ is the multiresolution space associated with the wavelet basis. This makes perfect sense as soon as K_δ restricted to V_j is invertible, but, although this is true for K, the fact that $\delta \neq 0$ requires extra care.

We introduce the Galerkin projection (*stiffness matrix*) of an operator T onto V_j by setting $T_j := P_j T|_{V_j}$, where P_j is the orthogonal projection onto V_j. The existence of a unique solution to (4.1) is equivalent to the invertibility of $K_{\delta,j}$. Choose some $\tau > 0$ and set

$$\hat{f}_{\delta,\varepsilon} := \begin{cases} K_{\delta,j}^{-1} P_j\, g_\varepsilon, & \text{if } \|K_{\delta,j}^{-1}\|_{V_j \to V_j} \leqslant \tau 2^{jt}, \\ \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

Definition 4.1. *Let us introduce the rate exponent*

$$r(s,t,d) := \frac{2s}{2s+2t+d}$$

and the mapping constants

$$c_K := \sup_{j \geqslant 0} 2^{-jt} \|K_j^{-1}\|_{V_j \to V_j}, \quad c'_K := \sup_{j \geqslant 0,\, h \in V_j, \|h\|_{H^t}=1} \|K_j^{-1} h\|_{L^2}.$$

Proposition 4.2. *Let $s > 0$, $M > 0$ and a $(\lfloor s \rfloor + 1)$-regular multiresolution analysis (V_j) be given. Grant Assumptions 3.1 and 3.3 for K with parameters $(s,p) = (0,2)$. Then the following asymptotic bound holds:*

$$\sup_{f \in W^s(M)} \mathcal{R}(\hat{f}_{\delta,\varepsilon}, f) \lesssim \max\{\delta, \varepsilon\}^{2r(s,t,d)},$$

as soon as the estimator $\hat{f}_{\delta,\varepsilon}$ is specified by $2^j \sim \max\{\delta,\varepsilon\}^{-2/(2s+2t+d)}$ and $\tau > c_K$.

Remark 4.3. *Assumption 3.1 ensures that c_K is finite, see Lemma 11.1 below.*

The normalised rate $\max\{\delta,\varepsilon\}^{r(s,t,d)}$ gives the explicit interplay between ε and δ and is indeed optimal under some restriction on K, cf. Section 7. Proposition 4.2 is essentially contained in [16] but is proved in Appendix 11.1 as central reference for the proposed nonlinear methods.

5 Two nonlinear estimation methods

5.1 Nonlinear estimation I

For $x > 0$ define the level-dependent hard thresholding rule $\mathcal{S}_x$ for $h \in L^2$ by

$$h \mapsto \mathcal{S}_x(h) := \sum_\lambda \langle h, \psi_\lambda \rangle \mathbf{1}_{\{|\langle h, \psi_\lambda \rangle| \geqslant 2^{|\lambda| t} \mathcal{T}(x)\}} \psi_\lambda, \tag{5.1}$$

where the threshold is defined by $\mathcal{T}(x) := \kappa x \sqrt{|\log x|}$ for $\kappa > 0$. Our first nonlinear estimator is defined by

$$\hat{f}^I_{\delta,\varepsilon} := \mathcal{S}_{\max\{\delta,\varepsilon\}}(\hat{f}_{\delta,\varepsilon}), \tag{5.2}$$

where $\hat{f}_{\delta,\varepsilon}$ is the linear estimator (4.2) specified by the level $J = J(\delta, \varepsilon)$ such that

$$2^J \sim \min\{\varepsilon^{-1/t}, \delta^{-1/(t+d)}\} \tag{5.3}$$

and $2^J \leqslant c_J \delta^{-1/(t+d)}$ for some small constant $c_J > 0$. Thus, $\hat{f}^I_{\delta,\varepsilon}$ is specified by c_J, τ and κ.

Remark 5.1.

(a) In practice, the computation of $\hat{f}_{\delta,\varepsilon}$ and thus $\hat{f}^I_{\delta,\varepsilon}$ is heavy since the data is inverted on a large space V_J.

(b) Since we assume K to be selfadjoint, we can reduce the error in the observation K_δ by considering the symmetrisation $\frac{1}{2}(K_\delta + K_\delta^*)$.

(c) Concerning the tuning parameters: the non-asymptotic choice of the resolution level J should be such that the smallest eigenvalue of the matrix $K_{\delta,J}$ is larger than $\delta 2^{Jd}$, i.e. the noise level δ multiplied by the matrix dimension. A proper choice of κ could be estimated from the data, but is already difficult in theory, cf. Abramovich and Silverman [1]. The last tuning parameter τ exists only for theoretical reasons and will not be enforced unless large deviations occur.

5.2 Nonlinear estimation II

Our second method is conceptually different: we use matrix compression techniques to remove the operator noise by thresholding K_δ in a first step and then apply the Galerkin inversion on the smoothed data g_ε. From a computational point of view, this approach is more efficient than the first one since the linear system we need to solve will be sparse and fast iterative solvers can be used (Dahmen *et al.* [11]). Also theoretically, thresholding K_δ enables us to take advantage of the possible sparsity of K in the wavelet basis. Let

$$\hat{K}_\delta := \mathcal{S}^J_\delta(K_{\delta,J}), \tag{5.4}$$

where $K_{\delta,J} = P_J K_\delta|_{V_J}$ is the Galerkin projection and $\mathcal{S}^J_\delta$ is a hard-thresholding rule applied to the entries in the wavelet representation:

$$T_J \mapsto \mathcal{S}^J_\delta(T_J) := \sum_{|\lambda|,|\lambda'|\leqslant J} T_{\lambda,\lambda'}\mathbf{1}_{\{|T_{\lambda,\lambda'}|\geqslant \mathcal{T}(\delta)\}}\langle \bullet, \psi_\lambda\rangle\psi_{\lambda'}, \tag{5.5}$$

$\mathcal{T}(\delta)$ is defined in Section 5.1 and the $T_{\lambda,\lambda'} := \langle T\psi_\lambda, \psi'_\lambda\rangle$ are the entries of the matrix of the operator T represented in the wavelet basis.

The estimator $\hat{g}_\varepsilon$ of the data is obtained by the classical hard-thresholding rule for noisy signals:

$$\hat{g}_\varepsilon := \sum_{|\lambda|\leqslant J} \langle g_\varepsilon, \psi_\lambda\rangle \mathbf{1}_{\{|\langle g_\varepsilon,\psi_\lambda\rangle|\geqslant \mathcal{T}(\varepsilon)\}}\psi_\lambda. \tag{5.6}$$

After this preliminary step, we invert the linear system to obtain our second nonlinear estimator:

$$\hat{f}^{II}_{\delta,\varepsilon} := \begin{cases} \hat{K}_\delta^{-1}\hat{g}_\varepsilon, & \|\hat{K}_\delta^{-1}\|_{V_J\to V_J} \leqslant \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{5.7}$$

Here $\tau > 0$ is a large cut-off value. This time, we take $J = J(\delta, \varepsilon)$ such that

$$2^J \sim \min\left\{\varepsilon^{-1/t}, \left(\delta\sqrt{|\log\delta|}\right)^{-1/(t+d)}\right\} \tag{5.8}$$

and $2^J \leqslant c_J\left(\delta\sqrt{|\log\delta|}\right)^{-1/(t+d)}$ for a small constant $c_J > 0$. We choose J a little bit smaller than in the previous method, in order to guarantee with overwhelming probability the invertibility of

$$\hat{K}_\delta : (V_J, \|\bullet\|_{L^2}) \to (V_J, \|\bullet\|_{H^t}),$$

see Lemma 10.3. Thus $\hat{f}^{II}_{\delta,\varepsilon}$ is specified by c_J, τ and κ.

Remark 5.2.

(a) Observe that this time we do not use level-dependent thresholds since we perform the thresholding before the inversion step such that the noise level is the same for all coefficients.

(b) The choice of the tuning parameters for this procedure should follow the same theoretical guidelines as that for method I. Let us stress that in practice, contrary to the first method $\hat{f}^{I}_{\delta,\varepsilon}$, an efficient numerical scheme to construct $\hat{f}^{II}_{\delta,\varepsilon}$, based on an iterative inversion method with thresholding in each step, could be used, as developed by Cohen et al. [10].

6 Results

In the following we fix $s_+ > 0$ and pick a wavelet basis $(\psi_\lambda)_\lambda$ associated to a $(\lfloor s_+\rfloor + 1)$-regular multiresolution analysis (V_j). We need to specify a restriction on the linear approximation error expressed in terms of the regularity in H^α:

$$\alpha \geqslant s\Big(\frac{t+d}{s+t+d/2}\Big)\min\left\{\frac{\log\varepsilon}{\log\delta}, 1\right\} \quad \text{in the case } \delta > \varepsilon^{1+d/t}. \tag{6.1}$$

6.1 Nonlinear estimation I

Theorem 6.1. *Let $0 \leqslant s \leqslant s_+$, $p \geqslant 1$ satisfy* (3.2) *and α satisfy* (6.1). *Grant for K Assumption 3.1 and Assumption 3.3 in both parameters (s,p) and $(0,2)$. Then for sufficiently small c_J:*

$$\sup_{f \in V_p^s(M) \bigcap W^\alpha(M)} \mathcal{R}(f^I_{\delta,\varepsilon}, f) \lesssim \max\left\{\delta\sqrt{|\log\delta|}, \varepsilon\sqrt{|\log\varepsilon|}\right\}^{2r(s,t,d)},$$

provided $f^I_{\delta,\varepsilon}$ is specified by J in (5.3), *$\tau > c_K$ and $\kappa \geq c(c'_K, M) > 0$. The constant $c(c'_K, M)$ is continuous, increasing in its arguments, and explicitly computable from Lemmas 9.5 and 9.6 below.*

Remark 6.2. *Assumption 3.3 with $(s,p) = (0,2)$ ensures that c'_K is finite by Lemma 11.1.*

For $\delta \leqslant \varepsilon^{1+d/t}$ there is no restriction on α and we obtain the optimal rate for the estimation in the class $V_p^s(M)$ up to logarithmic terms, cf. Section 7. If $\delta > \varepsilon^{1+d/t}$, we can get rid of the linear restriction (6.1) by Sobolev embeddings (Appendix 11.3) when excluding sparse functions f, that is, simultaneously small values of p and s:

Corollary 6.3. *Let $\delta > \varepsilon^{1+d/t}$. In the setting of Theorem 6.1, $f^I_{\delta,\varepsilon}$ attains the near-optimal rate over the scale $V_p^s(M)$*

$$\sup_{f \in V_p^s(M)} \mathcal{R}(\hat{f}^I_{\delta,\varepsilon}, f) \lesssim \max\left\{\delta\sqrt{|\log\delta|}, \varepsilon\sqrt{|\log\varepsilon|}\right\}^{4s/(2s+2t+d)}$$

under the additional restriction

$$\left\{\begin{array}{lcl} s & \geqslant & \frac{d}{2} - (t+d)\left(1 - \frac{\log\varepsilon}{\log\delta}\right)^+, \\ \frac{1}{p} & \leqslant & \frac{1}{2} + \frac{s}{d}\frac{s-d/2+(t+d)\left(1-\frac{\log\varepsilon}{\log\delta}\right)^+}{s+t+d/2}. \end{array}\right. \tag{6.2}$$

If $s \geqslant d + d^2/2t$, the additional conditions are automatically fulfilled for all $p \geqslant 1$ obeying (3.2).

6.2 Nonlinear estimation II

Also the second nonlinear method attains the optimal rate of convergence under certain parameter restrictions. For given s and p we further impose

$$\frac{2\bar{s} + d - 2d/\bar{p}}{2\bar{s} + 2t + d} \leqslant \frac{2s - d}{2t + d} \quad \text{with strict inequality for } p > 1, \tag{6.3}$$

where $(\bar{s}, \bar{p})$ are the sparsity coefficients of K from Assumption 3.4. In Corollary 6.5 below, we obtain an upper bound for $\hat{f}^{II}_{\delta,\varepsilon}$ in analogy to Theorem 6.1.

We first state a general result which gives separate estimates for the two error levels of $\hat{f}^{II}_{\delta,\varepsilon}$ associated with δ and ε, respectively, and that leads to faster rates of convergence than in Theorem 6.1 in the case of sparse operator discretisations, cf. Section 8.

Theorem 6.4. *Let $0 \leqslant s, \bar{s} \leqslant s_+$, $p \geqslant 1$, $\bar{p} \geqslant 0$ be such that (s,p) satisfy (3.2) and $(\bar{s},\bar{p})$ satisfy (3.2) and (6.3). Let $\alpha \geqslant 0$ satisfy (6.1). If K fulfills Assumption 3.3 for both parameters (s,p) and $(0,2)$ and Assumption 3.4 with sparsity parameters $(\bar{s},\bar{p})$, then for all $M \geqslant 0$ and sufficiently small c_J*

$$\sup_{f \in V_p^s(M) \bigcap W^\alpha(M)} \mathcal{R}(\hat{f}_{\delta,\varepsilon}^{II}, f) \lesssim \left(\varepsilon\sqrt{|\log \varepsilon|}\right)^{2r(s,t,d)} + \left(\delta\sqrt{|\log \delta|}\right)^{2r(\bar{s},t,d)},$$

provided $\hat{f}_{\delta,\varepsilon}^{II}$ is specified by J in (5.8), $\tau > c_K'$ and $\kappa > 0$.

