

Adaptive wavelet Galerkin methods for linear inverse problems

Albert Cohen* Marc Hoffmann† Markus Reiß‡§

May 14, 2002

Abstract

We introduce and analyse numerical methods for the treatment of inverse problems, based on an adaptive wavelet Galerkin discretization. These methods combine the theoretical advantages of the wavelet-vaguelette decomposition (WVD) in terms of optimally adapting to the unknown smoothness of the solution, together with the numerical simplicity of Galerkin methods. Two strategies are proposed: the first one simply combines a thresholding algorithm on the data with a Galerkin inversion on a fixed linear space, while the second one performs the inversion through an adaptive procedure in which a smaller space adapted to the solution is iteratively constructed. For both methods, we recover the same minimax rates achieved by WVD for various function classes modeling the solution.

1. Introduction

Let $K : X \rightarrow Y$ be a compact linear operator between Hilbert spaces X and Y . The classical ill-posed inverse problem consists of recovering a good

*Laboratoire d'analyse numérique, Université Paris VI

†Laboratoire de probabilités et modèles aléatoires, Université Paris VII

‡Institut für Mathematik, Humboldt-Universität Berlin

§Support by the EU research network Dynstoch and the DFG-Sonderforschungsbereich 373 is gratefully acknowledged.

approximation f_ε for the solution $f \in X$ of the equation

$$Kf = g \tag{1}$$

when only a perturbation g_ε of $g \in Y$ is known. In order to assess the quality of a method, the convergence of f_ε to f is examined when g_ε tends to g in a deterministic or stochastic sense. In this paper, we shall suppose that g_ε is a perturbation of g by an additive white noise of level ε . We are then interested in the behaviour of the mean-square error $E(\|f_\varepsilon - f\|^2)$ as ε tends to zero.

The *singular value decomposition* (SVD) of K

$$Kf = \sum_{k=0}^{\infty} s_k \langle f, \tilde{f}_k \rangle \tilde{g}_k, \tag{2}$$

where $(\tilde{f}_k)_{k \geq 0}$ and $(\tilde{g}_k)_{k \geq 0}$ are orthonormal bases of X and Y , and the singular values $(s_k)_{k \geq 0}$ are arranged in decreasing order of moduli, gives rise to the approximation

$$f_\varepsilon := \sum_{k=0}^N s_k^{-1} \langle g_\varepsilon, \tilde{g}_k \rangle \tilde{f}_k, \tag{3}$$

where $N = N(\varepsilon)$ has to be chosen properly. This SVD (or spectral cut-off) method is theoretically very attractive and can be shown to be asymptotically optimal in many cases, e.g. [16] and [14]. However, it suffers from two types of limitations. First of all, the bases $(\tilde{f}_k)_{k \geq 0}$ and $(\tilde{g}_k)_{k \geq 0}$ might be difficult to determine and to manipulate numerically. Secondly, while these bases are fully adapted to describe the action of K , they might not be so appropriate for the accurate description of the solution with a small number of parameters.

Concerning the first limitation, i.e. numerical simplicity, *projection methods* are more appealing. Given finite dimensional subspaces $X_h \subset X$ and $Y_h \subset Y$ such that $\dim(X_h) = \dim(Y_h)$, one defines the approximation f_ε as the solution in X_h of the problem

$$\langle Kf_\varepsilon, g_h \rangle = \langle g_\varepsilon, g_h \rangle, \quad \text{for all } g_h \in Y_h. \tag{4}$$

If $(u_k)_{k=0, \dots, N}$ and $(v_k)_{k=0, \dots, N}$ are bases of X_h and Y_h , this amounts to solving the $(N+1) \times (N+1)$ -dimensional linear system

$$K_h F_\varepsilon = G_\varepsilon, \tag{5}$$

where $K_h := (\langle Ku_i, v_j \rangle)_{i,j=0,\dots,N}$ and $G_\varepsilon := (\langle g_\varepsilon, v_j \rangle)_{j=0,\dots,N}$, and to define $f_\varepsilon := \sum_{i=0}^N F_{\varepsilon,i} u_i$. A general approach to projection methods was developed in [15].

In the case where $X = Y$ and K is a selfadjoint positive definite operator, we choose $Y_h = X_h$ and the linear system (5) is particularly simple to solve since K_h is symmetric positive definite. This is the so-called Galerkin method. In the case of general $X \neq Y$ and injective K , one may choose $Y_h := K(X_h)$ which amounts to solving (5) with the matrix $K_h := (\langle K^* K u_i, v_j \rangle)_{i,j=1,\dots,N}$ and the right hand side $G_\varepsilon := (\langle K^* g_\varepsilon, v_j \rangle)_{j=0,\dots,N}$, where K^* denotes the adjoint of K . Hence, we are led back to the Galerkin method applied to the least-square equation $K^* K f = K^* g$ with data $K^* g_\varepsilon$, which is equivalent to finding the minimum of the functional $\|Kf - g_\varepsilon\|^2$ over X_h . Note that minimizing over the restricted set X_h has an effect similar to *Tychonov regularization*, i.e. adding a penalization term (typically $\lambda \|\nabla f\|^2$) in the least square functional and minimizing over the whole space X . Further, note that the least-square method reduces to the SVD-method if the basis $(u_k)_{k=0,\dots,N}$ of X_h coincides with the $N + 1$ first vectors of the basis $(f_k)_{k \geq 0}$ in the SVD decomposition of K . We also refer to [12] for a detailed introduction to regularization methods for inverse problems.

The numerical simplicity of projection methods comes from the fact that X_h and Y_h are typically finite element spaces equipped with standard local bases. As in the SVD-method, the discretization parameter h has to be properly chosen. The choice of finite element spaces for X_h and Y_h is also beneficial with respect to the second limitation of SVD, since the approximation properties of finite elements can be exploited when the solution has a certain smoothness. To be more specific, assume that $X = Y = L^2(\Omega)$ where $\Omega \subset \mathbb{R}^d$ is a domain, and that K is self-adjoint and acts along the scale of Sobolev spaces H^s with a degree of ill-posedness $t > 0$ (i.e. maps H^s onto H^{s+t}). Then, under the assumption that the solution f belongs to some Sobolev ball $B := \{f \in H^s ; \|f\|_{H^s} \leq M\}$, it is known that the Galerkin method using finite element spaces X_h with the proper choice $h(\varepsilon) := \varepsilon^{\frac{2}{2s+2t+d}}$ attains the rate of convergence

$$\sup_{f \in B} E(\|f_\varepsilon - f\|_{L^2}^2) \leq C \varepsilon^{\frac{4s}{2s+2t+d}}, \quad (6)$$

which is optimal over all estimators (see e.g. [9] for the upper bound and [13] for the lower bound).

