SIAM J. Numer. Anal.
Vol. 42, No. 4, pp. 1479–1501

# ADAPTIVE WAVELET GALERKIN METHODS FOR LINEAR INVERSE PROBLEMS[*]

ALBERT COHEN[†], MARC HOFFMANN[‡], AND MARKUS REIß[§]

**Abstract.** We introduce and analyze 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. In a first step, we simply combine a thresholding algorithm on the data with a Galerkin inversion on a fixed linear space. In a second step, a more elaborate method performs the inversion by an adaptive procedure in which a smaller space adapted to the solution is iteratively constructed; this leads to a significant reduction of the computational cost.

**Key words.** statistical inverse problems, Galerkin methods, wavelets and nonlinear methods, Besov spaces, minimax estimation

**AMS subject classifications.** 65J20, 62G07

**DOI.** 10.1137/S0036142902411793

## 1. Introduction.

**1.1. Statistical model.** We want to recover a function $f$ in $L^2(\Omega_X)$, where $\Omega_X$ is a certain bounded domain in $\mathbb{R}^d$, but we are able to observe data about $Kf$ only, where $K : L^2(\Omega_X) \to L^2(\Omega_Y)$ is a compact linear operator and $\Omega_X$ and $\Omega_Y$ are two bounded domains in $\mathbb{R}^d$ and $\mathbb{R}^q$, respectively. In the following, when mentioning $L^2$ or more general function spaces, we shall omit the domain $\Omega_X$ or $\Omega_Y$ when this information is obvious from the context.

We are interested in the statistical formulation of linear inverse problems: we assume that the data are noisy, so that we observe

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

where $\dot{W}$ is a white noise and $\varepsilon$ a noise level. In rigorous probabilistic terms, we observe a Gaussian measure on $L^2$ with drift $Kf$ and intensity $\varepsilon^2$ (see, e.g., [20]). Observable quantities take the form

$$\langle g_\varepsilon, v\rangle = \langle Kf, v\rangle + \varepsilon\, \eta(v),$$

where $v \in L^2$ is a test function and $\eta(v)$ is a Gaussian centered random variable with variance $\|v\|_{L^2}^2$. For $v_1, v_2 \in L^2$ the covariance between $\eta(v_1)$ and $\eta(v_2)$ is given by the scalar product $\langle v_1, v_2\rangle$. In particular, if $v_1$ and $v_2$ are orthogonal, the random variables $\eta(v_1)$ and $\eta(v_2)$ are stochastically independent. The cognitive value of the white noise model (1.1) is discussed in detail in [4], [26] and the references therein.

[*] Received by the editors July 22, 2002; accepted for publication (in revised form) October 22, 2003; published electronically December 16, 2004. This research was partially supported by the European network Dynstoch and the DFG-Sonderforschungsbereich 383.

http://www.siam.org/journals/sinum/42-4/41179.html

[†] Laboratoire J.L. Lions, Université Pierre et Marie Curie, 175 rue du Chevaleret, 75013, Paris, France (cohen@ann.jussieu.fr).

[‡] Laboratoire d'Analyse et de Mathématiques Appliquées, Université Marne-la-Vallée, 5 Blvd. Descartes, 77454, Marne-la-Vallée, Cedex 2, France (hoffmann@math.univ-mlv.fr).

[§] Institut für Mathematik, Humboldt-Universität zu Berlin, unter den Linden 6, D-10099, Berlin, Germany (reiss@mathematik.hu-berlin.de).

If $f_\varepsilon \in L^2$ is an estimator of $f$, that is a measurable quantity w.r.t. the data $g_\varepsilon$, we measure its accuracy by the mean-square error $E(\|f_\varepsilon - f\|^2_{L^2})$ as $\varepsilon \to 0$, with $E(\cdot)$ denoting the expectation operator.

**1.2. SVD and Galerkin projection.** Among the most popular regularization methods, let us first mention the *singular value decomposition* (SVD); see, e.g., [21], [22], and [24]. Although very attractive theoretically, the SVD suffers from two limitations. First, the singular basis functions may be difficult to determine and manipulate numerically. Second, while these bases are fully adapted to describe the action of $K$, they might not be appropriate for the accurate description of the solution with a small number of parameters (see, e.g., [16]). Concerning numerical simplicity, *projection methods* are more appealing. Given finite dimensional subspaces $X_h \subset L^2(\Omega_X)$ and $Y_h \subset L^2(\Omega_Y)$ with $\dim(X_h) = \dim(Y_h)$, one defines the approximation $f_\varepsilon$ as the solution in $X_h$ of the problem

$$\langle K f_\varepsilon, g_h\rangle = \langle g, g_h\rangle \quad \text{for all } \ g_h \in Y_h, \tag{1.2}$$

which amounts to solving a linear system (see [25] for a general approach to projection methods). In the case where $\Omega_X = \Omega_Y$ and $K$ is a self-adjoint positive definite operator, we choose $Y_h = X_h$ and the linear system is particularly simple to solve since the corresponding discretized operator $K_h$ is symmetric positive definite: this is the so-called *Galerkin method*. In the case of general $\Omega_X \neq \Omega_Y$ and injective $K$, one may choose $Y_h := K(X_h)$ and we are led back to the Galerkin method applied to the least squares equation $K^*Kf = K^*g$ with data $K^*g_\varepsilon$, where $K^*$ denotes the adjoint of $K$. 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 some smoothness.