Corollary 6.5. *In the setting of Theorem 6.4 the following asymptotic risk bound holds:*

$$\sup_{f \in V_p^s(M) \bigcap W^\alpha(M)} \mathcal{R}(\hat{f}_{\delta,\varepsilon}^{II}, f) \lesssim \max\left\{\varepsilon\sqrt{|\log \varepsilon|}, \delta\sqrt{|\log \delta|}\right\}^{2r(s,t,d)}$$

under the additional restriction

$$s \geqslant \tfrac{1}{4}\left(d^2 + 8(2t+d)(d - d/p)\right)^{1/2}. \tag{6.4}$$

Proof. Set $\bar{s} = s$ and $\bar{p} = p$ and use that Assumption 3.3 implies Assumption 3.4. Then (6.4) implies restriction (6.3) and Theorem 6.4 gives the result. □

Remark 6.6. *The restriction (6.4) is automatically satisfied for $p = 1$ or for $s > \frac{d}{2}(1 + \frac{2t}{d})^{1/2}$ since s and p always satisfy restriction (3.2). Whenever (6.2) is fulfilled the linear restriction $f \in W^\alpha(M)$ can be avoided using Sobolev embeddings as in Corollary 6.3.*

7 Lower bounds

The lower bound in the case $\delta = 0$ is classical (Nussbaum and Pereverzev [27]). The lower bounds will not decrease for increasing noise levels δ or ε, whence it suffices to provide lower bounds for the case $\varepsilon = 0$. In the remainder of this section we consider the model given by the observation of $Y = Kf$ and $K_\delta = K + \delta\dot{B}$.

7.1 The dense case $p = 2$

The case $p = 2$ is addressed in Efromovich and Kolchinskii [16]. The following result can be derived from their study:

$$\inf_{\hat{f}_\delta} \sup_{(f,K) \in \mathcal{F}_{s,2,t}} \mathcal{R}(\hat{f}_\delta, f) \gtrsim \delta^{2r(s,t,d)},$$

where the nonparametric class $\mathcal{F}_{s,2,t} = \mathcal{F}_{s,2,t}(M,C)$ takes the form $\mathcal{F}_{s,2,t} = W^s(M) \times \mathcal{K}_t(C)$, where $\mathcal{K}_t(C)$ is the class of operators satisfying Assumption 3.1 with $c_K \leq C$ for some $C > 0$.

Remark 7.1. *The same lower bound applies for $\mathcal{F}_{s,p,t} := V_p^s(M) \times \mathcal{K}_t(C)$ for $s > 0$ and $p \in [1, \infty]$: it is universal over p and matches the upper bound attained by $\hat{f}_{\delta,\varepsilon}^I$ and $\hat{f}_{\delta,\varepsilon}^{II}$ for $\frac{1}{p} \leq \frac{1}{2} + \frac{s}{2t+d}$ up to a logarithmic factor.*

7.2 The sparse case $p < 2$

In the sparse case the logarithmic factor is necessary. In Appendix 11.2 we prove the following lower bound:

Theorem 7.2. *In the setting described above, the following asymptotic lower bound holds:*

$$\inf_{\hat{f}_\delta} \sup_{(K,f)\in\mathcal{F}_{s,p,t}} \mathcal{R}(\hat{f}_\delta, f) \gtrsim \left(\delta\sqrt{|\log\delta|}\right)^{\left(s+\frac{d}{2}-\frac{d}{p}\right)/\left(s+t+\frac{d}{2}-\frac{d}{p}\right)}. \tag{7.1}$$

In particular, this sparse lower bound matches the dense rate with exponent $r(s,t,d)$ for those values of (s,p) that satisfy

$$(\tfrac{d}{2}-\tfrac{d}{p})(s+t+\tfrac{d}{2}) = -s\tfrac{d}{p} \iff \tfrac{1}{p} = \tfrac{1}{2} + \tfrac{s}{2t+d}. \tag{7.2}$$

The equivalence (7.2) corresponds to the critical case of (3.2). For values of p smaller than the critical value we cannot attain the dense rate. In that case the upper bound matches the lower bound by embedding into $B^{\sigma}_{\pi,\pi}$, $\sigma < s$, with $\sigma - d\pi^{-1} = s - dp^{-1}$ and (σ, π) satisfying restriction (3.2) with equality.

8 Discussion

8.1 Concerning nonlinear estimation I

The first nonlinear method produces an estimator that attains optimal rates in the case of sparse data, as measured by Besov spaces $B^s_{p,p}$ with $p \in [1,2)$. In the case of negligible error in the operator $\delta \ll \varepsilon$ this result complements the procedures discussed in Cohen *et al.* [10] by showing that from an asymptotic viewpoint the order of thresholding and inversion in the estimation procedure does not matter.

Let us now focus on the error in the operator and assume for simplicity $\varepsilon = 0$. Then the only theoretical drawback is the linear restriction (6.1) in terms of $W^\alpha(M)$ for small values of s and p. Otherwise the bias would deteriorate, for only the choice $2^J \sim \delta^{-1/t}$ avoids this extra assumption. The reason for the smaller choice (5.3) of J is that the inverse of the Galerkin matrix $K_{\delta,J}$ cannot be controlled for larger values. That we should not choose $2^J \sim \delta^{-1/t}$ is intuitively clear because

$$K_{\delta,J} = K_J + \delta\dot{B}_J$$

is the sum of a positive-definite operator with smallest eigenvalues of order 2^{-Jt} and a random operator of norm $\delta 2^{Jd/2}$, cf. Lemmas 11.1 and 9.1, and $K_{\delta,J}$ is not likely to be invertible for $2^{J(t+d/2)} \gtrsim \delta$. Our even smaller choice is caused by the nonlinearity of the noise in the inverse $K_{\delta,J}^{-1}$. Looking at the toy example $K_J = 2^{-Jt}\,\mathrm{Id}$ and symmetrising $K_{\delta,J}$ and thus $\dot{B}_J$ (keeping the same notation), we know by spectral calculus and Lemma 9.1 (cf. also the Wigner law, e.g. in Davidson and Szarek [12]) that $K_{\delta,J}^{-1}K_J$ is symmetric with a normalised trace that satisfies with overwhelming probability

$$\begin{aligned}2^{-Jd}\mathrm{tr}(K_{\delta,J}^{-1}K_J) &\sim \frac{1}{2^{Jd/2+1}\delta}\int_{-2^{Jd/2}\delta}^{2^{Jd/2}\delta}\frac{2^{-Jt}}{2^{-Jt}+x}\,dx\\ &= \frac{\log(1+2^{J(t+d/2)}\delta)-\log(1-2^{J(t+d/2)}\delta)}{2^{J(t+d/2)+1}\delta}\\ &= 1+\tfrac{1}{3}2^{J(2t+d)}\delta^2+\mathcal{O}(2^{3J(t+d/2)}\delta^3).\end{aligned}$$

Hence, invariance in law of $\dot{B}_J$ under orthogonal transformations shows that all diagonal elements of $(K_{\delta,J}^{-1}-K_J^{-1})K_J$ in the orthonormal basis $(\psi_\lambda)_\lambda$ are also at least of order $2^{J(2t+d)}\delta^2$, which is only below the error level $2^{Jt}\delta$ of its linearisation $K_J^{-1}\delta\dot{B}_J$ if $2^{J(t+d)}\delta \lesssim 1$. We thus see that our technical result in Lemma 9.5, which has given rise to the choice (5.3) of J, is due to the inherent nonlinearity of the inversion $K_{\delta,J}\mapsto K_{\delta,J}^{-1}$.

8.2 Concerning nonlinear estimation II

In the case of a known operator K (i.e. when we set $\delta=0$ in (1.4)) the second nonlinear estimator $\hat{f}_{\delta,\varepsilon}^{II}$ is only a combination of signal denoising and operator compression using wavelet bases. The thresholding of the operator is very natural and leads to a significant gain in the speed of inversion using iterative solvers. This idea is used as a-posteriori compression scheme in Dahmen *et al.* [11], which discusses numerical issues in detail and provides mathematical results for more specific integral operators with a finger-like wavelet representation.

Let us turn to the case of significant errors in the operator and set $\varepsilon=0$ for simplicity. To overcome the problems of inverting $K_{\delta,J}$ in a large approximation space V_J, as observed in the first nonlinear method, it is plausible that we should first reduce the stochastic error and only then apply the inversion procedure. For the purpose of variance reduction, thresholding the operator seems a natural choice, as advocated in numerical analysis and successfully applied for variance reduction in signal detection.

Unfortunately, the method in general will not reduce the error in the operator so much that we can choose J as large as needed to render additional linear approximation conditions unnecessary. The reason for this lack of error reduction is a *stability* problem. We need that the estimated operator $\hat{K}_\delta$ is an L^2-H^t-isomorphism with uniformly bounded norm constants for $\delta\to 0$. This property is needed in order to guarantee the inversion estimate

$$\|(\hat{K}_\delta^{-1}-K_J^{-1})P_Jg\|_{L^2}=\|\hat{K}_\delta^{-1}(K_\delta-K_J)f_J\|_{L^2}\lesssim\|(K_\delta-K_J)f_J\|_{H^t}.$$

Let us consider for some level J an operator K with entries

$$\langle K\psi_\lambda,\psi_\lambda\rangle=2^{-|\lambda|t},\ \langle K\psi_\mu,\psi_\lambda\rangle=\sigma_{\mu,\lambda}2^{-J(t+d/2)}\text{ for }|\mu|=|\lambda|=J,\ \lambda\neq\mu,$$

and all other entries equal to zero, where $\sigma_{\mu,\lambda}=\sigma_{\lambda,\mu}\in\{-\eta c,c\}$ for some $\eta\in(0,1)$ and $c>0$ are chosen in such a way that $\|(\sigma_{\mu,\lambda})_{|\mu|=|\lambda|=J}\|_{l^2\to l^2}<2^{Jd/2}$ holds, a typical result when choosing the constant c small and independently $\sigma_{\lambda,\mu}=c$

with probability $\eta/(1+\eta)$ and $\sigma_{\lambda,\mu} = -\eta c$ with probability $1/(1+\eta)$, cf. Bennet *et al.* [3]. Then K satisfies Assumptions 3.1 and 3.3 for all $(s, 2)$. Let us suppose $\delta = 2^{-J(t+d/2)} c(1+\eta)/2$ and consider as reference the oracle estimate $\hat{K}_\delta^{or}$ which sets to zero all entries of K_δ for which the modulus of the corresponding entry of K is smaller than δ. Then $\hat{K}_\delta^{or}$ is obtained from K_δ by setting all entries with $\sigma_{\mu,\lambda} = -\eta c$ to zero. Since still approximately $2^{Jd}\eta/(1+\eta)$ entries are of order $2^{-J(t+d/2)}$ and of the same sign, we get with operator norms taken on V_J

$$\|\hat{K}_\delta^{or} - K_J\|_{L^2 \to H^t} \gtrsim \|(\sigma_{\mu,\lambda} 2^{-Jd/2} \mathbf{1}_{\{\sigma_{\mu,\lambda}=c\}})\|_{l^2 \to l^2} \sim 2^{Jd/2},$$

which explodes for $J \to \infty$. This means that even for $2^J \sim \delta^{-1/(t+d/2)}$ wavelet thresholding does not reduce the error sufficiently to guarantee a stable inversion of the Galerkin matrix. A more detailed analysis shows that this is only accomplished with the choice (5.8) of J, cf. Lemma 10.3. The fundamental reason for this property is that component-wise thresholding does not perform a shrinkage in operator norm such that operators with a large discrepancy between their operator norm and their Hilbert-Schmidt norm, which equals the Euclidean norm when represented in an orthonormal basis, behave badly under thresholding.

A similar example of an operator, which in the case $p = 2$ is a diagonal operator up to a random perturbation for $|\mu| = J(s - d/2)$ and $|\lambda| = J$, also shows that the additional restriction on s in (6.4) is very likely due to the method used and not the product of a suboptimal proof. Note that this restriction is void for $p = 1$, since we have then $s \geqslant t + d/2$ anyway due to the restriction (3.2) on p. When the estimation methods are judged by their minimax-type upper bounds along the scale of Besov spaces $B^s_{p,p}$ with $p \geqslant 1$, the second nonlinear estimation method has a worse behaviour than the first one, at least for all $p \approx 2$.