The above Galerkin projection method suffers from two drawbacks, which are encountered in all *linear* estimation methods. First of all, the choice of h with respect to the noise level ε depends on the regularity s of the solution which is almost always unknown in advance. Secondly, the use of a finite element space X_h imposes a uniform mesh size h and therefore does not allow any spatial adaptation. In recent years, *nonlinear* methods have been developed, with the objective of automatically adapting to unknown smoothness and local singular behaviour of the solution. In the case of simple denoising, i.e. when K is the identity, wavelet thresholding is probably one of the most attractive nonlinear method, since it is both numerically straightforward and asymptotically optimal for a large variety of Sobolev or Besov classes as models for the unknown smoothness of the solution, see e.g. [11]. This success strongly exploits the fact that wavelets provide unconditional bases for such smoothness spaces. In order to adapt this approach to the framework of ill-posed inverse problems, Donoho introduced in [10] a wavelet-like decomposition which is specifically adapted to describe the action of K , the so-called *wavelet-vaguelette decomposition* (WVD), and proposed to apply a thresholding algorithm on this decomposition. In [1], this method was compared with the similar *vaguelette-wavelet decomposition* (VWD) algorithm. Both methods rely on an orthogonal wavelet basis (ψ_λ) and associated Riesz bases of “vaguelettes” defined as

$$v_\lambda = \beta_\lambda K^{-1} \psi_\lambda \quad \text{and} \quad u_\lambda = \beta_\lambda (K^*)^{-1} \psi_\lambda, \quad (7)$$

where the scaling coefficients β_λ typically depend on the order of ill-posedness of K . We thus have the WVD and VWD decompositions

$$f = \sum_\lambda \beta_\lambda^{-1} \langle Kf, u_\lambda \rangle \psi_\lambda = \sum_\lambda \beta_\lambda^{-1} \langle Kf, \psi_\lambda \rangle v_\lambda. \quad (8)$$

In the WVD method, for a white noise model, one estimates the coefficients $\langle Kf, u_\lambda \rangle$ by applying a thresholding operator $T_\varepsilon : x \mapsto x \chi(|x| \geq t(\varepsilon))$ (where $\chi(P)$ stands for the boolean function which is 1 if P holds and 0 otherwise) to the observed coefficients $g_\lambda^\varepsilon := \langle g_\varepsilon, u_\lambda \rangle$. The estimator is thus given by

$$f_\varepsilon := \sum_\lambda \beta_\lambda^{-1} T_\varepsilon(g_\lambda^\varepsilon) \psi_\lambda. \quad (9)$$

In the VWD method, the same procedure is used, but now with $g_\lambda^\varepsilon := \langle g_\varepsilon, \psi_\lambda \rangle$

and obtains the estimator

$$f_\varepsilon := \sum_{\lambda} \beta_{\lambda}^{-1} T_{\varepsilon}(g_{\lambda}^{\varepsilon}) v_{\lambda}. \quad (10)$$

Note that the latter simply amounts to denoising the data by wavelet thresholding, followed by an exact application K^{-1} . Observe also that this is equivalent to applying a least-square method with X_h spanned by the vaguelettes v_{λ} (then K_h is diagonal) to such denoised data.

On a theoretical level, similarly to wavelet thresholding in the case of simple denoising, both WVD and VWD methods allow to recover the same rate $\varepsilon^{\frac{4s}{2s+2t+d}}$ (up to a logarithmic factor) as for the projection method, under weaker smoothness conditions which reflect their ability for spatial adaptivity. More precisely, this rate is obtained if the model for the unknown function is a Besov ball $B := \{f \in B_{p,p}^s ; \|f\|_{B_{p,p}^s} \leq M\}$ with $1/p = 1/2 + s/(2t + d)$. On a more applied level, numerical implementations in [1] and [9] have illustrated the efficiency of both WVD and VWD methods, in the case of the integration operator $Kf(x) = \int_0^x f(t)dt$. For more general operators, however, the assumption that $K^{-1}\psi_{\lambda}$ or $(K^*)^{-1}\psi_{\lambda}$ are known for all indices λ only transfers the inversion problem. If an integral operator has a kernel with a complicated structure, or if this kernel is itself derived from observations (see [17]), this inversion has to be done numerically with additional computational cost and error. In other words the vaguelettes u_{λ} and v_{λ} might be difficult to handle numerically (similar to the SVD functions), and in particular they are not ensured to have compact support. In this context, a natural goal is to build a method which combines the numerical simplicity of linear Galerkin projection methods with the optimality of adaptive wavelet thresholding methods.

Note that adaptive Galerkin methods are well established in the context of solving operator equation *without noise*: typically, the finite element space is locally refined based on *a-posteriori error analysis* of the current numerical solution. Such adaptive algorithms were recently extended to the context of wavelet discretizations, exploiting both the characterization of function spaces and the sparse representation of the operator by wavelet bases, see e.g. [4]. Our goal in this paper is to introduce and analyse similar adaptive wavelet Galerkin algorithms in the context of statistical inverse problems. Such adaptive algorithms only involve the wavelet system (ψ_{λ}) and are therefore easier to handle numerically than WVD and VWD in general

situations. On the other hand, their optimality will essentially rely on the assumption that K has certain mapping properties with respect to the relevant function spaces, a fact which is also implicitly used in the WVD and VWD approaches. Last but not least, we shall exploit in addition the fact that the Galerkin discretization of K in the wavelet basis might be sparse even for nonlocal integral operators, in order to improve the computational efficiency of our algorithm.