However, the Galerkin projection method suffers from two drawbacks which are encountered in all *linear* estimation methods, including, in particular, the SVD. First, the choice of $h$ with respect to the noise level $\varepsilon$ depends on the regularity of the solution which is almost always unknown in advance. Second, the use of a finite element space $X_h$ with a fixed uniform mesh size $h$ does not provide any spatial adaptation.

**1.3. Wavelet-vaguelette decomposition.** In recent years, *nonlinear* methods have been developed, with the objective of automatically adapting to unknown smoothness and locally singular behavior 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 methods, 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., [18]. 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 [16] a wavelet-like decomposition which is specifically adapted to describe the action of $K$, the so-called *wavelet-vaguelette decomposition* (WVD), and proposed applying a thresholding algorithm on this decomposition. In [1] Donoho's method was compared with the similar *vaguelette-wavelet decomposition* (VWD) algorithm. Both methods rely on an orthogonal wavelet basis $(\psi_\lambda)$ and associated Riesz bases of "vaguelettes" defined as

$$v_\lambda = \beta_\lambda K^{-1}\psi_\lambda \;\text{ and }\; u_\lambda = \beta_\lambda (K^*)^{-1}\psi_\lambda, \tag{1.3}$$

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

$$f = \sum_\lambda \beta_\lambda^{-1}\langle Kf, u_\lambda\rangle \psi_\lambda = \sum_\lambda \beta_\lambda^{-1}\langle Kf, \psi_\lambda\rangle v_\lambda. \tag{1.4}$$

The WVD and VWD estimation methods amount to estimating the coefficients in these expansions from the observed data and applying a thresholding procedure. On a theoretical level, similarly to wavelet thresholding in the case of simple denoising, both WVD and VWD allow recovery at the same rate as the projection method, under weaker smoothness conditions, a fact that reflects their ability for spatial adaptivity.

On a more applied level, numerical implementations in [1] and [15] have illustrated the efficiency of both WVD and VWD methods, in the case of operators that behave like integration $Kf(x) = \int_0^x f(t)dt$. For more general operators, however, the assumption that $K^{-1}\psi_\lambda$ or $(K^*)^{-1}\psi_\lambda$ are known for all indices $\lambda$ may simply result in putting forward the inversion problem: if an integral operator has a kernel with a complicated structure (see [5]), or if this kernel is itself derived from observations (see [27]), this inversion has to be done numerically with additional computational cost and error. In other words, the vaguelettes $u_\lambda$ and $v_\lambda$ might be difficult to handle numerically (similar to the SVD functions), and in particular they are not ensured to have compact support.

### 1.4. Our approach: Adaptive wavelet Galerkin.

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. This is the goal of the paper.

Adaptive Galerkin methods are well established in the context of solving operator equations *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., [8]. Our goal is to introduce and analyze similar adaptive wavelet Galerkin algorithms in the context of statistical inverse problems. Such adaptive algorithms involve only the wavelet system $(\psi_\lambda)$ and are therefore often easier to handle numerically than WVD and VWD. 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, one can exploit 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 estimator.

Concerning the organization of the paper, we progressively develop our methodology. In section 2, we introduce general assumptions on model (1.1), in terms of mapping properties of the operator $K$ between smoothness spaces. After a brief recall of the analysis of the linear Galerkin method using a wavelet multiresolution space $V_j$ in section 3.1, a first nonlinear method is proposed in section 3.2, which initially operates in a way similar to the VWD, by thresholding the wavelet coefficients $g_\lambda^\varepsilon := \langle g_\varepsilon, \psi_\lambda\rangle$ with $\lambda$ 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 $V_j$. As $\varepsilon$ decreases, the scale level $j(\varepsilon)$ grows and the Galerkin approximation thus becomes computationally heavy, while the solution could still be well represented by a small adaptive set of wavelets within $V_j$. Therefore, we propose in section 4 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. As we completed the revision of this paper, we became aware of a related approach recently proposed in [13], based on least squares minimization with a nonquadratic penalization term in a deterministic setting, which results in a similar combination of gradient iteration with a thresholding procedure, yet operating in infinite dimension. Both methods in sections 3.2 and 4 are proved to achieve the same minimax rate as WVD and VWD under the same general assumptions on the operator $K$. We eventually compare the different estimators in section 5 numerically on the example of a singular integral equation of the first kind. For simplicity, we present our methods and results in the case where $K$ is elliptic and self-adjoint. The extension to more general operators, via a least squares approach, is the object of Appendix A. We also discuss in Appendix B several properties of multiresolution spaces which are used throughout the paper.

## 2. 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 $L^2$-continuous. This is 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., $W^s_p$, $B^s_{p,q}$) if this space is mapped onto $H^{s+t}$ (resp., $W^{s+t}_p$, $B^{s+t}_{p,q}$).