Typical operators, however, often do not have a huge number of small entries of the same size in their wavelet representation, but display rather a finger-like structure with about $J2^{Jd}$ entries of considerable size, cf. Dahmen *et al.* [11]. Hence, we can profit from this highly sparse structure and attain even faster rates, compare Theorem 6.4. For an illustration let us consider the extreme case with an operator K diagonalised in the chosen wavelet basis with eigenvalues of order $2^{-|\lambda|t}$. Then the estimator attains the rate $\delta^{2\bar{s}/(2\bar{s}+2t+d)}$ for all $\bar{s} \geqslant 0$ and $1/\bar{p} = 1/2 + \bar{s}/(2t+d)$ satisfying (6.3), that is

$$\tfrac{2\bar{s}+d-d(1+2\bar{s}/(2t+d))}{2\bar{s}+2t+d} \leqslant \tfrac{2s-d}{2t+d} \iff \tfrac{2\bar{s}}{2\bar{s}+2t+d} \leqslant \tfrac{2s-d}{2t}.$$

Hence, we obtain up to logarithmic factors the rate $\max\{\delta, \delta^{(2s-d)/(2t)}\}$, which is barely parametric for not too small s and can be shown to be optimal in a minimax sense. For such highly sparse operator representations the estimator $\hat{f}^{II}_{\delta,\varepsilon}$ significantly outperforms $\hat{f}_{\delta,\varepsilon}$ and $\hat{f}^{I}_{\delta,\varepsilon}$ without any special tuning.

8.3 Conclusion

We have proposed two nonlinear estimation methods for inverse problems with errors in the operator which outperform linear methods when the data are sparse and spatially inhomogeneous. From an algorithmic point of view the second method is preferable. In theory, both methods are provably rate-optimal (up to a logarithmic factor in some case) over a wide range of smoothness classes. They show, however, different behaviour when the error in the operator dominates and the degree of smoothness is small. In essence, the first method deals better with densely populated matrices and signals whereas the second method is better suited in the case of sparse structures.

The main ideas of the wavelet approach are transferable to other nonlinear smoothing methods. For instance, when the kernel of the operator is the subject of observation, e.g. in the case of instrumental variables, adaptive kernel methods with local bandwidth choice could be used. The analogue of the first nonlinear method would consist in smoothing the kernel adaptively, discretising it in a second step and then inverting the linear system. Alternatively, for method II an undersmoothed kernel estimate for the inversion and an adaptive kernel smoother after the inversion could be used. We believe that the results and limitations obtained for the wavelet methods will carry over to this approach. In particular, estimating the kernel function optimally in L^2- or H^t-loss corresponds to estimating the operator in a loss of Hilbert-Schmidt type. Due to dimensionality effects this estimation usually differs strongly from the required operator norm loss.

9 Proofs for nonlinear estimation I

9.1 Preparations

The following result is a classical, yet intriguing bound for Gaussian random matrices, see [12, Thm. II.4].

Lemma 9.1. *For universal constants β_0, c, $C > 0$ and all $\beta \geqslant \beta_0$, $\alpha \geqslant 0$, $j \in \mathbb{N}$ we have*

$$\begin{aligned}\mathbb{P}(2^{-jd/2}\|\dot{B}_j\|_{V_j\to V_j} \geqslant \beta) &\leqslant \exp(-c\beta^2 2^{2jd}),\\ \mathbb{P}(2^{-jd/2}\|\dot{B}_j\|_{V_j\to V_j} \leqslant \alpha) &\leqslant (C\alpha)^{2^{2jd}},\end{aligned}$$

where $\mathbb{P}(\bullet)$ stands for probability. In particular, for all $\gamma \geqslant 1$

$$\mathbb{E}[\|\dot{B}_j\|_{V_j\to V_j}^{\gamma}]^{1/\gamma} \lesssim 2^{jd/2}.$$

Lemma 9.2. *Under Assumption 3.1 for K we obtain for the Galerkin inversion $f_J := K_J^{-1}P_JKf$ uniformly over $J \in \mathbb{N}$, $\alpha \geqslant 0$ and $M > 0$ the estimate*

$$\sup_{f\in W^\alpha(M)} \|f - f_J\|_{L^2} \lesssim M2^{-J\alpha}.$$

Proof. See the argument in [10], Section 3.1, especially Eq. (3.11). □

Lemma 9.3. *Given the choice of J in* (5.3) *and the restrictions* (3.2), (6.1) *for $s \geqslant 0$, $p \geqslant 1$ and $\alpha \geqslant 0$, we have*

$$\sup_{f\in V_p^s(M)\cap W^\alpha(M)} \|f - f_J\|_{L^2} \lesssim \max\{\delta, \varepsilon\}^{2s/(2s+2t+d)}.$$

Proof. For $\delta \leqslant \varepsilon^{1+d/t}$ we use the embeddings from Appendix 11.3 which under restriction (3.2) yields $B^s_{p,p} \subset H^{st/(t+d/2)}$, and the result follows from Lemma 9.2 and $2^J \sim \varepsilon^{-1/t}$ in (5.3). If $\delta > \varepsilon^{1+d/t}$ holds, then (6.1) has been built exactly such that Lemma 9.2 gives the result. □

Lemma 9.4. *Grant Assumption 3.3 for* $(s,p) = (0,2)$ *and let* J *be chosen according to* (5.3). *Then there is a constant* $c_\Omega > 0$ *such that*

$$\mathbb{P}(\Omega^c_{\rho,\delta,J}) \leqslant \exp(-c_\Omega \rho \delta^{-d/(2t+2d)} 2^{2Jd}) \quad \forall\, \rho > 0,\, \delta > 0,\, J \in \mathbb{N}.$$

The proof of Lemma 9.4 is obtained along the same lines as Lemma 11.2 in the Appendix. On the event $\Omega_{\rho,\delta,J}$ the random operator $K_{\delta,J}$ is invertible with

$$K^{-1}_{\delta,J} = \Big(\operatorname{Id} - \delta K_J^{-1} \dot{B}_J + \sum_{n \geqslant 2} (-\delta K_J^{-1} \dot{B}_J)^n \Big) K_J^{-1}$$

by the usual Neumann series argument such that

$$\hat{f}_{\delta,\varepsilon} = K^{-1}_{\delta,J} P_j g_\varepsilon = K^{-1}_{\delta,J} P_j K f + \varepsilon K^{-1}_{\delta,J} P_J \dot{W}.$$

On $\Omega_{\rho,\delta,J}$ we thus obtain the decomposition

$$\hat{f}_{\delta,\varepsilon} = f_J - \delta K_J^{-1} \dot{B}_J f_J + \varepsilon K_J^{-1} P_J \dot{W} + + r^{(1)}_{\delta,J} + r^{(2)}_{\delta,\varepsilon,J}, \tag{9.1}$$

with

$$r^{(1)}_{\delta,J} := \sum_{n \geqslant 2} (-\delta K_J^{-1} \dot{B}_J)^n f_J, \tag{9.2}$$

$$r^{(2)}_{\delta,\varepsilon,J} := -\varepsilon \delta K_J^{-1} \dot{B}_J \sum_{n \geqslant 2} (-\delta K_J^{-1} \dot{B}_J)^n K_J^{-1} P_J \dot{W}. \tag{9.3}$$

Lemma 9.5. *Let* $|\lambda| \leqslant J$ *and let* $\rho \in (0, 1 - c_K/\tau)$. *Under Assumption 3.1 the following decomposition holds:*

$$\begin{aligned}
\delta \langle K_J^{-1} \dot{B}_J f_J, \psi_\lambda \rangle &= \delta 2^{|\lambda| t} \|f_J\|_{L^2}\, c_\lambda\, \xi_\lambda, \\
\varepsilon \langle K_J^{-1} P_J \dot{W}, \psi_\lambda \rangle &= \varepsilon 2^{|\lambda| t} \tilde{c}_\lambda\, \tilde{\xi}_\lambda, \\
\langle r^{(1)}_{\delta,J}, \psi_\lambda \rangle &= \delta^2\, 2^{|\lambda| t}\, \|f_J\|_{L^2}\, 2^{J(t+d)}\, \zeta_{\lambda,J}, \\
\langle r^{(2)}_{\delta,\varepsilon,J}, \psi_\lambda \rangle &= \delta \varepsilon\, 2^{|\lambda| t}\, 2^{J(t+d/2)}\, \tilde{\zeta}_{\lambda,J} \quad \textit{on} \quad \Omega_{\rho,\delta,J},
\end{aligned}$$

where $|c_\lambda|, |\tilde{c}_\lambda| \lesssim 1$, ξ_λ *and* $\tilde{\xi}_\lambda$ *are standard Gaussian variables and* $\zeta_{\lambda,J}$, $\tilde{\zeta}_{\lambda,J}$ *are random variables satisfying*

$$\max\{\mathbb{P}(\{|\zeta_{\lambda,J}| \geqslant \beta\} \cap \Omega_{\rho,\delta,J}), \mathbb{P}(\{|\zeta_{\lambda,J}| \geqslant \beta\} \cap \Omega_{\rho,\delta,J})\} \leqslant \exp(-c\beta 2^{2Jd})$$

for all $\beta \geqslant \beta_0$ *with some constants* β_0, $c > 0$.

Proof. By Assumption 3.1, K_J is symmetric and thus $\delta \langle K_J^{-1} \dot{B}_J f_J, \psi_\lambda \rangle = \delta \langle \dot{B}_J f_J, K_J^{-1} \psi_\lambda \rangle$. By definition of $\dot{B}$, the last quantity is a centred Gaussian random variable with variance $\delta^2 \|f_J\|^2_{L^2} \|K_J^{-1} \psi_\lambda\|^2_{L^2}$. Moreover $\|K_J^{-1} \psi_\lambda\|^2_{L^2} \lesssim$

$\|\psi_\lambda\|_{H^t}^2 \lesssim 2^{2|\lambda|t}$ by the mapping property (11.1), and the first equality follows from Lemma 9.2.

For the second equality we write $\varepsilon\langle K_J^{-1}P_J\dot W,\psi_\lambda\rangle = \varepsilon\langle \dot W, K_J^{-1}\psi_\lambda\rangle$, which is centred Gaussian with variance $\varepsilon^2\|K_J^{-1}\psi_\lambda\|_{L^2}^2$, and the foregoing arguments apply.

Concerning the third decomposition, on $\Omega_{\rho,\delta,J}$ the term $|\langle r_{\delta,J}^{(1)},\psi_\lambda\rangle|$ equals

$$\begin{aligned}
&|\langle(\delta K_J^{-1}\dot B_J)^2(\mathrm{Id}+\delta K_J^{-1}\dot B_J)^{-1}f_J,\psi_\lambda\rangle|\\
=&\delta^2|\langle \dot B_J K_J^{-1}\dot B_J(\mathrm{Id}+\delta K_J^{-1}\dot B_J)^{-1}f_J, K_J^{-1}\psi_\lambda\rangle|\\
\leqslant&\delta^2\|\dot B_J\|_{V_J\to V_J}^2\|K_J^{-1}\|_{V_J\to V_J}\|(\mathrm{Id}+\delta K_J^{-1}\dot B_J)^{-1}\|_{V_J\to V_J}\|f_J\|_{L^2}\|K_J^{-1}\psi_\lambda\|_{L^2}\\
\lesssim&\delta^2\|\dot B_J\|_{V_J\to V_J}^2 2^{Jt}2^{|\lambda|t},
\end{aligned}$$

where we successively applied the Cauchy-Schwarz inequality, Lemma 11.1, estimate (11.2) on $\Omega_{\rho,\delta,J}$ and Lemma 9.2 together with the same arguments as before to bound $\|K_J^{-1}\psi_\lambda\|_{L^2}$. Lemma 9.1 yields the result.