Our paper is organized as follows. In §2, we introduce some general assumptions on the models which are used in this paper, both in terms of the mapping properties of the operator K between smoothness spaces, and of wavelet characterization of these spaces. In §3, we briefly recall the analysis of the linear Galerkin method, using the wavelet multiresolution spaces V_j . A first nonlinear method is proposed in §4, which initially operates in a similar way as VWD, by thresholding the wavelet coefficients $g_\lambda^\varepsilon := \langle g_\varepsilon, \psi_\lambda \rangle$ with λ restricted up to a maximal scale level $j = j(\varepsilon)$, and then applies a linear Galerkin inversion of these denoised data on the multiresolution space V_j . The dimension of V_j behaves like $\varepsilon^{-d/t}$, so that this Galerkin approximation can become computationally heavy, while the solution could still be well represented by a small adaptive set of wavelets within V_j . Therefore, we propose in §5 an adaptive algorithm which iteratively produces such a set together with the corresponding Galerkin estimator. This algorithm intertwines the process of thresholding with an iterative resolution of the Galerkin system, and it exploits in addition the sparse representation of K in the wavelet basis. Both methods are proved to achieve the same minimax rate as WVD and VWD, under the same general assumptions on the operator K . For the sake of simplicity, we present our methods and results in the case where K is elliptic. The generalization of these methods and results to non-elliptic operators, via a least square approach, is feasible and the objective of a forthcoming paper, where we shall also compare the different estimators on the example of a singular integral equation of the first kind.

2. Models and notations

We are interested in the statistical formulation of linear inverse problems: we observe the action of a compact operator $K : L^2(\Omega_X) \rightarrow L^2(\Omega_Y)$ on an

unknown function $f \in L^2(\Omega_X)$ polluted by an additive white gaussian noise of variance ε^2 , i.e.

$$g_\varepsilon = Kf + \varepsilon dW. \quad (11)$$

The unknown and observation domains Ω_X and Ω_Y have dimension d and might differ from one another. In the following when mentioning L^2 or more general function spaces, we shall omit to specify the domain Ω_X or Ω_Y when this information is obvious from the context. Note that observing g_ε does not make sense pointwise but in a measure sense: for any $v \in L^2$, we observe

$$\langle g_\varepsilon, v \rangle = \langle Kf, v \rangle + \varepsilon \|v\|_{L^2} \eta_v, \quad (12)$$

where η_v is a standard normal variable with covariance $E(\eta_v \eta_{v'}) = 0$ for $\langle v, v' \rangle = 0$.

2.1. Assumptions on the operator K

The ill-posed nature of the problem comes from the assumption that K is compact and therefore its inverse is not L^2 -bounded. This is generally expressed by a smoothing action: K typically maps L^2 into H^t for some $t > 0$. More generally we say that K has the smoothing property of order t with respect to some smoothness space H^s (resp. $W_p^s, B_{p,q}^s$) if this space is mapped onto H^{s+t} (resp. $W_p^{s+t}, B_{p,q}^{s+t}$).

In order to avoid a blow up of the noise, the estimator f_ε will be searched within a finite dimensional subspace V of $L^2(\Omega_X)$ based on the projection method which was recalled in the introduction. In the case where $\Omega_X = \Omega_Y = \Omega$ and K is self-adjoint positive definite, we shall use the Galerkin method, i.e.

$$\text{find } f_\varepsilon \in V \text{ such that } \langle Kf_\varepsilon, v \rangle = \langle g_\varepsilon, v \rangle \text{ for all } v \in V. \quad (13)$$

The smoothing property of order t will then be expressed by the ellipticity property

$$\langle Kf, f \rangle \sim \|f\|_{H^{-t/2}}^2 \quad (14)$$

where $H^{-t/2}$ stands for the dual space of the Sobolev space $H^{t/2}$ appended with boundary conditions that might vary depending on the considered problem (homogeneous, periodic...).

In the case where $\Omega_X \neq \Omega_Y$ or when K is not self-adjoint positive definite, one can consider the least square method, i.e.

$$\text{find } f_\varepsilon \in V \text{ which minimizes } \|Kv - g_\varepsilon\|_{L^2}^2 \text{ among all } v \in V. \quad (15)$$

As already remarked, this amounts to applying the Galerkin method on the equation $K^*Kf = K^*g$ with data K^*g_ε and trial space $K(V)$. The smoothing property of order t will then be expressed by the ellipticity property

$$\|Kf\|_{L^2}^2 = \langle K^*Kf, f \rangle \sim \|f\|_{H^{-t}}^2. \quad (16)$$

We shall not deal with this general situation here, and we therefore assume for the next sections that K is self-adjoint positive definite and verifies (14). As already explained, our goal is to propose an adaptive choice of V through the selection of appropriate basis functions within a standard wavelet basis. A key ingredient in the analysis of our methods will be the fact that such bases allow the characterization of the function spaces which describe both the smoothness of the solution and the smoothing action of the operator. We now recall these properties.

2.2. Wavelet bases

Wavelet bases have been documented in numerous textbooks and survey papers (see [7] for a general treatment). With a little effort they can be adapted to fairly general domains $\Omega \subset \mathbb{R}^d$ (see [3] for a survey of these adaptations as well as a discussion of the characterizations of function spaces on Ω by wavelet coefficients).

A wavelet basis consists of two types of functions: scaling functions φ_λ and wavelet functions ψ_λ . The index λ concatenates the usual scale and space parameters j and k . Thus for standard wavelet bases on \mathbb{R} , we simply have $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$ (and similarly for φ_λ). However, the notation ψ_λ takes into account the possible adaptations of wavelets to multivariate bounded domains in which case the functions ψ_λ and φ_λ usually change form near the boundary. At the scale level j which corresponds to resolution 2^{-j} , the scaling functions $(\varphi_\lambda)_{\lambda \in \Gamma_j}$ span a space V_j within a hierarchy $V_0 \subset V_1 \subset \dots \subset L^2$ of nested approximation spaces. The space V_j should be thought of as a finite element space V_h with uniform mesh size $h \sim 2^{-j}$, and local

basis $(\varphi_\lambda)_{\lambda \in \Gamma_j}$. The wavelets $(\psi_\lambda)_{\lambda \in \nabla_j}$ span a complement W_j of V_j into V_{j+1} . With these notations, the wavelet decomposition takes the form

$$f = \sum_{\lambda \in \Gamma_{j_0}} \alpha_\lambda \varphi_\lambda + \sum_{j \geq j_0} \sum_{\lambda \in \nabla_j} f_\lambda \psi_\lambda, \quad (17)$$

where $(\varphi_\lambda)_{\lambda \in \Gamma_j}$ is the scaling function basis spanning the approximation at level j , $(\psi_\lambda)_{\lambda \in \nabla_j}$ is the wavelet basis spanning the details at level j , and $\alpha = \alpha(f)$ and $f_\lambda = f_\lambda(f)$ are the scaling function and wavelet coefficients of f , respectively. In what follows, we shall (merely for notational convenience) always take $j_0 := 0$.