The estimator $f_\varepsilon$ will be searched within a finite dimensional subspace $V$ of $L^2(\Omega_X)$ based on the projection method. In the case where $\Omega_X = \Omega_Y = \Omega$ and $K$ is self-adjoint positive definite, we shall use the Galerkin method, that is,

$$\text{find } f_\varepsilon \in V \text{ such that } \langle Kf_\varepsilon, v\rangle = \langle g_\varepsilon, v\rangle \text{ for all } v \in V. \tag{2.1}$$

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

$$\langle Kf, f\rangle \sim \|f\|^2_{H^{-t/2}}, \tag{2.2}$$

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 Dirichlet, periodic, and so on). The symbol $a \sim b$ means that there exists $c > 0$ independent of $f$ such that $c^{-1}b \leq a \leq cb$.

In the case where $\Omega_X \neq \Omega_Y$ or when $K$ is not self-adjoint positive definite, we shall consider the least squares method, that is,

$$\text{find } f_\varepsilon \in V \text{ which minimizes } \|Kv - g_\varepsilon\|^2_{L^2} \text{ among all } v \in V. \tag{2.3}$$

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

$$\|Kf\|^2_{L^2} = \langle K^*Kf, f\rangle \sim \|f\|^2_{H^{-t}}. \tag{2.4}$$

We shall only deal with this general situation in Appendix A, and we therefore assume for the next sections that $K$ is self-adjoint positive definite and satisfies (2.2).

**3. Nonlinear estimation by linear Galerkin.** Wavelet bases have been documented in numerous textbooks and survey papers (see [12] for a general treatment). With a little effort, they can be adapted to fairly general domains $\Omega \subset \mathbb{R}^d$ (see [7] for a survey of these adaptations as well as a discussion of the characterizations of function spaces on $\Omega$ by wavelet coefficients).

The wavelet decomposition of a function $f \in L^2$ takes the form

$$f = \sum_{\lambda \in \Gamma_{j_0}} \alpha_\lambda \varphi_\lambda + \sum_{j \geq j_0} \sum_{\lambda \in \nabla_j} f_\lambda \psi_\lambda, \tag{3.1}$$

where $(\varphi_\lambda)_{\lambda \in \Gamma_j}$ is the scaling function basis spanning the approximation at level $j$ with appropriate boundary modification, and $(\psi_\lambda)_{\lambda \in \nabla_j}$ is the wavelet basis spanning the details at level $j$. The index $\lambda$ concatenates the usual scale and space indices $j$ and $k$. The coefficients of $f$ can be evaluated according to

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

where $\tilde{\varphi}_\lambda$ and $\tilde{\psi}_\lambda$ are the corresponding dual scaling functions and wavelets. In what follows, we shall (merely for notational convenience) always take $j_0 := 0$. 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, if we write $|\lambda| = j$ if $\lambda \in \nabla_j$, we simply have

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

**3.1. Preliminaries: Linear estimation by linear Galerkin.** We first recall some classical results on the linear Galerkin projection method. For some scale $j > 0$ to be chosen further, let $V_j$ be the linear space spanned by $(\varphi_\lambda)_{\lambda \in \Gamma_j}$. 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

$$\text{find } f_\varepsilon \in V_j \text{ such that } \ \langle K f_\varepsilon, v \rangle = \langle g_\varepsilon, v \rangle \ \text{ for all } v \in V_j. \tag{3.2}$$

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

$$K_j F_\varepsilon = G_\varepsilon. \tag{3.3}$$

The analysis of the method is summarized in the following classical result; see, for instance, [21], [22], [25], [15], and the references therein.

Proposition 3.1. *Assuming that $f$ belongs to the Sobolev ball $B := \{f \in H^s \,;\, \|f\|_{H^s} \leq M\}$ and choosing $j = j(\varepsilon)$ with $2^{-j(\varepsilon)} \sim \varepsilon^{2/(2s+2t+d)}$, we have*

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

*and this rate is minimax over the class $B$.*

The symbol $\lesssim$ means that the left-hand side is bounded by a constant multiple of the right-hand side where the constant possibly depends on $s$ and $M$, but not on

$\varepsilon$. In order to be self contained, we give a proof of Proposition 3.1. The techniques we use here will prove helpful in the sequel.

*Proof.* The analysis of this method can be done by decomposing $f_\varepsilon$ according to

$$f_\varepsilon = f_j + h_\varepsilon, \tag{3.4}$$

where the terms $f_j$ and $h_\varepsilon$ are, respectively, solutions of (3.2) with $Kf$ and $\varepsilon\dot{W}$ in place of $g_\varepsilon$ as the right-hand side. This gives the classical decomposition of the estimation error into a bias and variance term

$$E(\|f - f_\varepsilon\|_{L^2}^2) = \|f - f_j\|_{L^2}^2 + E(\|h_\varepsilon\|_{L^2}^2). \tag{3.5}$$

Both terms are estimated by inverse and direct estimates with respect to Sobolev spaces, which are recalled in Appendix B. The variance term can be estimated as follows: we first use the ellipticity property (2.2), which gives

$$\|h_\varepsilon\|_{H^{-t/2}}^2 \lesssim \langle Kh_\varepsilon, h_\varepsilon\rangle = \varepsilon\langle \dot{W}, h_\varepsilon\rangle \le \varepsilon\|P_j\dot{W}\|_{L^2}\|h_\varepsilon\|_{L^2}. \tag{3.6}$$