Finally, since $\dot W$ and $\dot B$ are independent, we have that, conditional on $\dot B$, the random variable $\langle r_{\delta,\varepsilon,J}^{(2)},\psi_\lambda\rangle\mathbf{1}_{\Omega_{\rho,\delta,J}}$ is centred Gaussian with conditional variance

$$\begin{aligned}
&\delta^2\varepsilon^2\|(K_J^{-1}\dot B_J(\mathrm{Id}+\delta K_j^{-1}\dot B_j)^{-1}K_J^{-1})^*\psi_\lambda\|_{L^2}^2\mathbf{1}_{\Omega_{\rho,\delta,J}}\\
&=\delta^2\varepsilon^2\|(\dot B_J(\mathrm{Id}+\delta K_j^{-1}\dot B_j)^{-1}K_J^{-1})^*K_J^{-1}\psi_\lambda\|_{L^2}^2\mathbf{1}_{\Omega_{\rho,\delta,J}}\\
&\lesssim\delta^2\varepsilon^2\|(\dot B_J(\mathrm{Id}+\delta K_j^{-1}\dot B_j)^{-1}K_J^{-1})^*\|_{V_J\to V_J}^2 2^{2|\lambda|t}\mathbf{1}_{\Omega_{\rho,\delta,J}}\\
&\lesssim\delta^2\varepsilon^2 2^{2(|\lambda|+J)t}\|\dot B_J^*\|_{V_J\to V_J}^2
\end{aligned}$$

by Lemma 11.1 and estimate (11.2), which is not affected when passing to the adjoint, up to an appropriate modification of $\Omega_{\rho,\delta,J}$ incorporating B_J^*. We conclude by applying Lemma 9.1 which is also not affected when passing to the adjoint. □

Lemma 9.6. *Let $|\lambda|\leqslant J$ and $\gamma>0$. For sufficiently large κ, depending on c_K' and γ, we have the uniform estimate*

$$\mathbb{P}\left(\{|\langle\hat f_{\delta,\varepsilon},\psi_\lambda\rangle-\langle f_J,\psi_\lambda\rangle|\geqslant\tfrac14 2^{|\lambda|t}\mathcal{T}(\max\{\delta,\varepsilon\})\}\cap\Omega_{\rho,\delta,J}\right)\lesssim\max\{\delta,\varepsilon\}^\gamma.$$

Proof. By decomposition (9.1) and Lemma 9.5, the above probability is bounded by the sum of four terms $I+II+III+IV$ with

$$\begin{aligned}
I&:=\mathbb{P}\left(\delta\|f_J\|_{L^2}c_\lambda|\xi_\lambda|\geqslant\tfrac{1}{16}\mathcal{T}(\max\{\delta,\varepsilon\})\right),\\
II&:=\mathbb{P}\left(\varepsilon|\tilde c_\lambda\tilde\xi_\lambda|\geqslant\tfrac{1}{16}\mathcal{T}(\max\{\delta,\varepsilon\})\right),\\
III&:=\mathbb{P}\left(\{\delta^2 2^{J(t+d)}\|f_J\|_{L^2}\zeta_{\lambda,J}\geqslant\tfrac{1}{16}\mathcal{T}(\max\{\delta,\varepsilon\})\}\cap\Omega_{\rho,\delta,J}\right),\\
IV&:=\mathbb{P}\left(\{\delta\varepsilon 2^{J(t+d/2)}\tilde\zeta_{\lambda,J}\geqslant\tfrac{1}{16}\mathcal{T}(\max\{\delta,\varepsilon\})\}\cap\Omega_{\rho,\delta,J}\right).
\end{aligned}$$

We bound I thanks to the standard estimate $\mathbb{P}(|\xi_\lambda|\geqslant x)\leqslant\exp(-x^2/2)$, $x\geqslant0$, and thus obtain by straightforward calculation

$$I\leqslant\max\{\delta,\varepsilon\}^{\kappa/c_I}\lesssim\max\{\delta,\varepsilon\}^\gamma$$

with some constant $c_I>0$ and for $\kappa\geqslant\gamma c_I$; likewise, for some $c_{II}>0$ and $\kappa\geqslant\gamma c_{II}$

$$II\leqslant\max\{\delta,\varepsilon\}^{\kappa/c_{II}}\lesssim\max\{\delta,\varepsilon\}^\gamma.$$

We bound III by the large deviation estimate, using the definition (5.3) of c_J for all sufficiently large J:

$$III \leqslant \exp\big(- cc_J^{-1} \|f_J\|_{L^2}^{-1} \tfrac{1}{16} \kappa |\log(\max\{\delta, \varepsilon\})|^{1/2} 2^{2Jd} \big).$$

The condition $2^J \gtrsim \min\{\varepsilon^{-1/t}, \delta^{-1/(t+d)}\}$ implies that term III is asymptotically negligible. The same argument applied to the large deviation estimate of $\tilde{\zeta}_{\lambda,J}$ shows that also term IV is asymptotically negligible. □

Lemma 9.7. *Grant Assumption 3.3 for $s = 0$, $p = 2$ and suppose $f \in H^\alpha$. Then we have for all coefficients with $|\lambda| \leqslant J$*

$$2^{-|\lambda| t} |f^\lambda - f_J^\lambda| \lesssim 2^{-J(t+\alpha)} \|f\|_{H^\alpha}$$

where $f^\lambda := \langle f, \psi_\lambda \rangle$ and $f_J^\lambda := \langle f_J, \psi_\lambda \rangle$.

Proof. We proceed using Lemma 11.1, the mapping property $K : H^{-t} \to L^2$, which we derive by duality from Assumption 3.3 with $(s, p) = (0, 2)$, and the following inverse estimate:

$$\begin{aligned} 2^{-|\lambda| t} |f^\lambda - f_J^\lambda| &\lesssim \|P_J f - f_J\|_{H^{-t}} \lesssim \|K_J(P_J f - f_J)\|_{L^2} \\ &\leqslant \|K P_J f - K f\|_{L^2} \lesssim \|f - P_J f\|_{H^{-t}}. \end{aligned}$$

A simple direct estimate from Appendix 11.3 gives the result. □

9.2 Proof of Theorem 6.1

The error $\mathcal{R}(\hat{f}_{\delta,\varepsilon}^I, f)$ is bounded by a constant times the sum of three terms $I + II + III$, with

$$\begin{aligned} I &:= \|f - f_J\|_{L^2}^2, \\ II &:= \mathbb{E}\left[\|\mathcal{S}_{\kappa, \max\{\delta,\varepsilon\}}(\hat{f}_{\delta,\varepsilon}) - f_J\|_{L^2}^2 \mathbf{1}_{\{\|K_{\delta,J}^{-1}\|_{V_J \to V_J} \leqslant \tau 2^{Jt}\}} \right], \\ III &:= \|f\|_{L^2}^2 \, \mathbb{P}\left(\|K_{\delta,J}^{-1}\|_{V_J \to V_J} > \tau 2^{Jt} \right). \end{aligned}$$

We bound the bias term I by Lemma 9.3. The term III is proved to be negligible by exactly the same arguments as for Theorem 4.2, using now Lemma 9.4. Likewise, by introducing the event $\Omega_{\rho,\delta,J}$, using Lemma 9.4 and repeating the argument of Theorem 4.2, the control of the term II amounts to showing that

$$\begin{aligned} \widetilde{II} &:= \mathbb{E}\left[\|\mathcal{S}_{\kappa, \max\{\delta,\varepsilon\}}(\hat{f}_{\delta,\varepsilon}) - f_J\|_{L^2}^2 \mathbf{1}_{\Omega_{\rho,\delta,J}} \right] \\ &= \sum_{|\lambda| \leqslant J} \mathbb{E}\left[\big(f_{\delta,\varepsilon}^\lambda \mathbf{1}_{\{|f_{\delta,\varepsilon}^\lambda| \geqslant 2^{|\lambda| t} \mathcal{T}(\max\{\delta,\varepsilon\})\}} - f_J^\lambda \big)^2 \mathbf{1}_{\Omega_{\rho,\delta,J}} \right] \end{aligned}$$

has the right order ($f^\lambda_{\hat{\delta},\varepsilon} = \langle \hat{f}_{\delta,\varepsilon}, \psi_\lambda\rangle$). We split $\widetilde{II}$ into the sum of four terms:

$$\begin{aligned}
\widetilde{II}_A &:= \sum_{\lambda\in\mathcal{I}_1} \mathbb{E}[(f^\lambda_{\hat{\delta},\varepsilon} - f^\lambda_J)^2 \mathbf{1}_{\{|f^\lambda_{\hat{\delta},\varepsilon}|\geqslant 2^{|\lambda|t}\mathcal{T}(\max\{\delta,\varepsilon\})\}} \cap \Omega_{\rho,\delta,J}],\\
\widetilde{II}_B &:= \sum_{\lambda\in\mathcal{I}_2} \mathbb{E}[(f^\lambda_{\hat{\delta},\varepsilon} - f^\lambda_J)^2 \mathbf{1}_{\{|f^\lambda_{\hat{\delta},\varepsilon}|\geqslant 2^{|\lambda|t}\mathcal{T}(\max\{\delta,\varepsilon\})\}} \cap \Omega_{\rho,\delta,J}],\\
\widetilde{II}_C &:= \sum_{\lambda\in\mathcal{I}_3} (f^\lambda_J)^2\, \mathbb{P}(\{|f^\lambda_{\hat{\delta},\varepsilon}| < 2^{|\lambda|t}\mathcal{T}(\max\{\delta,\varepsilon\})\} \cap \Omega_{\rho,\delta,J}),\\
\widetilde{II}_D &:= \sum_{\lambda\in\mathcal{I}_4} (f^\lambda_J)^2\, \mathbb{P}(\{|f^\lambda_{\hat{\delta},\varepsilon}| < 2^{|\lambda|t}\mathcal{T}(\max\{\delta,\varepsilon\})\} \cap \Omega_{\rho,\delta,J}),
\end{aligned}$$

where

$$\begin{aligned}
&\mathcal{I}_1 := \big\{|f^\lambda| > 2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})\big\},\ \mathcal{I}_2 := \big\{|f^\lambda| \leqslant 2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})\big\},\\
&\mathcal{I}_3 := \big\{|f^\lambda_J| > 2^{|\lambda|t+1}\mathcal{T}(\max\{\delta,\varepsilon\})\big\},\ \mathcal{I}_4 := \big\{|f^\lambda_J| \leqslant 2^{|\lambda|t+1}\mathcal{T}(\max\{\delta,\varepsilon\})\big\}.
\end{aligned}$$

Observe the different splitting decision, according to the size of the coefficients f^λ and f^λ_J, respectively. By doing so, we avoid an additional control on $\|f_J\|_{B^s_{p,p}}$.

We first bound the nonlinear approximation term $\widetilde{II}_D$ by $\sum_\lambda (f^\lambda_J)^2 \mathbf{1}_{\{|f^\lambda_J|\leqslant 2^{|\lambda|t+1}\mathcal{T}(\max\{\delta,\varepsilon\})\}}$. Using for $|a| \leqslant \eta$ and arbitrary $b \in \mathbb{R}$ the general inequality $|a| \leqslant |a-b| + |b|\mathbf{1}_{\{|b|<2\eta\}}$, we further bound $\widetilde{II}_D$ by

$$2\sum_\lambda \big((f^\lambda_J - f^\lambda)^2 + (f^\lambda)^2 \mathbf{1}_{\{|f^\lambda|\leqslant 2^{|\lambda|t+2}\mathcal{T}(\max\{\delta,\varepsilon\})\}}\big).$$

The sum over the first term equals $2\|f - f_J\|^2_{L^2}$ and is by Lemma 9.3 of the right order. The sum over the second term is under restriction (3.2) for p bounded in order by $\|f\|^2_{B^s_{p,p}}\mathcal{T}(\max\{\delta,\varepsilon\})^{2-p}$, which is classical and follows e.g. from Theorem 7.1 in [9]. Thus, $\widetilde{II}_D$ has the right order.

Concerning the second approximation term $\widetilde{II}_A$, we have

$$\widetilde{II}_A \leqslant \sum_{|f^\lambda|>2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})} \mathbb{E}[(f^\lambda_{\hat{\delta},\varepsilon} - f^\lambda_J)^2 \mathbf{1}_{\Omega_{\rho,\delta,J}}].$$

By decomposition (9.1) and Lemma 9.5 we obtain on $\Omega_{\rho,\delta,J}$

$$\begin{aligned}
f^\lambda_{\hat{\delta},\varepsilon} - f^\lambda_J &= \delta 2^{|\lambda|t}\|f_J\|_{L^2}\, c_\lambda\, \xi_\lambda + \varepsilon 2^{|\lambda|t}\tilde{c}_\lambda\, \tilde{\xi}_\lambda\\
&\quad + \delta^2\, 2^{|\lambda|t}\, \|f_J\|_{L^2}\, 2^{J[d+t])}\, \zeta_{\lambda J} + \delta\varepsilon\, 2^{|\lambda|t}\, 2^{J[t+d/2]}\, \tilde{\zeta}_{\lambda,J}.
\end{aligned}$$

Therefore, by Lemma 9.5 we find $\mathbb{E}[(f^\lambda_{\hat{\delta},\varepsilon} - f^\lambda_J)^2 \mathbf{1}_{\Omega_{\rho,\delta,J}}] \lesssim \max\{\delta,\varepsilon\}^2 2^{2|\lambda|t}$. It follows that $\widetilde{II}_A$ is bounded by

$$\begin{aligned}
\widetilde{II}_A &\lesssim \max\{\delta,\varepsilon\}^2 \sum_{|f^\lambda|>2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})} 2^{2|\lambda|t}\\
&\leqslant \max\{\delta,\varepsilon\}^2 \mathcal{T}(\max\{\delta,\varepsilon\})^{-p} \sum_{|f^\lambda|>2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})} 2^{(2-p)|\lambda|t}|f^\lambda_J|^p\\
&\lesssim \mathcal{T}(\max\{\delta,\varepsilon\})^{2-p} \sum_{j\leqslant J} 2^{(2-p)jt} \sum_{|\lambda|=j} |f^\lambda|^p.
\end{aligned}$$