The approximation and detail coefficients of f are linear functionals of f which can be evaluated according to

$$\alpha_\lambda = \langle f, \tilde{\varphi}_\lambda \rangle \quad \text{and} \quad f_\lambda = \langle f, \tilde{\psi}_\lambda \rangle, \quad (18)$$

where $\tilde{\varphi}_\lambda$ and $\tilde{\psi}_\lambda$ are the corresponding dual scaling functions and wavelets. In the orthonormal case, which we do not impose here, these are the same as the primal scaling functions and wavelets φ_λ and ψ_λ . We also use the notation $|\lambda| = j$ if $\lambda \in \Gamma_j$ or $\lambda \in \nabla_j$. Finally, to simplify notation even more, we incorporate the first layer of scaling functions $(\varphi_\lambda)_{\lambda \in \Gamma_0}$ into the wavelet layer $(\psi_\lambda)_{\lambda \in \nabla_0}$ and define $\nabla = \cup_{j \geq 0} \nabla_j$, so that we simply have

$$f = \sum_{\lambda \in \nabla} f_\lambda \psi_\lambda = \sum_{j=0}^{\infty} \sum_{|\lambda|=j} f_\lambda \psi_\lambda. \quad (19)$$

If we truncate this expansion at level j , we obtain the oblique projector onto the space V_j

$$P_j f = \sum_{|\lambda|=j} \alpha_\lambda \varphi_\lambda = \sum_{|\lambda| < j} f_\lambda \psi_\lambda. \quad (20)$$

It is well known (see e.g. [3]) that wavelet bases provide characterizations of smoothness spaces such as the Hölder spaces C^s , Sobolev spaces W_p^s and Besov spaces $B_{p,q}^s$ for a range of indices s that depend both on the smoothness properties of ψ and $\tilde{\psi}$. For the scale of Besov spaces (which includes as particular cases $C^s = B_{\infty,\infty}^s$ and $W_p^s = B_{p,p}^s$ if $s \notin \mathbb{N}$ or if $p = 2$), the characterization result has the form

$$\|f\|_{B_{q,p}^s} \sim \|(2^{s|\lambda|} 2^{d|\lambda|(1/2-1/p)} \|(f_\lambda)_{|\lambda|=j}\|_{\ell_p})_{j \geq 0}\|_{\ell_q}, \quad (21)$$

where d is the space dimension ($\Omega \subset \mathbb{R}^d$).

3. Linear Galerkin estimation

We recall here some classical results on the linear Galerkin projection method. For some scale $j > 0$ to be fixed later, we define our first estimator $f_\varepsilon = \sum_{\gamma \in \Gamma_j} f_{\varepsilon, \gamma} \varphi_\gamma \in V_j$ as the unique solution of the finite dimensional linear problem

$$\text{find } f_\varepsilon \in V_j \text{ such that } \langle K f_\varepsilon, v \rangle = \langle g_\varepsilon, v \rangle \text{ for all } v \in V_j. \quad (22)$$

Defining the data vector $G_\varepsilon := (\langle g_\varepsilon, \varphi_\gamma \rangle)_{\gamma \in \Gamma_j}$, and the Galerkin stiffness matrix $K_j := (\langle K \varphi_\gamma, \varphi_\mu \rangle)_{\gamma, \mu \in \Gamma_j}$, the coordinate vector $F_\varepsilon := (f_{\varepsilon, \gamma})_{\gamma \in \Gamma_j}$ of f_ε is therefore the solution of the linear system

$$K_j F_\varepsilon = G_\varepsilon. \quad (23)$$

The analysis of this method can be done by decomposing f_ε according to

$$f_\varepsilon = f_j + h_\varepsilon, \quad (24)$$

where the terms f_j and h_ε are respectively solutions of (22) with $g = Kf$ and εdW in place of g_ε as the right hand side. This gives the classical decomposition of the estimation error into a bias and variance term

$$E(\|f - f_\varepsilon\|_{L^2}^2) = \|f - f_j\|_{L^2}^2 + E(\|h_\varepsilon\|_{L^2}^2). \quad (25)$$

Both terms are estimated with the help of some specific properties of the spaces V_j , namely inverse and direct estimates with respect to Sobolev spaces. The variance term can be estimated as follows: we first use the ellipticity property (14), which gives

$$\|h_\varepsilon\|_{H^{-t/2}}^2 \lesssim \langle K h_\varepsilon, h_\varepsilon \rangle = \varepsilon \langle dW, h_\varepsilon \rangle \leq \varepsilon \|P_j dW\|_{L^2} \|h_\varepsilon\|_{L^2}. \quad (26)$$

Using the Bernstein inequality which states that $\|g\|_{L^2} \lesssim 2^{tj/2} \|g\|_{H^{-t/2}}$ for all $g \in V_j$ and dividing by $\|h_\varepsilon\|_{L^2}$, we obtain

$$\|h_\varepsilon\|_{L^2} \lesssim \varepsilon 2^{tj} \|P_j dW\|_{L^2}, \quad (27)$$

and therefore

$$E(\|h_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^2 2^{2tj} \dim(V_j) \lesssim \varepsilon^2 2^{(2t+d)j}. \quad (28)$$

For the bias term, we take any $g_j \in V_j$ and write

$$\begin{aligned} \|f - f_j\|_{L^2} &\leq \|f - g_j\|_{L^2} + \|f_j - g_j\|_{L^2} \\ &\lesssim \|f - g_j\|_{L^2} + 2^{tj/2} \|f_j - g_j\|_{H^{-t/2}} \\ &\lesssim \|f - g_j\|_{L^2} + 2^{tj/2} \|f - g_j\|_{H^{-t/2}}, \end{aligned}$$

where we have again used the Bernstein inequality and the fact that the Galerkin projection satisfies $\|f - f_j\|_{H^{-t/2}} \lesssim \|f - g_j\|_{H^{-t/2}}$ for any $g_j \in V_j$. It follows that

$$\|f - f_j\|_{L^2} \lesssim \inf_{g_j \in V_j} [\|f - g_j\|_{L^2} + 2^{tj/2} \|f - g_j\|_{H^{-t/2}}]. \quad (29)$$

Assuming that f belongs to the Sobolev ball $B := \{f \in H^s ; \|f\|_{H^s} \leq M\}$, we obtain from approximation theory the estimate

$$\inf_{g_j \in V_j} [\|f - g_j\|_{L^2} + 2^{tj/2} \|f - g_j\|_{H^{-t/2}}] \lesssim 2^{-sj}, \quad (30)$$

and therefore

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-2sj}. \quad (31)$$

Equilibrating the bias and variance terms gives the optimal choice of resolution

$$2^{-j(\varepsilon)} \sim \varepsilon^{2/(2s+2t+d)}, \quad (32)$$

and the rate of convergence

$$E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^{4s/(2s+2t+d)}, \quad (33)$$

which is known to be minimax over the class B . As we already explained, the main defects of this linear method are its lack of adaptation to unknown smoothness ($j(\varepsilon)$ depends on s) and its lack of spatial adaptation in the presence of singularities.