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

$$\|h_\varepsilon\|_{L^2} \lesssim \varepsilon 2^{tj}\|P_j\dot{W}\|_{L^2}, \tag{3.7}$$

and therefore

$$E(\|h_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^2 2^{2tj}\dim(V_j) \lesssim \varepsilon^2 2^{(2t+d)j}. \tag{3.8}$$

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

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

where we have again used the inverse 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

$$\|f - f_j\|_{L^2} \lesssim \inf_{g_j\in V_j}[\|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}}]. \tag{3.9}$$

Assuming that $f$ belongs to $B$ we obtain the direct estimate

$$\inf_{g_j\in V_j}[\|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}}] \lesssim 2^{-sj}, \tag{3.10}$$

and therefore

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-2sj}. \tag{3.11}$$

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

$$2^{-j(\varepsilon)} \sim \varepsilon^{2/(2s+2t+d)}, \tag{3.12}$$

and the rate of convergence

$$E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^{4s/(2s+2t+d)}, \tag{3.13}$$

which ends the proof of Proposition 3.1. □

**3.2. Nonlinear estimation by linear Galerkin.** Our first nonlinear estimator $f_\varepsilon$ simply consists of applying a thresholding algorithm on the observed data before performing the linear Galerkin inversion which was described in the previous section: for some $j \geq 0$ to be chosen later, we define $f_\varepsilon = \sum_{|\lambda|<j} f_{\varepsilon,\lambda}\psi_\lambda \in V_j$ such that

$$\langle Kf_\varepsilon, \psi_\lambda\rangle = T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle) \tag{3.14}$$

for all $|\lambda| < j$. Here $T_\varepsilon$ is the hard thresholding operator

$$T_\varepsilon(x) = x\chi(|x| \geq t(\varepsilon)) \tag{3.15}$$

(where $\chi(P)$ is 1 if $P$ is true and 0 otherwise), and the threshold $t(\varepsilon)$ has the usual size

$$t(\varepsilon) := 8\varepsilon\sqrt{|\log\varepsilon|}. \tag{3.16}$$

Defining the data vector $G_\varepsilon := (\langle g_\varepsilon, \psi_\lambda\rangle)_{|\lambda|<j}$, and the Galerkin stiffness matrix $K_j := (\langle K\psi_\lambda, \psi_\mu\rangle)_{|\lambda|,|\mu|<j}$, the coordinate vector $F_\varepsilon := (f_{\varepsilon,\lambda})_{|\lambda|<j}$ of $f_\varepsilon$ in the wavelet basis $(\psi_\lambda)_{|\lambda|<j}$ is the solution of the linear system

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon), \tag{3.17}$$

where $T_\varepsilon(G_\varepsilon) := (T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle))_{|\lambda|<j}$. 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 $(\psi_\lambda)$ is an orthonormal basis) by

$$f_\varepsilon := \sum_{|\lambda|<j} T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle)K^{-1}\psi_\lambda. \tag{3.18}$$

The solution $f_\varepsilon$ of (3.14) has a similar form with the vaguelettes $K^{-1}\psi_\lambda$ replaced by their Galerkin approximations $u_\lambda^j \in V_j$ such that

$$\langle Ku_\lambda^j, v\rangle = \langle\psi_\lambda, v\rangle \text{ for all } v \in V_j. \tag{3.19}$$

We therefore expect that this estimator behaves 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 operator. It should be noted that the lower bound on $j$ does not depend on the unknown smoothness of $f$, in contrast to the classical thresholding for signal denoising.

THEOREM 3.2. *Assume that $f$ belongs to $B := \{f\ ;\ \|f\|_{B^s_{p,p}} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$. 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 $B^s_{p,p}$. Then the estimator from equation* (3.14) *satisfies the minimax rate*

$$\sup_{f\in B} E(\|f - f_\varepsilon\|^2_{L^2}) \ \lesssim\ [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}, \tag{3.20}$$

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

*Proof.* We write again $f_\varepsilon = f_j + h_\varepsilon$, where $f_j \in V_j$ is the solution of the linear problem with data $g_\varepsilon$

$$\text{find } f_j \in V_j \text{ such that } \langle Kf_j, v\rangle = \langle g, v\rangle \text{ for all } v \in V_j, \tag{3.21}$$

where $g = Kf$. Correspondingly, the term $h_\varepsilon$ represents the solution of the linear problem with the thresholding error as data, in other words $h_\varepsilon \in V_j$ such that

$$\langle Kh_\varepsilon, \psi_\lambda\rangle = T_\varepsilon(\langle g_\varepsilon, \psi_\lambda\rangle) - \langle g, \psi_\lambda\rangle \tag{3.22}$$