Next,

$$\sum_{j \leqslant J} 2^{j(sp+d(p/2-1))} \sum_{|\lambda|=j} |f^\lambda|^p \lesssim \|f\|^p_{B^s_{p,p}} \leqslant M^p \tag{9.4}$$

yields

$$\widetilde{II_A} \lesssim \mathcal{T}(\max\{\delta,\varepsilon\})^{2-p} \sum_{j \leqslant J} 2^{j(-sp-d(p/2-1)+t(2-p))}\, 2^{j(sp+d(p/2-1))} \sum_{|\lambda|=j} |f^\lambda|^p.$$

Noting that $-sp - d(p/2-1) + t(2-p) = 0$ in view of (9.4), we derive

$$\widetilde{II_A} \lesssim \mathcal{T}(\max\{\delta,\varepsilon\})^{2-p} = \max\{\varepsilon\sqrt{|\log\varepsilon|}, \delta\sqrt{|\log\delta|}\}^{4s/(2s+2t+d)}.$$

We now turn to the first deviation term $\widetilde{II_B}$. By the Cauchy-Schwarz inequality $\widetilde{II_B}$ is less than

$$\sum_{|\lambda| \leqslant J} \mathbb{E}[|f^\lambda_{\delta,\varepsilon} - f^\lambda_J|^4 \mathbf{1}_{\Omega_{\rho,\delta,J}}]^{\frac{1}{2}}\, \mathbb{P}\left(\{|f^\lambda_{\delta,\varepsilon} - f^\lambda| \geqslant 2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})\} \cap \Omega_{\rho,\delta,J}\right)^{\frac{1}{2}}.$$

By the same argument as for $\widetilde{II_A}$ to bound the expectation term, but using now moments of order 4, we further bound $\widetilde{II_B}$ by

$$\max\{\delta,\varepsilon\}^2 \sum_{|\lambda| \leqslant J} 2^{2t|\lambda|}\, \mathbb{P}\left(\{|f^\lambda_{\delta,\varepsilon} - f^\lambda| \geqslant 2^{|\lambda|t-1}\mathcal{T}(\max\{\delta,\varepsilon\})\} \cap \Omega_{\rho,\delta,J}\right)^{1/2}.$$

We infer from Lemma 9.7 and restriction (6.1) the estimate $2^{-|\lambda|t}|f^\lambda - f^\lambda_J| \lesssim 2^{-J(t+\alpha)} \lesssim \max\{\delta,\varepsilon\}$ for all λ such that this difference is asymptotically small compared to $\mathcal{T}(\max\{\delta,\varepsilon\})$. Consequently, the triangle inequality and Lemma 9.6 yield for any $\gamma > 0$ a bound of order

$$\max\{\delta,\varepsilon\}^2 \sum_{|\lambda| \leqslant J} 2^{2t|\lambda|} \max\{\delta,\varepsilon\}^{\gamma/2} \lesssim \max\{\delta,\varepsilon\}^{2+\gamma/2} 2^{J(2t+d)}.$$

For $\gamma = 4 + 2d/t$ this yields the bound $\max\{\delta,\varepsilon\}^2$ and $\widetilde{II_B}$ proves negligible. We eventually consider $\widetilde{II_C}$:

$$\widetilde{II_C} \lesssim \sum_{|\lambda| \leqslant J} (f^\lambda_J)^2\, \mathbb{P}(\{|f^\lambda_{\delta,\varepsilon} - f^\lambda_J| \geqslant 2^{|\lambda|t+1}\mathcal{T}(\max\{\delta,\varepsilon\})\} \cap \Omega_{\rho,\delta,J}),$$

and a straightforward application of Lemma 9.6 shows that this term is also negligible.

10 Proof for nonlinear estimation II

10.1 Deviation bounds in H^t-norm

We need precise deviation bound of the hard thresholding estimator in H^t-loss and must also deal with increasing signal noise intensity. The following bounds seem to be new.

Lemma 10.1. *Assume $\kappa > 4\sqrt{t/d}$ and $2^J \lesssim \varepsilon^{-1/t}$. Let $s \geqslant 0$, $p > 0$ satisfy restriction (3.2). There exist constants c_0, η_0, $R_0 > 0$ such that for all functions $g \in B^{s+t}_{p,p}$ the hard thresholding estimate $\hat{g}_\varepsilon$ satisfies for all $\eta > \eta_0$ and $R > R_0$:*

$$\begin{aligned}\mathbb{P}\left(\mathcal{T}(\varepsilon)^{-r(s,t,d)}\|\hat{g}_\varepsilon - P_J g\|_{H^t} \geqslant \eta \max\{\|P_J g\|_{B^{s+t}_{pp}}, \|P_J g\|^{p/2}_{B^{s+t}_{p,p}}\}^{\frac{2t+d}{2s+2t+d}}\right)&\\ \lesssim \varepsilon^{c_0\eta} + \varepsilon^{\kappa^2/8-d/t},&\\ \mathbb{P}\left(\|\hat{g}_\varepsilon - P_J g\|_{H^t} \geqslant R\max\{\|P_J g\|_{B^{s+t}_{pp}}, \|P_J g\|^{p/2}_{B^{s+t}_{p,p}}\}\right) \lesssim \varepsilon^{\kappa^2/16-d/t}R^{-4}.&\end{aligned}$$

Proof. Denote by g^λ and g^λ_ε the wavelet coefficients of g and g_ε. We have

$$\|\hat{g}_\varepsilon - P_J g\|^2_{H^t} \sim \sum_{|\lambda| \leqslant J} 2^{2|\lambda|t}(g^\lambda_\varepsilon \mathbf{1}_{|g^\lambda_\varepsilon| \geqslant \mathcal{T}(\varepsilon)} - g^\lambda)^2.$$

The usual decomposition yields a bound of the right-hand side by the sum of four terms $I + II + III + IV$ with

$$\begin{aligned} I &:= \sum 2^{2|\lambda|t}(g^\lambda_\varepsilon - g^\lambda)^2 \mathbf{1}_{\{|g^\lambda_\varepsilon| \geqslant \mathcal{T}(\varepsilon)\}}\mathbf{1}_{\{|g^\lambda| \geqslant \frac{1}{2}\mathcal{T}(\varepsilon)\}}\\ &\leqslant \sum 2^{2|\lambda|t}(g^\lambda_\varepsilon - g^\lambda)^2 \mathbf{1}_{\{|g^\lambda| \geqslant \frac{1}{2}\mathcal{T}(\varepsilon)\}},\\ II &:= \sum 2^{2|\lambda|t}(g^\lambda_\varepsilon - g^\lambda)^2 \mathbf{1}_{\{|g^\lambda_\varepsilon| \geqslant \mathcal{T}(\varepsilon)\}}\mathbf{1}_{\{|g^\lambda| < \frac{1}{2}\mathcal{T}(\varepsilon)\}}\\ &\leqslant \sum 2^{2|\lambda|t}(g^\lambda_\varepsilon - g^\lambda)^2 \mathbf{1}_{\{|g^\lambda_\varepsilon - g^\lambda| > \frac{1}{2}\mathcal{T}(\varepsilon)\}},\\ III &:= \sum 2^{2|\lambda|t}(g^\lambda)^2 \mathbf{1}_{\{|g^\lambda_\varepsilon| < \mathcal{T}(\varepsilon)\}}\mathbf{1}_{\{|g^\lambda| \geqslant 2\mathcal{T}(\varepsilon)\}}\\ &\leqslant \sum 2^{2|\lambda|t}(g^\lambda)^2 \mathbf{1}_{\{|g^\lambda_\varepsilon - g^\lambda| > \mathcal{T}(\varepsilon)\}},\\ IV &:= \sum 2^{2|\lambda|t}(g^\lambda)^2 \mathbf{1}_{\{|g^\lambda_\varepsilon| < \mathcal{T}(\varepsilon)\}}\mathbf{1}_{\{|g^\lambda| < 2\mathcal{T}(\varepsilon)\}}\\ &\leqslant \sum 2^{2|\lambda|t}(g^\lambda)^2 \mathbf{1}_{\{|g^\lambda| < 2\mathcal{T}(\varepsilon)\}} \end{aligned}$$

and where the sums in λ range through the set $\{|\lambda| \leq J\}$. The approximation term IV is bounded by

$$\begin{aligned} &\mathcal{T}(\varepsilon)^{2-p}\sum_{j \leqslant J} 2^{2jt} \sum_{|\lambda| = j} \min\{(g^\lambda)^p, (2\mathcal{T}(\varepsilon))^p\}\\ \lesssim &\mathcal{T}(\varepsilon)^{2-p}\sum_{j \leqslant J} 2^{2jt}\min\{\|P_J g\|^p_{B^{s+t}_{p,p}} 2^{-j(s+t+d/2-d/p)p}, 2^{jd}\mathcal{T}(\varepsilon)^p\} \end{aligned}$$

which is of order $\mathcal{T}(\varepsilon)^2 2^{j_0(2t+d)}$ with

$$2^{j_0(2s+2t+d)} \sim \min\left\{\|P_J g\|^2_{B^{s+t}_{p,p}}\mathcal{T}(\varepsilon)^{-2}, 2^{J(2s+2t+d)}\right\}.$$

Therefore, we obtain

$$IV \lesssim \|P_J g\|^2_{B^{s+t}_{p,p}}\left(\frac{\varepsilon\sqrt{|\log\varepsilon|}}{\|P_J g\|_{B^{s+t}_{p,p}}}\right)^{2r(s,t,d)}.$$

For the second approximation term I we need to introduce the random variables

$$\xi_j := \frac{\varepsilon^{-2}}{\#\{|\lambda| = j,\, |g^\lambda| \geqslant \frac{1}{2}\mathcal{T}(\varepsilon)\}}\sum_{|\lambda| = j}(g^\lambda_\varepsilon - g^\lambda)^2 \mathbf{1}_{\{|g^\lambda| \geqslant \frac{1}{2}\mathcal{T}(\varepsilon)\}}.$$

Using $\mathbf{1}_{\{|g^\lambda|\geqslant\frac{1}{2}\mathcal{T}(\varepsilon)\}} \leqslant |2g^\lambda/\mathcal{T}(\varepsilon)|^p$, we obtain for $1/p = 1/2 + s/(2t+d)$ that I is bounded by

$$\begin{aligned}\sum_{j\leqslant J} 2^{2jt}\varepsilon^2\xi_j \sum_{|\lambda|=j} \mathbf{1}_{\{|g^\lambda|\geqslant\frac{1}{2}\mathcal{T}(\varepsilon)\}} &\lesssim \sum_{j\leqslant J} 2^{2jt}\varepsilon^2\xi_j \min\Big\{\mathcal{T}(\varepsilon)^{-p}\sum_{|\lambda|=j}|g^\lambda|^p, 2^{jd}\Big\}\\ &\lesssim \sum_{j\leqslant J}\varepsilon^2\xi_j \min\Big\{\mathcal{T}(\varepsilon)^{-p}2^{-j(s+t+d/2-d/p)p+2jt}\|P_J g\|^p_{B^{s+t}_{p,p}}, 2^{j(2t+d)}\Big\}.\end{aligned}$$

Now observe that, as before, the following inequality holds:

$$\sum_{j\leqslant J}\varepsilon^2 \min\Big\{\mathcal{T}(\varepsilon)^{-p}2^{-j(s+t+d/2-d/p)p+2jt}\|P_J g\|^p_{B^{s+t}_{p,p}}, 2^{j(2t+d)}\Big\} \sim \varepsilon^2 2^{j_1(2t+d)},$$

$$2^{j_1(2s+2t+d)} \sim \min\big\{\|P_J g\|^p_{B^{s+t}_{p,p}}\mathcal{T}(\varepsilon)^{-2}, 2^{J(2s+2t+d)}\big\}.$$

By definition, each ξ_j is the arithmetic mean of the squares of independent normalized Gaussian random variables. By independence and standard Gaussian estimates, for any sequence (a_j) with $\|(a_j)\|_{l^1} = 1$, Markov's inequality yields for any $\eta > 0$:

$$\mathbb{P}\big(\sum_j a_j\xi_j \geqslant \eta\big) \leqslant \exp(-\eta/3)\prod_j \mathbb{E}[\exp(a_j\xi_j/3)] \leqslant \exp(-\eta/3).$$

Consequently, we obtain $\mathbb{P}(I \geqslant \eta\varepsilon^2 2^{j_1(2t+d)}) \leqslant \exp(-c_1\eta)$ with a constant $c_1 > 0$. Substituting for j_1, we conclude with another constant $c_2 > 0$ that

$$\mathbb{P}\big(I \geqslant \eta\|P_J g\|^p_{B^s_{p,p}}\big(\mathcal{T}(\varepsilon)\|P_J g\|^{-p/2}_{B^{s+t}_{p,p}}\big)^{2r(s,t,d)}\big) \leqslant \exp\big(-c_2\eta|\log\varepsilon|\big).$$