4. Nonlinear estimation by linear Galerkin

Our first nonlinear estimator f_ε will simply consist of applying a thresholding algorithm on the data to which we apply the linear Galerkin inversion

which was described in the previous section: for some $j \geq 0$ to be fixed further, we define $f_\varepsilon = \sum_{|\lambda| < j} f_{\varepsilon, \lambda} \psi_\lambda \in V_j$ such that

$$\langle K f_\varepsilon, \psi_\lambda \rangle = T_\varepsilon(\langle g_\varepsilon, \psi_\lambda \rangle) \quad (34)$$

for all $|\lambda| < j$. In the thresholding operator $T_\varepsilon(x) = x\chi(|x| \geq t(\varepsilon))$, the threshold $t(\varepsilon)$ has the usual form

$$t(\varepsilon) = C\varepsilon\sqrt{|\log \varepsilon|}. \quad (35)$$

Note that such an estimator can be viewed as a variant of the vaguelette-wavelet estimator truncated at level j . Such an estimator would indeed be given (in the case where (ψ_λ) is an orthonormal basis) by

$$f_\varepsilon := \sum_{|\lambda| < j} T_\varepsilon(\langle g_\varepsilon, \psi_\lambda \rangle) K^{-1} \psi_\lambda. \quad (36)$$

The solution f_ε of (34) has a similar form with the vaguelettes $K^{-1}\psi_\lambda$ replaced by their Galerkin approximations $u_\lambda^j \in V_j$ such that

$$\langle K u_\lambda^j, v \rangle = \langle \psi_\lambda, v \rangle \quad (37)$$

for all $v \in V_j$. We therefore expect that this estimator behave in the same optimal way as the VWD estimator provided that j is large enough. The following theorem shows that this is indeed true if $2^{-j} \leq \varepsilon^{1/t}$ where t is the degree of ill-posedness of the estimator. It should be noted that the lower bound on j does not depend on the unknown smoothness of f .

Theorem 1. *Let us assume that f belongs to $B := \{f ; \|f\|_{B_{p,p}^s} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t + d)$. Also assume in addition that K is an isomorphism between L^2 and H^t and that it has the smoothing property of order t with respect to the space $B_{p,p}^s$. Then, the above described estimator satisfies the minimax rate*

$$E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \left[\varepsilon \sqrt{|\log \varepsilon|} \right]^{4s/(2s+2t+d)}, \quad (38)$$

provided that j is such that $2^{-j} \leq \varepsilon^{1/t}$.

Proof: We again write $f_\varepsilon^1 = f_j + h_\varepsilon$, where $f_j \in V_j$ is the solution of the linear problem with data g

$$\text{find } f_j \in V_j \text{ such that } \langle Kf_j, v \rangle = \langle g, v \rangle \text{ for all } v \in V_j. \quad (39)$$

Now the term h_ε represents the solution of the linear problem with the thresholding error as data, in other words $h_\varepsilon \in V_j$ such that

$$\langle Kh_\varepsilon, \psi_\lambda \rangle = T_\varepsilon(\langle g_\varepsilon, \psi_\lambda \rangle) - \langle g, \psi_\lambda \rangle, \quad (40)$$

for all $|\lambda| < j$. Similarly to the analysis described in the previous section, we need to estimate $\|f - f_j\|_{L^2}^2$ and $E(\|h_\varepsilon\|_{L^2}^2)$. For the deterministic term, we remark that the space $B_{p,p}^s$ is continuously imbedded into H^α whenever

$$\alpha \leq s + d/2 - d/p = 2ts/(2t + d). \quad (41)$$

By the same arguments as in the previous section, we therefore obtain the estimate

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-4sj \frac{t}{d+2t}} \quad (42)$$

This gives the optimal order $\varepsilon^{4s/(2s+2t+d)}$ if j is large enough so that

$$2^{-j} \leq \varepsilon^{\frac{d+2t}{t(2s+2t+d)}}. \quad (43)$$

For $\varepsilon < 1$, we have $\varepsilon^{1/t} \leq \varepsilon^{\frac{d+2t}{t(2s+2t+d)}}$ since $s \geq 0$, with equality if $s = 0$. Therefore, if $2^{-j} \leq \varepsilon^{1/t}$, we obtain

$$\|f - f_j\|_{L^2}^2 \lesssim \varepsilon^{4s/(2s+2t+d)}. \quad (44)$$

We next turn to the stochastic term $E(\|h_\varepsilon\|_{L^2}^2)$. If H_ε is the coordinate vector of h_ε in the basis $(\psi_\lambda)_{|\lambda| < j}$, we therefore want to estimate $E(\|H_\varepsilon\|_{\ell^2}^2)$. We can write

$$H_\varepsilon = K_j^{-1}(T_\varepsilon(G_\varepsilon) - G), \quad (45)$$

where $G_j := (\langle g, \psi_\lambda \rangle)_{|\lambda| < j}$ and $T_\varepsilon(G_\varepsilon) := (T_\varepsilon(\langle g_\varepsilon, \psi_\lambda \rangle))_{|\lambda| < j}$. We remark that $T_\varepsilon(G_\varepsilon) - G_j$ is exactly the error when estimating G_j by the thresholding procedure on the data G_ε . We shall take into account the action of K_j^{-1} by measuring this error in the wavelet version of the H^t norm

$$\|U\|_{h^t}^2 := \sum_{|\lambda| < j} 2^{2t|\lambda|} |u_\lambda|^2. \quad (46)$$

Indeed, we shall see that the stability property

$$\|K_j^{-1}U\|_{\ell^2} \lesssim \|U\|_{H^t}. \quad (47)$$

holds under the assumption that K^{-1} maps H^t onto L^2 . Our result will therefore follow from

$$E(\|T_\varepsilon(G_\varepsilon) - G_j\|_{h^t}^2) \lesssim \left[\varepsilon \sqrt{|\log \varepsilon|} \right]^{4s/(2s+2t+d)}. \quad (48)$$