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 $B^s_{p,p}$ is continuously imbedded in $H^\alpha$ whenever

$$\alpha \le s + d/2 - d/p = 2ts/(2t + d). \tag{3.23}$$

By the same arguments as in the previous section we obtain

$$\|f - f_j\|_{L^2}^2 \lesssim 2^{-4sj\frac{t}{d+2t}}. \tag{3.24}$$

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

$$2^{-j} \le \varepsilon^{\frac{d+2t}{t(2s+2t+d)}}. \tag{3.25}$$

We have $\varepsilon^{1/t} \le \varepsilon^{\frac{d+2t}{t(2s+2t+d)}}$ for all $s \ge 0$. Therefore the choice $2^{-j} \le \varepsilon^{1/t}$ yields

$$\|f - f_j\|_{L^2}^2 \lesssim \varepsilon^{4s/(2s+2t+d)}. \tag{3.26}$$

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

$$H_\varepsilon = K_j^{-1}(T_\varepsilon(G_\varepsilon) - G_j) \tag{3.27}$$

and remark that $T_\varepsilon(G_\varepsilon) - G_j$ is exactly the error when estimating $G_j$ by the thresholding procedure on the data $G_\varepsilon$. We shall take into account the action of $K_j^{-1}$ by measuring this error in the wavelet version of the $H^t$ norm

$$\|U\|_{h^t}^2 := \sum_{|\lambda|<j} 2^{2t|\lambda|}|u_\lambda|^2. \tag{3.28}$$

Indeed, we shall see that the stability property

$$\|K_j^{-1}U\|_{\ell^2} \lesssim \|U\|_{h^t} \tag{3.29}$$

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

$$E(\|T_\varepsilon(G_\varepsilon) - G_j\|_{h^t}^2) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{3.30}$$

Such a rate is a particular case of classical results on wavelet thresholding, using the fact that $g$ belongs to a Besov ball $\tilde{B} = \{g \in B^{s+t}_{p,p}\ ;\ \|g\|_{B^{s+t}_{p,p}} \le \tilde{M}\}$. For this model, (3.30) follows, e.g., from Theorem 4 in [10]. We are thus left with proving the stability property (3.29). To do so, we remark that if

$$K_j^{-1}U = V = (v_\lambda)_{|\lambda|<j}, \tag{3.31}$$

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

$$U = (\langle u, \psi_\lambda \rangle)_{|\lambda|<j} \text{ and } \langle u, \psi_\lambda \rangle = 0 \text{ if } |\lambda| \geq j. \tag{3.32}$$

It follows that

$$\|K^{-1}u - v\|_{H^{-t/2}} \lesssim 2^{-jt/2} \|K^{-1}u\|_{L^2} \lesssim 2^{-jt/2} \|u\|_{H^t}. \tag{3.33}$$

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

$$\|K^{-1}u - P_j K^{-1}u\|_{H^{-t/2}} \lesssim 2^{-jt/2} \|K^{-1}u\|_{L^2} \lesssim 2^{-jt/2} \|u\|_{H^t}. \tag{3.34}$$

It follows that

$$\|v - P_j K^{-1}u\|_{H^{-t/2}} \lesssim 2^{-jt/2} \|u\|_{H^t}. \tag{3.35}$$

Using the inverse estimate, we obtain

$$\|v - P_j K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t}, \tag{3.36}$$

so that

$$\|v\|_{L^2} \lesssim \|u\|_{H^t} + \|P_j K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t} + \|K^{-1}u\|_{L^2} \lesssim \|u\|_{H^t}. \tag{3.37}$$

Using the wavelet characterization of $L^2$ and $H^t$, this yields (3.29). □

*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

$$v_\lambda = \beta_\lambda K^{-1} \psi_\lambda = 2^{-t|\lambda|} K^{-1} \psi_\lambda \tag{3.38}$$

constitute a Riesz basis of $L^2$.

## 4. Nonlinear estimation by adaptive Galerkin.

The main defect of the method 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 operates only in subspaces of $V_j$ generated by appropriate wavelets and which exploits, in addition, the possibility of compressing the matrix $K_j$ when discretized in the wavelet basis. Our estimator $f_\varepsilon$ will therefore belong to an adaptive subspace of $V_j$

$$V_{\Lambda_\varepsilon} = \text{Span}\{\psi_\lambda \ ; \ \lambda \in \Lambda_\varepsilon\}, \tag{4.1}$$

where $\Lambda_\varepsilon$ is a data-driven subset of $\{|\lambda| < j\}$. A first intuitive guess for $\Lambda_\varepsilon$ is the set obtained by the thresholding procedure applied on $g_\varepsilon$ in the previous section, namely

$$\Lambda_\varepsilon := \{|\lambda| < j \ ; \ |\langle g_\varepsilon, \psi_\lambda \rangle| \geq t(\varepsilon)\}. \tag{4.2}$$

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

$$\langle f_\varepsilon, \psi_\lambda \rangle = \langle g_\varepsilon, \psi_\lambda \rangle \text{ for all } \lambda \in \Lambda_\varepsilon. \tag{4.3}$$

However, it is by no means ensured that such an estimator $f_\varepsilon$ will achieve the optimal convergence rate in the case of nonlocal operators $K$. 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 $S_\varepsilon$ to be applied in the solution domain (in contrast to $T_\varepsilon$ which operates in the observation domain) according to

$$S_\varepsilon(u_\lambda) = u_\lambda \chi(|u_\lambda| \geq 2^{t|\lambda|} t(\varepsilon)). \tag{4.4}$$