Considering the deviation terms II and III, we observe

$$\begin{aligned}\mathbb{P}(\{II=0\}\cap\{III=0\}) \geqslant &\mathbb{P}(|g^\varepsilon_\lambda - g_\lambda| \leqslant \tfrac{1}{2}\mathcal{T}(\varepsilon) \text{ for all } |\lambda| \leqslant J)\\ \geqslant &\big(1-\exp(-\kappa^2|\log\varepsilon|/8)\big)^{2^{Jd}}.\end{aligned}$$

Using $2^{Jd} \lesssim \varepsilon^{-d/t}$, we derive $\mathbb{P}(II+III > 0) \leqslant 1-(1-\varepsilon^{\kappa^2/8})^{2^{Jd}}$ which is of order $\varepsilon^{\kappa^2/8-d/t}$. Therefore we obtain for some constants $c_3 > 0$ and $\eta_1 > 0$ and for all $\eta > \eta_1$:

$$\begin{aligned}&\mathbb{P}\big(\|\hat{g}_\varepsilon - P_J g\|_{H^t} \geqslant \eta\max\{\|P_J g\|_{B^{s+t}_{pp}}, \|P_J g\|^{p/2}_{B^{s+t}_{pp}}\}^{\frac{2t+d}{2s+2t+d}}\mathcal{T}(\varepsilon)^{r(s,t,d)}\big)\\ &\leqslant \mathbb{P}\big(I > \tfrac{1}{3}\eta^2\|P_J g\|^{\frac{p(2t+d)}{2s+2t+d}}_{B^{s+t}_{p,p}}\mathcal{T}(\varepsilon)^{2r(s,t,d)}\big) + \mathbb{P}(II+III > 0)\\ &\quad + \mathbb{P}\big(IV > \|P_J g\|^{\frac{2(2t+d)}{2s+2t+d}}_{B^{s+t}_{p,p}}\mathcal{T}(\varepsilon)^{2r(s,t,d)}\big)\\ &\lesssim \varepsilon^{c_2\eta} + 0 + \varepsilon^{\kappa^2/8-d/t}.\end{aligned}$$

On the other hand, the deviation terms are well bounded in probability. While obviously $III \leqslant \|P_J g\|^2_{B^{s+t}_{p,p}}$ holds by the Cauchy-Schwarz inequality, the term $\mathbb{E}[II]$ is less than

$$\sum_{|\lambda|\leqslant J} 2^{2|\lambda|t}\,\mathbb{E}[(g^\lambda_\varepsilon - g^\lambda)^4]^{1/2}\,\mathbb{P}(|g^\lambda_\varepsilon - g^\lambda| > \mathcal{T}(\varepsilon)/2)^{1/2}.$$

This is bounded in order by $2^{J(2t+d)}\varepsilon^2 \exp(\kappa^2|\log\varepsilon|/8)^{1/2} \sim \varepsilon^{\kappa^2/16-d/t}$ due to $2^J \lesssim \varepsilon^{-1/t}$. In the same way we find that $\operatorname{Var}[II]$ is less than

$$\sum_{|\lambda|\leqslant J} 2^{4|\lambda|t}\,\mathbb{E}[(g_\varepsilon^\lambda - g^\lambda)^8]^{1/2}\,\mathbb{P}(|g_\varepsilon^\lambda - g^\lambda| > \mathcal{T}(\varepsilon)/2)^{1/2} \lesssim \varepsilon^{\kappa^2/16-d/t}.$$

By Chebyshev's inequality, we infer $\mathbb{P}(II \geqslant R) \lesssim \varepsilon^{\kappa^2/16-d/t}R^{-2}$ for $R > 0$. Since the above estimates of the approximation terms yield superoptimal deviation bounds, we obtain altogether for any $R > 2$:

$$\mathbb{P}\left(\|\hat{g}_\varepsilon - P_J g\|_{H^t} \geqslant R\max\{\|P_J g\|_{B^{s+t}_{pp}}, \|P_J g\|^{p/2}_{B^{s+t}_{pp}}\}\right) \lesssim \varepsilon^{\kappa^2/16-d/t}R^{-4}.$$

□

10.2 Estimation in operator norm

Proposition 10.2. *Suppose $\kappa^2 > 32\max\{d/t, 1 + t(2t+d)/(4t(t+d))\}$. Grant Assumption 3.3 with $s \geqslant 0$, $p \geqslant 1$ and Assumption 3.4 with $\bar{s} > 0, \bar{p} > 0$, satisfying in addition restriction* (6.3) *with strict inequality for $p > 1$. We have*

$$\mathbb{E}\left[\|\hat{K}_\delta - K_J\|^2_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t}\right] \lesssim \left(\delta\sqrt{|\log\delta|}\right)^{2r(\bar{s},t,d)}.$$

Proof. The wavelet characterisation of Besov spaces (cf. Appendix 11.3) together with Hölder's inequality for $p^{-1} + q^{-1} = 1$ yields

$$\begin{aligned}
&\|\hat{K}_\delta - K_J\|_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t}\\
&\sim \sup_{\|(a_\mu)\|_{l^p}=1} \|(\hat{K}_\delta - K)\big(\sum_{|\mu|\leqslant J} 2^{-\mu(s+d/2-d/p)}a_\mu\psi_\mu\big)\|_{H^t}\\
&\leqslant \|\big(2^{-|\mu|(s+d/2-d/p)}\|(\hat{K}_\delta - K)\psi_\mu\|_{H^t}\big)_{|\mu|\leqslant J}\|_{l^q}\\
&\leqslant \|(2^{|\mu|(-(s+d/2-d/p)+(\bar{s}+d/2-d/\bar{p})(2t+d)/(2\bar{s}+2t+d))})_{|\mu|\leqslant J}\|_{l^q}\\
&\quad\bullet \sup_{|\mu|\leqslant J} 2^{-|\mu|(\bar{s}+d/2-d/\bar{p})(2t+d)/(2\bar{s}+2t+d)}\|(\hat{K}_\delta - K)\psi_\mu\|_{H^t}.
\end{aligned}$$

The l^q-norm of the powers of 2 evaluates to

$$\|(2^{j(-(s-d/2)+(\bar{s}+d/2-d/\bar{p})(2t+d)/(2\bar{s}+2t+d))})_{j\leqslant J}\|_{l^q},$$

which is of order one whenever restriction (6.3) is fulfilled with strict inequality for $q < \infty$.

By construction, $\hat{K}_\delta\psi_\mu$ is the hard thresholding estimator for $K\psi_\mu$ given the observation of $K_\delta\psi_\mu$, which is $K\psi_\mu$ corrupted by white noise of level δ. Therefore, under Assumption 3.4, Lemma 10.1 applied to $K\psi_\mu$ and δ gives for any $\eta \geqslant \eta_0$:

$$\begin{aligned}
\mathbb{P}\left(\|(\hat{K}_\delta - K)\psi_\mu\|_{H^t} \geqslant \eta\|K\psi_\mu\|^{(2t+d)/(2s+2t+d)}_{B^{\bar{s}+t}_{\bar{p},\bar{p}}}\mathcal{T}(\delta)^{r(\bar{s},t,d)}\right)&\\
\lesssim \delta^{c_0\eta} + \delta^{\kappa^2/8-d/t}&.
\end{aligned}$$

By estimating the probability of the supremum by the sum over the probabilities, we obtain with a constant $c_1 > 0$ for all $\eta \geqslant \eta_0$:

$$\begin{aligned}
&\mathbb{P}\left(\|\hat{K}_\delta - K_J\|_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t} \geqslant \eta\mathcal{T}(\delta)^{r(\bar{s},t,d)}\right)\\
&\leqslant \sum_{|\mu|\leqslant J} \mathbb{P}\left(\|(\hat{K}_\delta - K)\psi_\mu\|_{H^t} \geqslant c_1 2^{|\mu|(\bar{s}+d/2-d/\bar{p})\frac{2t+d}{2s+2t+d}}\eta\mathcal{T}(\delta)^{r(\bar{s},t,d)}\right)\\
&\lesssim 2^{Jd}(\delta^{c_0\eta} + \delta^{\kappa^2/8-d/t})\\
&\lesssim \delta^{c_0\eta - d/(t+d)} + \delta^{\kappa^2/8-d(2t+d)/(t(t+d))}.
\end{aligned}$$

For a sufficiently large $\eta_1 > \eta_0$, depending only on c_0, d and t, with $\gamma := \kappa^2/8 - d(2t+d)/(t(t+d)) > 0$, we thus obtain:

$$\mathbb{P}\left(\|\hat{K}_\delta - K_J\|_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t} \geqslant \eta_1\mathcal{T}(\delta)^{r(\bar{s},t,d)}\right) \lesssim \delta^\gamma.$$

By the above bound on the operator norm and Hölder's inequality for $q := \gamma/2 > 2$ and $\rho^{-1} + q^{-1} = 1$ together with the second estimate in Lemma 10.1, we find for some constant $R_0 > 0$:

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{K}_\delta - K_J\|^2_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t}\right]\\
&\lesssim \eta_1\mathcal{T}(\delta)^{2r(\bar{s},t,d)} + \mathbb{E}\left[\|\hat{K}_\delta - K_J\|^{2\rho}_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t}\right]^{1/\rho}\delta^{\gamma/q}\\
&\lesssim \mathcal{T}(\delta)^{2r(\bar{s},t,d)} + \Big(\int_0^\infty R^{2\rho-1}\,\mathbb{P}(\|\hat{K}_\delta - K_J\|_{(V_J,\|\bullet\|_{B^s_{p,p}})\to H^t} \geqslant R)\,dR\Big)^{1/\rho}\delta^2\\
&\lesssim \mathcal{T}(\delta)^{2r(\bar{s},t,d)} + \Big(R_0 + \int_{R_0}^\infty R^{2\rho-1}2^{Jd}\delta^{\kappa^2/16-d/t}R^{-4}\,dR\Big)^{1/\rho}\delta^2\\
&\lesssim \mathcal{T}(\delta)^{2r(\bar{s},t,d)} + \max\{\delta^{(\kappa^2/16-2d/t)/\rho}, 1\}\delta^2
\end{aligned}$$

which is of order $\mathcal{T}(\delta)^{2r(\bar{s},t,d)}$ by assumption on κ. □

Lemma 10.3. *Grant Assumption 3.3 and assume $\kappa^2 > 4d/(t+d)$. Then*

$$\mathbb{P}\left(\|\hat{K}_\delta - K_J\|_{L^2\to H^t} \geqslant c_0(c_J^{t+d} + \eta)\right) \lesssim \delta^{c_0\eta},$$

where $c_0 > 0$ is a constant independent of J, c_J from (5.8)*, δ and η, but depending on κ.*

Proof. For $|\mu|, |\lambda| \leqslant J$ we have for the entries in the wavelet representation

$$|(\hat{K}_\delta)_{\mu,\lambda} - K_{\mu,\lambda}| = |K_{\mu,\lambda}|\mathbf{1}_{\{|(K_\delta)_{\mu,\lambda}|\leqslant\mathcal{T}(\delta)\}} + \delta|\dot{B}_{\mu,\lambda}|\mathbf{1}_{\{|(K_\delta)_{\mu,\lambda}|>\mathcal{T}(\delta)\}}.$$

A simple rough estimate yields

$$|(\hat{K}_\delta)_{\mu,\lambda} - K_{\mu,\lambda}| \leqslant 2\mathcal{T}(\delta) + |K_{\mu,\lambda}|\mathbf{1}_{\{|(K_\delta - K)_{\mu,\lambda}|\geqslant\mathcal{T}(\delta)\}} + \delta|\dot{B}_{\mu,\lambda}|.$$

We now bound the operator norm by the corresponding Hilbert-Schmidt norm and apply the estimate $2^{|\lambda|t}|K_{\mu,\lambda}| \leqslant \|K\psi_\mu\|_{B^{s+t}_{p,p}} \lesssim 1$ derived from Assumption

3.3 to obtain

$$\begin{aligned}\|\hat{K}_\delta - K_J\|^2_{(V_J,\|\bullet\|_{L^2})\to H^t} &\leqslant \sum_{|\mu|,\,|\lambda|\leqslant J} 2^{2|\lambda|t}\big((\hat{K}_\delta)_{\mu,\lambda} - K_{\mu,\lambda}\big)^2\\ &\lesssim 2^{2J(t+d)}\mathcal{T}(\delta)^2 + \#\{\delta|(K_\delta - K)_{\mu,\lambda}| \geqslant \mathcal{T}(\delta)\} + \delta^2 2^{2Jt}\sum_{|\mu|,\,|\lambda|\leqslant J} \dot{B}^2_{\mu,\lambda}\\ &= 2^{2J(t+d)}\mathcal{T}(\delta)^2 + \#\{|\dot{B}_{\mu,\lambda}| \geqslant \kappa\sqrt{|\log\delta|}\} + \delta^2 2^{2Jt}\sum_{|\mu|,\,|\lambda|\leqslant J} \dot{B}^2_{\mu,\lambda}\end{aligned}$$

where the cardinality is taken for multi-indices (λ,μ) such that $|\lambda|,|\mu|\leq J$. The first term is bounded by $c_J^{2(t+d)}$. The second term is a Binomial random variable with expectation $2^{2Jd}\mathbb{P}(|\dot{B}_{\mu,\lambda}|\geqslant \kappa|\log\delta|^{1/2}) \lesssim \delta^{-2d/(t+d)+\kappa^2/2}$. An exponential moment bound for the Binomial distribution yields

$$\mathbb{P}\big(\#\{|\dot{B}_{\mu,\lambda}|\geqslant \kappa\sqrt{|\log\delta|}\}\geqslant \eta\big)\lesssim \exp(-\eta(-2d/(t+d)+\kappa^2/2)|\log\delta|),$$

which evaluates to $\delta^{\eta(\kappa^2/2-2d/(t+d))}$.