Such a rate is a particular case of classical results on wavelet thresholding, using the fact that g belongs to a Besov ball $B = \{g \in B_{p,p}^{s+t}; \|g\|_{B_{p,p}^{s+t}} \leq M\}$. For this model, (48) follows e.g. from Theorem 4 in [6]. We are thus left with proving the stability property (47). To do so, we remark that if

$$K_j^{-1}U =: V = (v_\lambda)_{|\lambda| < j}, \quad (49)$$

then the function $v = \sum_{|\lambda| < j} v_\lambda \psi_\lambda$, is the Galerkin approximation of $K^{-1}u$, where u is the function defined by

$$U = (\langle u, \psi_\lambda \rangle)_{|\lambda| < j} \text{ and } \langle u, \psi_\lambda \rangle = 0 \text{ if } |\lambda| \geq j. \quad (50)$$

It follows that

$$\|K^{-1}u - v\|_{H^{-t/2}} \lesssim 2^{-jt/2} \|K^{-1}u\|_{L^2} \lesssim 2^{-jt/2} \|u\|_{H^t}. \quad (51)$$

For the projection $P_j K^{-1}u$, we also have the error estimate

$$\|K^{-1}u - P_j K^{-1}u\|_{H^{-t/2}} \lesssim 2^{-jt/2} \|K^{-1}u\|_{L^2} \lesssim 2^{-jt/2} \|u\|_{H^t}. \quad (52)$$

It follows that

$$\|v - P_j K^{-1}u\|_{H^{-t/2}} \lesssim 2^{-jt/2} \|u\|_{H^t}. \quad (53)$$

Using the Bernstein estimate, we obtain

$$\|v - P_j K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t}, \quad (54)$$

so that

$$\|v\|_{L^2} \lesssim \|u\|_{H^t} + \|P_j K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t} + \|K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t}. \quad (55)$$

Using the wavelet characterization of L^2 and H^t , this yields

$$\|V\|_{\ell^2} \lesssim \|U\|_{h^s}, \quad (56)$$

i.e. the stability estimate (47). \square

Remark: The assumption that K^{-1} maps H^t into L^2 which we are using in the above result is also implicit in the vaguelette-wavelet method when assuming that the vaguelettes

$$v_\lambda = \beta_\lambda K^{-1} \psi_\lambda = 2^{-t|\lambda|} K^{-1} \psi_\lambda, \quad (57)$$

constitute a Riesz basis of L^2 .

5. Nonlinear estimation by adaptive Galerkin

The main defect of the method that we have described in the previous section remains its computational cost: the dimension of V_j is of order $N_j = 2^{dj} \sim \varepsilon^{-d/t}$ and might therefore be quite large. Moreover, in the case of an integral operator the stiffness matrix K_j might be densely populated. In this section we shall try to circumvent this problem by replacing the full Galerkin inversion by an adaptive algorithm which operate only in subspaces of V_j generated by appropriate wavelets, and exploits in addition the possibility of compressing the matrix K_j when discretized in the wavelet basis. Our estimator f_ε will therefore belong to an adaptive subspace of V_j

$$V_{\Lambda_\varepsilon} = \text{Span}\{\psi_\lambda ; \lambda \in \Lambda_\varepsilon\}, \quad (58)$$

where Λ_ε is an appropriate subset $\{|\lambda| < j\}$. A first intuitive guess for Λ_ε is the set obtained by the thresholding procedure applied on g_ε in the previous section, namely

$$\Lambda_\varepsilon := \{|\lambda| < j ; |\langle g_\varepsilon, \psi_\lambda \rangle| \geq t(\varepsilon)\}. \quad (59)$$

It would thus be tempting to define then $f_\varepsilon \in V_{\Lambda_\varepsilon}$ by applying the Galerkin inversion in this adaptive subspace:

$$\langle K f_\varepsilon, \psi_\lambda \rangle = \langle g_\varepsilon, \psi_\lambda \rangle \text{ for all } \lambda \in \Lambda_\varepsilon. \quad (60)$$

However, it is by no means ensured that such an estimator f_ε will achieve the optimal convergence rate obtained in the previous section when inverting

the thresholded data in the full space V_j . Indeed, there are many instances of operator equations $Kf = g$ where the adapted wavelet set for the solution f differs significantly from the adapted set for the data g .

In order to build a better adapted set of wavelets, we shall introduce a level dependent thresholding operator S_ε to be applied in the solution domain (in contrast to T_ε which operates in the observation domain) according to

$$S_\varepsilon(u_\lambda) = u_\lambda \chi(|u_\lambda| \geq 2^{t|\lambda|} t(\varepsilon)). \quad (61)$$

The role of the weight $2^{t|\lambda|}$ is to take into account the amplification of the noise by the inversion process. The L^2 approximation error obtained by such level dependent thresholding procedures is well understood: see in particular theorem 7.1 in [5], which implies that for $f = \sum_{\lambda \in \nabla} f_\lambda \psi_\lambda \in B_{p,p}^s$, with $s > 0$ and $1/p = 1/2 + s/(2t + d)$ and $S_\varepsilon(f) = \sum_{\lambda \in \nabla} S_\varepsilon(f_\lambda) \psi_\lambda$ we have

$$\|f - S_\varepsilon(f)\|_{L^2}^2 \sim \sum_{|f_\lambda| < 2^{t|\lambda|} \eta} |f_\lambda|^2 \lesssim \|f\|_{B_{p,p}^s}^2 \eta^{2-p} = \|f\|_{B_{p,p}^s}^2 \eta^{4s/(2s+2t+d)}. \quad (62)$$

Our first result shows that S_ε is well adapted to build an adaptive solution of the inverse problem in the following sense: if we apply S_ε to the coordinates of the estimator f_ε defined in the previous section by (34), then the resulting estimator

$$S_\varepsilon(f_\varepsilon) := \sum_{|\lambda| < j} S_\varepsilon(f_{\varepsilon,\lambda}) \psi_\lambda, \quad (63)$$

also satisfies the optimal convergence rate.

Theorem 2. *Let us assume that f belongs to $B := \{f ; \|f\|_{B_{p,p}^s} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t + d)$. Then, we have the estimate*

$$E(\|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|_{L^2}^2) \lesssim \left[\varepsilon \sqrt{|\log \varepsilon|} \right]^{4s/(2s+2t+d)}. \quad (64)$$

It follows that the adaptive estimator $S_\varepsilon(f_\varepsilon) = \sum_{|\lambda| < j} S_\varepsilon(f_{\varepsilon,\lambda}) \psi_\lambda$ is also optimal.