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 [9], which implies that for $f = \sum_{\lambda\in\nabla} f_\lambda \psi_\lambda \in B^s_{p,p}$, 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

$$\|f - S_\varepsilon(f)\|_{L^2}^2 \sim \sum_{|f_\lambda| < 2^{t|\lambda|} t(\varepsilon)} |f_\lambda|^2$$

$$\lesssim \|f\|^2_{B^s_{p,p}} t(\varepsilon)^{2-p} = \|f\|^2_{B^s_{p,p}} t(\varepsilon)^{4s/(2s+2t+d)}. \tag{4.5}$$

Our first result shows that $S_\varepsilon$ is well adapted to build an adaptive solution of the inverse problem in the following sense: if we apply $S_\varepsilon$ to the coordinates of the estimator $f_\varepsilon$ defined in the previous section by (3.14), then the resulting estimator

$$S_\varepsilon(f_\varepsilon) := \sum_{|\lambda|<j} S_\varepsilon(f_{\varepsilon,\lambda})\psi_\lambda \tag{4.6}$$

still satisfies the optimal convergence rate.

Theorem 4.1. *Let us assume that $f$ belongs to $B := \{f \ ; \ \|f\|_{B^s_{p,p}} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$. Then, we have the estimate*

$$\sup_{f\in B} E(\|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|^2_{L^2}) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{4.7}$$

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

*Proof.* We want to estimate the expectation of

$$\|f_\varepsilon - S_\varepsilon(f_\varepsilon)\|^2_{L^2} \lesssim \sum_{|\lambda|<j, |f_{\varepsilon,\lambda}| < 2^{t|\lambda|} t(\varepsilon)} |f_{\varepsilon,\lambda}|^2. \tag{4.8}$$

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

$$|a| \leq |a - b\chi(|b| \geq 2\eta)|, \tag{4.9}$$

we derive

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

Taking the expectation, we obtain (4.7) and

$$E(\|f - S_\varepsilon(f_\varepsilon)\|_{L^2}^2) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)} \tag{4.10}$$

follows by the triangle inequality. □

Of course, computing $S_\varepsilon(f_\varepsilon)$ is more costly than computing $f_\varepsilon$, and we cannot be satisfied with this new estimator. However, it shows us that the level-dependent thresholding operator $S_\varepsilon$ maintains optimality. Based on this observation we now build an adaptive procedure which aims at reducing the computational cost. Let us note that many numerical methods are available in order to solve the system

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon) \tag{4.11}$$

with the optimal cost $\mathcal{O}(N_j)$, where $N_j = \dim(V_j) \sim 2^{dj}$. In particular, one can rely on multigrid methods [3] in the case of local elliptic operators and fast multipole or wavelet [2] methods in the case of integral operators. However, our goal here is to reduce further the computational cost to the order of the dimension of the compressed solution, i.e., the number of nonzero coefficients in $S_\varepsilon(f_\varepsilon)$. Therefore, we shall rather be inspired by the approach introduced in [8] for adaptively solving operator equations *without noise*: consider a simple method for solving

$$K_j F_\varepsilon = T_\varepsilon(G_\varepsilon), \tag{4.12}$$

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

$$F_\varepsilon^n = F_\varepsilon^{n-1} + \tau(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}) \tag{4.13}$$

with a sufficiently small enough relaxation parameter $\tau > 0$. The convergence rate of $F_\varepsilon^n$ to $F_\varepsilon$ might deteriorate for large $j$ due to the bad condition number of $K_j$. Wavelet discretization is well adapted to circumvent this problem, when using the preconditioned iteration

$$F_\varepsilon^n = F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1}), \tag{4.14}$$

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

$$\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2} \leq \rho\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}, \tag{4.15}$$

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

$$F_\varepsilon^n = S_\varepsilon[F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_j F_\varepsilon^{n-1})]. \tag{4.16}$$

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

$$f_\varepsilon^n = \sum_{\lambda\in\Lambda_\varepsilon^n} f_{\varepsilon,\lambda}^n \psi_\lambda. \tag{4.17}$$

When comparing (4.16) with (4.14), we observe a first obvious gain in computational time: the cost of the matrix-vector multiplication $K_j F_\varepsilon^{n-1}$ in (4.16) is of order $(\dim(V_j))^2 \sim 2^{2dj}$, while the cost of the matrix-vector multiplication $K_jF_\varepsilon^{n-1}$ in (4.16) 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 pseudodifferential operators and singular integral operators with Calderon–Zygmund-type kernel; see, e.g., Chapter 4 in [7] and [8]. For such operators, the entries $K_j(\lambda,\mu)$ can be estimated a priori, allowing us to predict in advance those coefficients in $F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_jF_\varepsilon^{n-1})$ which will be thresholded by $S_\varepsilon$ and to avoid their exact computation. With such an approach, the cost of each iteration (4.16) can therefore be pushed down to the order $\#(\Lambda_\varepsilon^n)^2$, and even to $\#(\Lambda_\varepsilon^n)$ using a fast matrix vector multiplication; see Chapter 4 in [7] and [8].