For the last term, we use a rough deviation bound for the χ^2-square distribution, namely

$$\mathbb{P}\big(2^{-2Jd}\sum_{|\mu|,\,|\lambda|\leqslant J}\dot{B}^2_{\mu,\lambda}\geqslant\eta\big)\leqslant \exp(-\eta/2)$$

to infer from $2^{J(t+d)}\lesssim \mathcal{T}(\delta)$ that

$$\mathbb{P}\big(\delta^2 2^{2Jt}\sum_{|\mu|,\,|\lambda|\leqslant J}\dot{B}^2_{\mu,\lambda}\geqslant\eta\big)\leqslant \exp\big(-2^{-2J(t+d)-1}\delta^{-2}\eta\big)\leqslant e^{-c_0\eta|\log\delta|}=\delta^{c_1\eta}$$

holds with a constant $c_1>0$. The choice $c_0=\min\{c_1,\kappa^2/2-2d/(t+d)\}$ gives the result. □

10.3 Proof of Theorem 6.4

For $\rho\in(0,1)$ we introduce the event

$$\Omega^{II}_{\rho,\delta,J} := \big\{\|\hat{K}_\delta - K_J\|_{(V_J,\|\bullet\|_{L^2})\to H^t}\leqslant \rho\|K_J^{-1}\|^{-1}_{(V_J,\|\bullet\|_{H^t})\to L^2}\big\}. \tag{10.1}$$

The Neumann series representation implies that on $\Omega^{II}_{\rho,\delta,J}$ the random operator

$$\hat{K}_\delta : (V_J,\|\bullet\|_{L^2})\to (V_J,\|\bullet\|_{H^t})$$

is invertible with norm $\|\hat{K}_\delta^{-1}\|\leqslant (1-\rho)^{-1}\|K_J\|^{-1}$. For the choice $\rho\in(0,1-c'_K/\tau)$ this bound is smaller than the cut-off value τ.

On $\Omega^{II}_{\rho,\delta,J}$ and assuming $\rho\in(0,1-c'_K/\tau)$, we bound $\|\hat{f}^{II}_{\delta,\varepsilon}-f\|_{L^2}$ by

$$\|\hat{K}_\delta^{-1}(\hat{g}_\varepsilon - P_J g)\|_{L^2} + \|(\hat{K}_\delta^{-1}-K_J^{-1})P_J g\|_{L^2} + \|f_J - f\|_{L^2}.$$

The first two stochastic errors are further bounded by

$$\|\hat{K}_\delta^{-1}\|_{(V_J,\|\bullet\|_{H^t})\to L^2}(\|\hat{g}_\varepsilon - P_J g\|_{H^t} + \|\hat{K}_\delta - K_J\|_{B^s_{p,p}\to H^t}\|f_J\|_{B^s_{p,p}}).$$

Because of $\|\hat{K}_\delta^{-1}\|_{(V_J,\|\bullet\|_{H^t})\to L^2} \lesssim 1$, the assertion on $\Omega^{II}_{\rho,\delta,J}$ follows from the classical moment estimate for hard thresholding in analogy to Lemma 10.1, from the operator norm estimate of Proposition 10.2 and the Galerkin estimate in Lemma 9.2.

On the complement $(\Omega^{II}_{\rho,\delta,J})^c$ the risk of $\hat{f}^{II}_{\delta,\varepsilon}$, conditional on $\dot{B}$, is uniformly bounded thanks to the cut-off rule in the construction. Consequently, Theorem 6.4 follows from

$$\mathbb{P}\left((\Omega^{II}_{\rho,\delta,J})^c\right) \lesssim \delta^2.$$

By Lemma 10.3, this last bound is fulfilled for a sufficiently small choice of c_J, depending on ρ and thus on c_K' and τ.

11 Appendix

11.1 Proofs for the linear estimator

Preparations

Lemma 11.1. *Under Assumption 3.3 with* $(s,p)=(0,2)$ *we have*

$$\|K_j^{-1}\|_{V_j\to V_j} \lesssim 2^{jt}.$$

Proof. Under Assumption 3.3 the following mapping property is proved in [10]:

$$\|K_j^{-1}h\|_{L^2} \lesssim \|h\|_{H^t} \text{ for } h\in V_j. \tag{11.1}$$

Therefore $\|K_j^{-1}\|^2_{V_j\to V_j} \lesssim \sup_{h\in V_j,\|h\|_{L^2}=1}\|h\|^2_{H^t} \lesssim 2^{2jt}$ follows from an inverse estimate (see Appendix 11.3). □

Let $\rho\in(0,1)$. On the event $\Omega_{\rho,\delta,j} := \{\delta\|K_j^{-1}\dot{B}_j\|_{V_j\to V_j} \leqslant \rho\}$, the random operator $(\mathrm{Id}+\delta K_j^{-1}\dot{B}_j)^{-1}$is well defined by the usual Neumann series argument and satisfies

$$\begin{aligned}\|(\mathrm{Id}+\delta K_j^{-1}\dot{B}_j)^{-1}\|_{V_j\to V_j}\mathbf{1}_{\Omega_{\rho,\delta,j}} &\leqslant (1-\delta\|K_j^{-1}\dot{B}_j\|_{V_j\to V_j})^{-1}\mathbf{1}_{\Omega_{\rho,\delta,j}}\\ &\leqslant (1-\rho)^{-1}.\end{aligned} \tag{11.2}$$

Lemma 11.2. *Grant Assumption 3.3 with* $(s,p)=(0,2)$. *Let* $\eta := 1-(2t+d)/(2s+2t+d) > 0$ *and let* j *be specified as in Proposition 4.2. There is a constant* $c_\Omega > 0$ *such that for all sufficiently small* $\delta > 0$

$$\mathbb{P}(\Omega^c_{\rho,\delta,j}) \leqslant \exp(-c_\Omega\rho\delta^{-\eta}2^{2jd}),\quad \rho>0,\, j\in\mathbb{N}.$$

Proof. By definition of c_K and Lemma 11.1 the inclusion

$$\begin{aligned}\Omega^c_{\rho,\delta,j} &\subset \{\|\dot{B}_j\|_{V_j\to V_j} > \rho\delta^{-1}c_K^{-1}2^{-jt}\}\\ &= \{2^{-jd/2}\|\dot{B}_j\|_{V_j\to V_j} > \rho c_K^{-1}\delta^{-1}2^{-j(2t+d)/2}\}\end{aligned}$$

is valid. Using $\delta^{-1}2^{-j(2t+d)/2} \gtrsim \delta^{-\eta}$ and Lemma 9.1 we obtain the result. □

Proof of Proposition 4.2

By definition, $\mathcal{R}(\hat{f}_{\delta,\varepsilon}, f)$ is bounded by a constant times the sum of three terms $I + II + III$, with

$$\begin{aligned} I &:= \|f - f_j\|_{L^2}^2, \\ II &:= \mathbb{E}[\|(K_{\delta,j}^{-1} P_j g_\varepsilon - f_j)\mathbf{1}_{\{\|K_{\delta,j}^{-1}\|_{V_j \to V_j} \leqslant \tau 2^{jt}\}}\|_{L^2}^2], \\ III &:= \|f\|_{L^2}^2 \,\mathbb{P}(\|K_{\delta,j}^{-1}\|_{V_j \to V_j} > \tau 2^{jt}) \end{aligned}$$

By Lemma 9.2 the bias term I satisfies

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-2js} \sim \max\{\delta, \varepsilon\}^{4s/(2s+2t+d)}$$

and has the right order. Concerning the third term III, we have

$$\mathbb{P}(\|K_{\delta,j}^{-1}\|_{V_j \to V_j} > \tau 2^{jt}) \leqslant \mathbb{P}(\{\|K_{\delta,j}^{-1}\|_{V_j \to V_j} > \tau 2^{jt}\} \cap \Omega_{\rho,\delta,j}) + \mathbb{P}(\Omega_{\rho,\delta,j}^c).$$

The second term of the right-hand side is asymptotically negligible by Lemma 11.2. For the first term we use that on $\Omega_{\rho,\delta,j}$ the operator $K_{\delta,j} = K_j(\mathrm{Id} + \delta K_j^{-1} \dot{B})$ is invertible with $\|K_{\delta,J}^{-1}\|_{V_j \to V_j} \leqslant (1-\rho)^{-1} \|K_j^{-1}\|_{V_j \to V_j}$. The restriction $c_K < \tau$ ensures that

$$\mathbb{P}(\{\|K_{\delta,j}^{-1}\|_{V_j \to V_j} > \tau 2^{jt}\} \cap \Omega_{\rho,\delta,j}) = 0,$$

provided $\rho \in (0, 1 - c_K/\tau)$, a choice we shall make from now on. We turn to the main term II. First, writing $P_j g_\varepsilon = P_j K f + \varepsilon P_j \dot{W}$, we have

$$\begin{aligned} &\mathbb{E}[\|K_{\delta,j}^{-1} P_j g_\varepsilon - f_j\|_{L^2}^2 \mathbf{1}_{\{\|K_{\delta,j}^{-1}\|_{V_j \to V_j} \leqslant \tau 2^{jt}\}} \mathbf{1}_{\Omega_{\rho,j,\delta}^c}] \\ &\lesssim 2^{2jt}(\|P_j K f\|_{L^2}^2 + \varepsilon^2\, \mathbb{E}[\|P_j \dot{W}\|_{L^2}^2] + \|f_j\|_{L^2}^2)\,\mathbb{P}(\Omega_{\rho,\delta,j}^c), \end{aligned}$$

where we used that the event $\Omega_{\rho,\delta,j}^c$ is independent of $P_j \dot{W}$ since $\dot{B}$ and $\dot{W}$ are independent. Using $\|P_j K f\|_{L^2}^2 + \|f_j\|_{L^2}^2 \lesssim M^2$ and $\mathbb{E}[\|P_j \dot{W}\|_{L^2}^2] \lesssim 2^{jd}$, we see by Lemma 11.2 that the above term is asymptotically negligible. Therefore, we are left with proving that $\mathbb{E}[\|K_{\delta,j}^{-1} P_j g_\varepsilon - f_j\|_{L^2}^2 \mathbf{1}_{\Omega_{\rho,j,\delta}}]$ has the right order. By the same Neumann series argument as in (11.2), we readily obtain on $\Omega_{\rho,j,\delta}$:

$$K_{\delta,j}^{-1} P_j g_\varepsilon - f_j = \sum_{n \geqslant 1} (-\delta K_j^{-1} \dot{B}_j)^n f_j + \varepsilon(\mathrm{Id} + \delta K_j^{-1} \dot{B}_j)^{-1} K_j^{-1} P_j \dot{W}. \qquad (11.3)$$

As for the second term in the right-hand side of (11.3), we have

$$\begin{aligned} &\mathbb{E}[\varepsilon^2 \|(\mathrm{Id} + \delta K_j^{-1} \dot{B}_j)^{-1} K_j^{-1} P_j \dot{W}\|_{L^2}^2 \mathbf{1}_{\Omega_{\rho,\delta,j}}] \\ &\leqslant \varepsilon^2\, \mathbb{E}[\|(\mathrm{Id} + \delta K_j^{-1} \dot{B}_j)^{-1}\|_{V_j \to V_j}^2 \mathbf{1}_{\Omega_{\rho,\delta,j}}] \|K_j^{-1}\|_{V_j \to V_j}^2\, \mathbb{E}[\|P_j \dot{W}\|_{L^2}^2] \\ &\lesssim \varepsilon^2 2^{jt} 2^{dj} \sim \max\{\delta, \varepsilon\}^{4s/(2s+2t+d)}, \end{aligned}$$

where we used again the independence of $\Omega_{\rho,\delta,j}$ and $P_j\dot{W}$, Lemma 11.1 and (11.2). The first term in the right-hand side of (11.3) is treated by

$$\begin{aligned}
&\mathbb{E}\,[\|\sum_{n\geqslant 1}(-\delta K_j^{-1}\dot{B}_j)^n f_j\|_{L^2}^2 \mathbf{1}_{\Omega_{\rho,\delta,j}}] \\
&= \mathbb{E}[\|\delta K_j^{-1}\dot{B}_j(\mathrm{Id}+\delta K_j^{-1}\dot{B}_j)^{-1}f_j\|_{L^2}^2\mathbf{1}_{\Omega_{\rho,\delta,j}}] \\
&\leqslant \delta^2\|K_j^{-1}\|_{V_j\to V_j}^2\,\mathbb{E}[\|\dot{B}_j\|_{V_j\to V_j}^2\|(\mathrm{Id}+\delta K_j^{-1}\dot{B}_j)^{-1}\|_{V_j\to V_j}^2\mathbf{1}_{\Omega_{\rho,\delta,j}}] \\
&\lesssim \delta^2\|K_j^{-1}\|_{V_j\to V_j}^2\,\mathbb{E}[\|\dot{B}_j\|_{V_j\to V_j}^2] \\
&\lesssim \delta^2 2^{2jt}2^{dj} \leqslant \max\{\delta,\varepsilon\}^{4s/(2s+2t+d)},
\end{aligned}$$

where we successively used Lemma 9.2, estimate (11.2) and Lemmas 11.1, 9.1.