Proof : We want to estimate the expectation of

$$\|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|_{L^2}^2 \lesssim \sum_{|\lambda| < j, |f_{\varepsilon,\lambda}| < 2^{t|\lambda|} t(\varepsilon)} |f_{\varepsilon,\lambda}|^2. \quad (65)$$

Using the fact that if $|a| \leq \eta$ we have for all b

$$|a| \leq |a - b\chi(|b| \geq 2\eta)|, \quad (66)$$

we derive

$$\begin{aligned} \|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|_{L^2}^2 &\lesssim \sum_{\lambda \in \nabla} |f_{\varepsilon,\lambda} - f_\lambda \chi(|f_\lambda| \geq 2^{t|\lambda|+1}t(\varepsilon))|^2 \\ &\lesssim \|f - f_\varepsilon\|_{L^2}^2 + \sum_{|f_\lambda| < 2^{t|\lambda|+1}t(\varepsilon)} |f_\lambda|^2 \\ &\lesssim \|f - f_\varepsilon\|_{L^2}^2 + [\varepsilon \sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)}. \end{aligned}$$

Taking the expectation, we therefore obtain (64)

$$E(\|f - S_\varepsilon(f_\varepsilon)\|_{L^2}^2) \lesssim \left[\varepsilon \sqrt{|\log \varepsilon|} \right]^{4s/(2s+2t+d)}, \quad (67)$$

and the last assertion follows by triangle inequality. \square

Of course, computing $S_\varepsilon(f_\varepsilon)$ is at least as costly as computing f_ε , and we cannot be satisfied with this new estimator. However, it shows us that the level dependent thresholding operator S_ε maintains optimality. Based on this observation we shall now build an adaptive procedure which aims at reducing the computational cost. To this effect, we consider a simple method for solving

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon), \quad (68)$$

namely the fixed step gradient iteration : $F_\varepsilon^0 = 0$ and

$$F_\varepsilon^n = F_\varepsilon^{n-1} + \tau(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}). \quad (69)$$

The convergence rate of F_ε^n to F_ε will slow down for large j due to the bad condition number of K_j . Wavelet discretizations are well adapted to circumvent this problem, using the preconditioned iteration

$$F_\varepsilon^n = F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}), \quad (70)$$

where $D_j = \text{Diag}(2^{-t|\lambda|})$. From the ellipticity of K and the wavelet characterization of $H^{-t/2}$, it indeed follows that the condition number $\kappa(D_j^{-1}K_j)$ remains bounded independently of j , so that a proper choice of τ that depends on K only will ensure a fixed error reduction rate

$$\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2} \leq \rho \|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}, \quad (71)$$

with $\rho \in]0, 1[$ independent of j . The idea is now to perturb this iteration by the thresholding operator S_ε , i.e. define

$$F_\varepsilon^n = S_\varepsilon[F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1})], \quad (72)$$

At each step n , the vector $F_\varepsilon^n = (f_{\varepsilon,\lambda}^n)$ is supported on an adaptive index set Λ_ε^n . The corresponding estimator for f is given

$$f_\varepsilon^n = \sum_{\lambda \in \Lambda_\varepsilon^n} f_{\varepsilon,\lambda}^n \psi_\lambda. \quad (73)$$

When comparing (72) with (70), we observe a first obvious gain in computational time : the cost of the matrix-vector multiplication $K_j F_\varepsilon^{n-1}$ in (72) is of order $(\dim(V_j))^2 \sim 2^{2dj}$, while the cost of the matrix-vector multiplication $K_j F_\varepsilon^{n-1}$ in (72) is of order $\dim(V_j) \times \#(\Lambda_\varepsilon^n) \sim 2^{dj} \#(\Lambda_\varepsilon^n)$. Additional computational time can be gained using the fact that for many relevant instances of operators K , the matrix K_j can be compressed by discarding most of its entries. Such instances include in particular pseudo-differential operators and singular integrals with Calderon-Zygmund type kernel, see e.g. Chapter IV in [3]. For such operators, the entries $K_j(\lambda, \mu)$ can be a priori estimated, allowing to predict in advance those coefficients in $F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1})$ which will be thresholded by S_ε and to avoid their exact computation. With such an approach, we can hope to decrease significantly the cost of each iteration (72).

We shall now prove that after a sufficient number of iterations independent of the unknown smoothness, the estimator f_ε^n also attains the optimal rate of convergence.

Theorem 3. *Let us assume that f belongs to $B := \{f ; \|f\|_{B_{p,p}^s} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$. For $n \geq \log(\varepsilon)/\log(\rho)$, we have*

$$E(\|f_\varepsilon - f_\varepsilon^n\|_{L^2}^2) \lesssim \left[\varepsilon \sqrt{|\log \varepsilon|} \right]^{4s/(2s+2t+d)}. \quad (74)$$

It follows that the adaptive estimator f_ε^n is also optimal.

Proof : The result will follow from proving the reduction estimate

$$E(\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2) \leq \tilde{\rho}^2 E(\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}^2) + C[\varepsilon \sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)} \quad (75)$$

for any $\tilde{\rho} > \rho$ where C depends on the closeness of $\tilde{\rho}$ to ρ . Indeed, assuming that this estimate holds, since

$$E(\|F_\varepsilon - F_\varepsilon^0\|_{\ell^2}^2) = E(\|F_\varepsilon\|_{\ell^2}^2) \lesssim \|F\|_{\ell^2}^2 \leq C, \quad (76)$$

we obtain after n steps

$$E(\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2) \lesssim \max\{\tilde{\rho}^{2n}, [\varepsilon\sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)}\}. \quad (77)$$

Since $4s/(2s+2t+d) < 2$, we have

$$\tilde{\rho}^{2\log(\varepsilon)/\log(\rho)} = \varepsilon^{2\log(\tilde{\rho})/\log(\rho)} \lesssim [\varepsilon\sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)} \quad (78)$$

if $\tilde{\rho}$ is chosen close enough to ρ , so that (74) holds. In order to prove (75), we introduce the intermediate vector

$$F_\varepsilon^{n-1/2} = F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}), \quad (79)$$

for which we have

$$\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2} \leq \rho \|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}. \quad (80)$$

We can then write

$$\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2 = \sum_{|\lambda| < j} |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2} \chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda} t(\varepsilon))|^2. \quad (81)$$