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

THEOREM 4.2. *Let us assume that $f$ belongs to $B := \{f \ ;\ \|f\|_{B^s_{p,p}} \le M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$. For $n \ge \log(\varepsilon)/\log(\rho)$, we have*

$$\sup_{f\in B} E(\|f_\varepsilon - f_\varepsilon^n\|_{L^2}^2) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{4.18}$$

*It follows that the adaptive estimator $f_\varepsilon^n$ is also rate-optimal.*

*Proof.* The result will follow from the reduction estimate

$$E(\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2) \le \tilde\rho^2 E(\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}^2) + C[\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)} \tag{4.19}$$

for any $\tilde\rho > \rho$, where $C$ depends of the closeness of $\tilde\rho$ to $\rho$. Indeed, assuming this estimate to hold, from

$$E(\|F_\varepsilon - F_\varepsilon^0\|_{\ell^2}^2) = E(\|F_\varepsilon\|_{\ell^2}^2) \lesssim \|F\|_{\ell^2}^2 \le C \tag{4.20}$$

we obtain after $n$ steps

$$E(\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2) \lesssim \max\{\tilde\rho^{2n}, [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}\}. \tag{4.21}$$

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

$$\tilde\rho^{2\log(\varepsilon)/\log(\rho)} = \varepsilon^{2\log(\tilde\rho)/\log(\rho)} \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)} \tag{4.22}$$

if $\tilde\rho$ is chosen close enough to $\rho$, and (4.18) follows. In order to prove (4.19), we introduce the intermediate vector

$$F_\varepsilon^{n-1/2} = F_\varepsilon^{n-1} + \tau D_j^{-1}(T_\varepsilon(G_\varepsilon) - K_jF_\varepsilon^{n-1}), \tag{4.23}$$

for which we have

$$\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2} \le \rho\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}. \tag{4.24}$$

We can then write

$$\|F_\varepsilon - F_\varepsilon^n\|_{\ell^2}^2 = \sum_{|\lambda|<j} |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \ge 2^{t|\lambda|}t(\varepsilon))|^2. \tag{4.25}$$

Denoting by $K > 1$ a constant to be fixed later, we split the above sum into three parts $\Sigma_1$, $\Sigma_2$, and $\Sigma_3$, respectively corresponding to the index sets

$$\begin{aligned}
I_1 &:= \{|\lambda| < j \ ;\ |f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda|}t(\varepsilon) \text{ and } |f_{\varepsilon,\lambda}| < K2^{t|\lambda|}t(\varepsilon)\},\\
I_2 &:= \{|\lambda| < j \ ;\ |f_{\varepsilon,\lambda}^{n-1/2}| \ge 2^{t|\lambda|}t(\varepsilon)\},\\
I_3 &:= \{|\lambda| < j \ ;\ |f_{\varepsilon,\lambda}^{n-1/2}| < 2^{t|\lambda|}t(\varepsilon) \text{ and } |f_{\varepsilon,\lambda}| \ge K2^{t|\lambda|}t(\varepsilon)\}.
\end{aligned}$$

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

$$\begin{aligned} |f_{\varepsilon,\lambda}| &\leq |f_{\varepsilon,\lambda} - f_\lambda \chi(|f_\lambda| \geq 2K2^{t|\lambda|}t(\varepsilon))| \\ &\leq |f_{\varepsilon,\lambda} - f_\lambda| + |f_\lambda - f_\lambda \chi(|f_\lambda| \geq 2K2^{t|\lambda|}t(\varepsilon))|. \end{aligned}$$

It follows that

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

so that

$$E(\Sigma_1) \lesssim [\varepsilon\sqrt{|\log\varepsilon|}]^{4s/(2s+2t+d)}. \tag{4.26}$$

If $\lambda \in I_2$, we have

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda|}t(\varepsilon))| = |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}|, \tag{4.27}$$

so that

$$\Sigma_2 = \|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2(\Lambda_1)}^2. \tag{4.28}$$

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

$$|f_{\varepsilon,\lambda}| \leq |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| + |f_{\varepsilon,\lambda}^{n-1/2}| \leq |f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| + 2^{t|\lambda|}t(\varepsilon). \tag{4.29}$$

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

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}| \geq (K-1)2^{t|\lambda|}t(\varepsilon). \tag{4.30}$$

It follows that

$$|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}\chi(|f_{\varepsilon,\lambda}^{n-1/2}| \geq 2^{t|\lambda|}t(\varepsilon))| < \tfrac{K}{K-1}|f_{\varepsilon,\lambda} - f_{\varepsilon,\lambda}^{n-1/2}|, \tag{4.31}$$

so that

$$\Sigma_3 \leq (\tfrac{K}{K-1})^2\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2(\Lambda_3)}^2. \tag{4.32}$$

Combining (4.28) and (4.32), we obtain

$$\Sigma_2 + \Sigma_3 \leq (\tfrac{K}{K-1})^2\|F_\varepsilon - F_\varepsilon^{n-1/2}\|_{\ell^2}^2 \leq (\tfrac{K}{K-1})^2\rho^2\|F_\varepsilon - F_\varepsilon^{n-1}\|_{\ell^2}^2. \tag{4.33}$$

Combined with (4.26), this yields the claimed estimate (4.19) with $\tilde{\rho} = \frac{K}{K-1}\rho$, which can be made arbitrarily close to $\rho$ by taking $K$ large enough, up to enlarging the constant $C$ which comes from the estimation of $\Sigma_1$. □