11.2 Proof of Theorem 7.2

To avoid singularity of the underlying probability measures we only consider the subclass $\mathcal{F}_0$ of parameters (K,f) such that $Kf = y_0$ for some fixed $y_0 \in L^2$, i.e. $\mathcal{F}_0 := \{f \in L^2 \,|\, K^{-1}y_0,\ K \in \mathcal{K}\}$, where $\mathcal{K} = \mathcal{K}_t(C)$ abbreviates the class of operators under consideration. We shall henceforth keep y_0 fixed and refer to the parameter (K,f) equivalently just by K.

The likelihood $\Lambda(\bullet)$ of $\mathbb{P}^{K^2}$ under the law $\mathbb{P}^{K^1}$ corresponding to the parameters K^i, $i = 1, 2$, is

$$\Lambda(K^2,K^1) = \exp\big(\delta^{-1}\langle K^2 - K^1, \dot{B}\rangle_{HS} - \tfrac{1}{2}\delta^{-2}\|K^1 - K^2\|_{HS}^2\big)$$

in terms of the scalar product and norm of the Hilbert space $HS(L^2)$ of Hilbert-Schmidt operators on L^2 and with a Gaussian white noise operator $\dot{B}$. In particular, the Kullback-Leibler divergence between the two measures equals $\frac{1}{2}\delta^{-2}\|K^1 - K^2\|_{HS}^2$ and the two models remain contiguous for $\delta \to 0$ as long as the Hilbert-Schmidt norm of the difference remains of order δ.

Let us fix the parameter $f_0 = \psi_{-1,0} = \mathbf{1}$ and the operator K^0 which, in a wavelet basis $(\psi_\lambda)_\lambda$, has diagonal form $K^0 = \mathrm{diag}(2^{-(|\lambda|+1)t})$. Then K^0 is ill-posed of degree t and trivially obeys all the mapping properties imposed for the upper bound. Henceforth, $y_0 := K^0 f_0 = \mathbf{1}$ remains fixed.

For any $k = 0, \ldots, 2^{Jd} - 1$ introduce the symmetric perturbation $H^\varepsilon = (H^\varepsilon_{\lambda,\mu})$ with vanishing coefficients except for $H^\varepsilon_{(0,0),(J,k)} = 1$ and $H^\varepsilon_{(J,k),(0,0)} = 1$. Put $K^\varepsilon = K^0 + \gamma H^\varepsilon$ for some $\gamma > 0$. By setting $\gamma := \delta J$ we enforce $\|K^\varepsilon - K^0\|_{HS} = \delta J$. For $f_\varepsilon := (K^\varepsilon)^{-1}y_0$, we obtain

$$f_\varepsilon - f_0 = \big((K^\varepsilon)^{-1} - (K^0)^{-1}\big)y_0 = \gamma(K^\varepsilon)^{-1}H^\varepsilon f_0 = \gamma(K^\varepsilon)^{-1}\psi_{J,0}.$$

Now observe that H^ε trivially satisfies the conditions

$$\begin{aligned}
&\tfrac{1}{2}|\langle H^\varepsilon f, f\rangle| \leqslant 2^{Jt}\|f\|_{H^{-t/2}}^2, \quad \tfrac{1}{2}\|H^\varepsilon\|_{L^2\to H^t} \leqslant 2^{Jt}, \\
&\tfrac{1}{2}\|H^\varepsilon\|_{B^s_{p,p}\to B^{s+t}_{p,p}} \leqslant 2^{J(t+s+\frac{d}{2}-\frac{d}{p})}.
\end{aligned}$$

This implies that for $\gamma 2^{J(t+s+\frac{d}{2}-\frac{d}{p})}$ sufficiently small K^{ε} inherits the mapping properties from K^0. Hence

$$\|f_\varepsilon - f_0\|_{L^2} \sim \gamma \|\psi_{J,0}\|_{H^t} = \gamma 2^{Jt},$$
$$\|f_\varepsilon - f_0\|_{B^s_{p,p}} \sim \gamma \|\psi_{J,0}\|_{B^{s+t}_{p,p}} = \gamma 2^{J(t+s+\frac{d}{2}-\frac{d}{p})}$$

follows. In order to apply the classical lower bound proof in the sparse case [24, Thm. 2.5.3] and thus to obtain the logarithmic correction, we nevertheless have to show that $f_\varepsilon - f_0$ is well localized. Using the fact that $\big((H^\varepsilon)^2\big)_{\lambda,\mu} = 1$ holds for coordinates $\lambda = \mu = (0,0)$ and $\lambda = \mu = (J,k)$, but vanishes elsewhere, we infer from the Neumann series representation

$$f_\varepsilon - f_0 = \sum_{m=1}^{\infty}(-\gamma H^\varepsilon)^m f_0 = \sum_{n=1}^{\infty}\gamma^{2n} f_0 - \sum_{n=0}^{\infty}\gamma^{2n+1}\psi_{J,k} = \frac{\gamma}{1-\gamma^2}(\gamma f_0 - \psi_{J,k}).$$

Consequently, the asymptotics for $\gamma \to 0$ are governed by the term $-\gamma\psi_{J,k}$, which is well localized. The choice $2^J < \gamma^{-1/(t+s+\frac{d}{2}-\frac{d}{p})}$ ensures that $\|f_\varepsilon\|_{B^s_{p,p}}$ remains bounded and we conclude by usual arguments, cf. Chapter 2 in [24] or the lower bound in [30].

11.3 Some tools from approximation theory

The material gathered here is classical, see e.g. [8]. An equivalent norming of the Besov space $B^s_{p,p}$ for all $s \in \mathbb{R}$ and $p > 0$ is given in terms of weighted wavelet coefficients, if the wavelet basis is $(\lfloor |s| \rfloor + 1)$-regular:

$$\|f\|_{B^s_{p,p}} \sim \Big(\sum_{j=-1}^{\infty} 2^{j(s+\frac{1}{2}-\frac{1}{p})p}\sum_k |\langle f, \psi_{jk}\rangle|^p\Big)^{1/p}.$$

Here, $k \in \mathbb{Z}^d$ is the location parameter and j the resolution level of the wavelet. For $p < 1$ the Besov spaces are only quasi-Banach spaces, but still coincide with the corresponding nonlinear approximation spaces, see Section 30 in [8]. If s is not an integer or if $p = 2$, the space $B^s_{p,p}$ equals the L^p-Sobolev space $W^{s,p}$, which for $p = 2$ is denoted by H^s. The Sobolev embedding generalizes to

$$B^s_{p,p} \subset B^{s'}_{p',p'} \text{ for } s \geqslant s' \text{ and } s - \tfrac{d}{p} \geqslant s' - \tfrac{d}{p'}.$$

Direct and inverse estimates are the main tools in approximation theory. In its simplest form, a direct inequality reads

$$\inf_{h_j \in V_j} \|f - h_j\|_{L^2} \lesssim 2^{-sj}|f|_{H^s},$$

and the inverse estimate states that for all $h_j \in V_j$

$$|h_j|_{H^s} \lesssim 2^{sj}\|h_j\|_{L^2}.$$

The direct and inverse estimate we use in the paper are less standard since they involve the Sobolev space of negative order $H^{-t/2}$. The inverse estimate states that for all $g_j \in V_j$

$$\|g_j\|_{L^2} \lesssim 2^{tj/2}\|g_j\|_{H^{-t/2}},$$

see [10]. The direct estimate states that

$$\inf_{h_j \in V_j} [\|f - h_j\|_{L^2} + 2^{tj/2} \|f - h_j\|_{H^{-t/2}}] \lesssim 2^{-sj} \|f\|_{H^s}.$$

References

[1] Abramovich, F. and Silverman, B.W. (1998): Wavelet decomposition approaches to statistical inverse problems, *Biometrika* **85**, 115–129.

[2] Antoniadis, A. and Bigot, J (2004): Poisson inverse problems. Preprint.

[3] Bennet, G., V. Goodman and Newman, C.M. (1975): Norms of random matrices. *Pacific Journal of Math* **59**, 359–365.

[4] Cai, T. and Hall, P. (2004): Prediction in functional linear regression. Available under `http://stat.wharton.upenn.edu/~tcai/research.html`.

[5] Cavalier, L., Golubev, G.K., Picard, D. and Tsybakov, A.B. (2002): Oracle inequalities for inverse problems. Dedicated to the memory of Lucien Le Cam. *Ann. Statist.* **30**, 843–874.

[6] Cavalier, L. and Hengartner, N.W. (2003): Adaptive estimation for inverse problems with noisy operators. *Preprint.*

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

[8] Cohen, A. (2000): *Wavelets in numerical analysis*, Handbook of Numerical Analysis, vol. VII, P.G. Ciarlet and J.L. Lions, eds., Elsevier, Amsterdam.

[9] Cohen, A., DeVore R. and Hochmuth R. (2000): Restricted nonlinear approximation. *Constr. Approx.* **16**, 85–113.

[10] Cohen, A., Hoffmann, M. and Reiß, M. (2004): Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.*, to appear.

[11] Dahmen, W., Harbrecht, H. and Schneider, R. (2002): Compression techniques for boundary integral equations — optimal complexity estimates. IGPM Report No. 218, RWTH Aachen.

[12] Davidson, K.R. and Szarek, S.J. (2001): Local operator theory, random matrices and Banach spaces. In *Handbook on the Geometry of Banach spaces,* Vol. 1, W. B. Johnson, J. Lindenstrauss eds., Elsevier, 317–366.

[13] Dicken V. and Maass P. (1996): Wavelet-Galerkin methods for ill-posed problems. *J. Inverse Ill-Posed Probl.* **4**, 203–221.

[14] Donoho, D. (1995): Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2**, 101–126.

[15] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996): Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508–539.

[16] Efromovich, S. and Koltchinskii, V. (2001): On inverse problems with unknown operators. *IEEE Trans. Inf. Theory* **47**, 2876–2894.

[17] Engl, H.W, Hanke, M. and Neubauer, A. (2000): *Regularization of inverse problems*, Kluwer Academic Press.

[18] Florens, J.P. (2003): Inverse problems and structural econometrics: the problem of instrumental variables. In *Advances in Economics and Econometrics – Theory and Applications*, Dewatripont, Hansen and Turnovski, eds., vol. 2, Cambridge University Press, 284–311.

[19] Gobet, E., Hoffmann, M and Reiß, M. (2004): Nonparametric estimation of scalar diffusions from low frequency data. *Ann. Statist.* **32**, 2223–2253.

[20] Goldenshluger, A. and Pereverzev, S.V. (2003): On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli* **9**, 783–807.

[21] Hall, P. and Horowitz, J.L. (2003): Nonparametric methods in the presence of instrumental variables. Preprint available under `http://www.faculty.econ.northwestern.edu/faculty/horowitz`.

[22] Johnstone, I.M., Kerkyacharian, G., Picard, D. and Raimondo, M. (2004): Wavelet deconvolution in a periodic setting. *J. Royal Statistical Society*, Ser B, **66**, 547–573.

[23] Johnstone, I.M. and Silverman, B.W. (1990): Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251–280.

[24] Korostelev, A. and Tsybakov, A.B. (1993): *Minimax theory of image reconstruction*, Lecture Notes in Statistics **82**, Springer, New York.

[25] Mathé, P. and Pereverzev, S.V. (2003): Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse problems* **19**, 789–803.

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

[27] Nußbaum, M. and Pereverzev, S.V. (1999): The degree of ill-posedness in stochastic and deterministic models. Preprint No. 509, Weierstraß-Institut Berlin.

[28] Pruessner, A. and O'Leary, D. (2003): Blind deconvolution using a regularized structured total least norm algorithm. *SIAM J. Matrix Anal. Appl.* **24**, 1018–1037.

[29] Ramsay, J.O. and Silverman, B.W. (1997): Functional data analysis, Springer Series in Statistics, New York.

[30] Reiß, M. (2004): Adaptive estimation for affine stochastic delay differential equations, *Bernoulli*, to appear.

[31] Tsybakov, A.B. (2000): On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris Sér. I Math.* **330**, 835–840.

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