Denoting by K a constant strictly larger than 1 to be fixed later, we split the above sum into three parts Σ_1 , Σ_2 and Σ_3 , respectively corresponding to the index sets

$$\begin{aligned} I_1 &:= \{|\lambda| < j \text{ s.t. } |f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda} t(\varepsilon) \text{ and } |f_{\varepsilon,\lambda}| < K 2^{t|\lambda} t(\varepsilon)\}, \\ I_2 &:= \{|\lambda| < j \text{ s.t. } |f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda} t(\varepsilon)\}, \\ I_3 &:= \{|\lambda| < j \text{ s.t. } |f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda} t(\varepsilon) \text{ and } |f_{\varepsilon,\lambda}| \geq K 2^{t|\lambda} t(\varepsilon)\}. \end{aligned}$$

If $\lambda \in I_1$, we have $|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2} \chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda} t(\varepsilon))| = |f_{\varepsilon,\lambda}|$. Using again the fact that if $|a| \leq \eta$ we have $|a| \leq |a - b \chi(|b| \geq 2\eta)|$ for all b , we can write

$$\begin{aligned} |f_{\varepsilon,\lambda}| &\leq |f_{\varepsilon,\lambda} - f_\lambda \chi(|f_\lambda| \geq 2K 2^{t|\lambda} t(\varepsilon))| \\ &\leq |f_{\varepsilon,\lambda} - f_\lambda| + |f_\lambda - f_\lambda \chi(|f_\lambda| \geq 2K 2^{t|\lambda} t(\varepsilon))|. \end{aligned}$$

It follows that

$$\begin{aligned}\Sigma_1 &\leq 2\|f_\varepsilon - f\|_{L^2}^2 + 2\sum_{|f_\lambda| < 2K2^{t|\lambda}t(\varepsilon)} |f_\lambda|^2 \\ &\lesssim \|f_\varepsilon - f\|_{L^2}^2 + [\varepsilon\sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)},\end{aligned}$$

so that

$$E(\Sigma_1) \lesssim [\varepsilon\sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)}. \quad (82)$$

If $\lambda \in I_2$, we have

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda}t(\varepsilon))| = |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}|, \quad (83)$$

so that

$$\Sigma_2 = \|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2(\Lambda_1)}^2 \quad (84)$$

If $\lambda \in I_3$, we have $|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda}t(\varepsilon))| = |f_{\varepsilon,\lambda}|$ and

$$\begin{aligned}|f_{\varepsilon,\lambda}| &\leq |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| + |f_{\varepsilon,\lambda}^{n-1/2}| \\ &\leq |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| + 2^{t|\lambda}t(\varepsilon).\end{aligned}$$

On the other hand, since $|f_{\varepsilon,\lambda}| > K2^{t|\lambda}t(\varepsilon)$ and $|f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda}t(\varepsilon)$, we also have

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| \geq (K-1)2^{t|\lambda}t(\varepsilon). \quad (85)$$

It follows that

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda}t(\varepsilon))| < \frac{K}{K-1}|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}|, \quad (86)$$

so that

$$\Sigma_3 \leq \left(\frac{K}{K-1}\right)^2 \|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2(\Lambda_3)}^2. \quad (87)$$

Combining (84) and (87), we obtain

$$\Sigma_2 + \Sigma_3 \leq \left(\frac{K}{K-1}\right)^2 \|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2}^2 \leq \left(\frac{K}{K-1}\right)^2 \rho^2 \|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}^2. \quad (88)$$

Combined with (82), this yields the claimed estimate (75) with $\tilde{\rho} = \frac{K}{K-1}\rho$, which can be made arbitrarily close to ρ by taking K large enough up to enlarging the constant C which comes from the estimation of Σ_1 . \square

References

- [1] Abramovitch F. and Silverman B.W. (1998), *wavelet decomposition approaches to statistical inverse problems*, Biometrika, **85**, 115-129.
- [2] Ciarlet P. (1978), *The finite element method for elliptic problems* Studies in Mathematics and its Applications, Vol. 4. North-Holland Publishing Co., Amsterdam-New York-Oxford.
- [3] Cohen, A. (2000), *Wavelets in numerical analysis*, Handbook of Numerical Analysis, vol. VII, P.G. Ciarlet and J.L. Lions, eds., Elsevier, Amsterdam.
- [4] Cohen A., Dahmen W. and DeVore R. (2001), *Adaptive wavelet methods for elliptic operator equations: convergence rates*, Math. Comp. **70**, 27-75.
- [5] Cohen, A., Devore R. and Hochmuth R. (2000), *Restricted nonlinear approximation*, Constr. Approx. **16**, 85-113.
- [6] Cohen A., DeVore R., Kerkyacharian G. and Picard D. (2001), *Maximal spaces with given rate of convergence for thresholding algorithms*, Appl. Comp. Harm. Anal. **11**, 167-191.
- [7] Daubechies, I. (1992), *Ten Lectures on Wavelets*, SIAM, Philadelphia.
- [8] DeVore D. (1998), *Nonlinear Approximation*, Acta Numerica **7**, 51-150.
- [9] Dicken V. and Maass P. (1996), *Wavelet-Galerkin methods for ill-posed problems*, J. Inverse Ill-Posed Probl. **4**, 203-221.
- [10] Donoho, D. (1995) *Nonlinear solution of linear inverse problems by Wavelet-Vaguelette decomposition*, Appl. Comput. Harmon. Anal. **2**, 101-126.
- [11] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995), *Wavelet Shrinkage: Asymptopia?* , J. Roy. Statist. Soc. **57**, 301-369.
- [12] Engl, H.W, Hanke, M. and A. Neubauer (2000), *Regularization of inverse problems*, Kluwer Academic Press.

- [13] Korostelev A. and Tsybakov A. (1993), *Minimax theory of image reconstruction*, Lecture Notes in Statistics. **82**, Springer, New-York.
- [14] Mair, B. and Ruymgaart F. (1996), *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math. **56**, 1424-1444.
- [15] Natterer, F. (1977), *Regularisierung schlecht gestellter Probleme durch Projektionsverfahren*, Numer. Math. **28**, 329-341.
- [16] Nussbaum M. and Pereverzev S. (1999), *The degree of ill-posedness in stochastic and deterministic models*, Preprint 509, Weierstraß-Institut, Berlin.
- [17] Reiß, M. (2001), *Nonparametric estimation for stochastic delay differential equations*. PhD Thesis, Humboldt Universität, Berlin.