*Remark.* As was already explained, the cost of each iteration can be, at most, pushed down to $\mathcal{O}(\#(\Lambda_\varepsilon^n))$, which allows the estimate of the computational cost of the algorithm by

$$\mathcal{C}(\varepsilon) \leq \sum_{n \leq \log(\varepsilon)/\log(\rho)} \#(\Lambda_\varepsilon^n). \tag{4.34}$$

In addition, a rough estimate of $\#(\Lambda_\varepsilon^n)$ can be obtained from the smoothness of $f$, assuming that the number of coefficients retained at each thresholding step is of the same order as the number of coefficients which would be retained when applying $S_\varepsilon$ to the exact $f$. In order to estimate this number, we note that if $f$ belongs to $B := \{f \; ; \; \|f\|_{B^s_{p,p}} \leq M\}$ with $s > 0$ and $1/p = 1/2 + s/(2t+d)$, it also belongs to $\tilde{B} := \{f \; ; \; \|f\|_{B^s_{q,q}} \leq M\}$ with $1/q = 1/2 + (s+t)/d$ (since $q \leq p$), or equivalently

$$\sum \|f_\lambda \psi_\lambda\|^q_{H^{-t}} \sim \sum |f_\lambda 2^{-t|\lambda|}|^q \leq M^q. \tag{4.35}$$

It follows that

$$\#(\Lambda_\varepsilon^n) \;\lesssim\; t(\varepsilon)^{-q} \sim [\varepsilon\sqrt{|\log\varepsilon|}]^{-\frac{2d}{d+2s+2t}}, \tag{4.36}$$

and therefore

$$\mathcal{C}(\varepsilon) \;\lesssim\; [\varepsilon\sqrt{|\log\varepsilon|}]^{-\frac{2d}{d+2s+2t}} \log(\varepsilon)/\log(\rho). \tag{4.37}$$

In contrast, the cost of a nonadaptive inversion of (4.11) when using an optimal solver is of order

$$\mathcal{C}(\varepsilon) \sim \dim(V_j) \sim 2^{dj} \;\lesssim\; \varepsilon^{-d/t}. \tag{4.38}$$

Since $\frac{2d}{d+2s+2t} < \frac{d}{t}$, we see that (4.37) always improves (4.38). Let us insist on the fact that this improvement relies on the compressibility of the stiffness matrix in the sense that the adaptive matrix-vector multiplication involved in the iteration only costs $\mathcal{O}\left(\#(\Lambda_\varepsilon^n)\right)$ operations. For arbitrary noncompressible matrices, this cost should be multiplied by $2^{jd}$. Note also that the cost for nonadaptive inversion may then also be substantially higher than $\dim(V_j)$.

Note also that, in addition to the fact that the algorithm adapts to unknown smoothness, its practical computational cost also decreases as the amount of smoothness increases.

## 5. A numerical example.

We focus on a simple example of a logarithmic potential kernel in dimension one and a single test function. The relatively simple analytical form of this operator gives the ability to approximate reasonably well its singular values. Therefore, the SVD method is feasible and can be compared with the Galerkin approach with reasonable accuracy. Our goals are the following:

1. To illustrate on a specific test case that (1) the oracle-SVD and the oracle linear Galerkin methods are comparable; (2) the nonlinear Galerkin method of section 4, obtained by thresholding the data in the image domain, achieves comparable numerical results as the oracle-SVD and the oracle linear Galerkin estimators.
2. To verify on an example that the empirical $L^2$ error stabilizes beyond a certain resolution level $j$, which is related to the noise level $\varepsilon$ and the degree of ill-posedness of the operator. In theory, we know that the condition $2^{-j} \;\lesssim\; \varepsilon^{1/t}$ is sufficient to obtain optimality and that there is no gain in increasing $j$ further.
3. To verify on an example that if we threshold further by $S_\varepsilon$ the estimator obtained by the nonlinear Galerkin inversion of section 4, we still have a good estimator, as predicted by our Theorem 4.1. This suggests that the iterative adaptive Galerkin method described in Theorem 4.2 shall be effective when dealing with more precise numerical studies.

**A logarithmic potential integral operator.** We consider a single-layer logarithmic potential operator that relates the density of electric charge on an infinite cylinder of a given radius $r > 0$: $\{(z, re^{i2\pi x}), z \in \mathbb{R}, x \in [0,1]\}$ to the induced potential on the same cylinder, when both functions are independent of the variable $z$. The associated kernel $k(x,y)$ of the operator that we take is

$$Kf(x) = \int_0^1 k(x,y)f(y)dy, \;\; k(x,y) = -\log\big(r|e^{i2\pi x} - e^{i2\pi y}|\big) \tag{5.1}$$

for some $r > 0$. We choose $r = \frac{1}{4}$ and we rewrite (5.1) as

$$k(x,y) = -\log\big[\tfrac{1}{2}|\sin\pi(y-x)|\big], \;\; x, y \in [0,1], \tag{5.2}$$

so that $k(x,y) \geq 0$ on the unit square. It is singular on the diagonal $\{x = y\}$ but integrable. The single layer potential is known to be an elliptic operator of order $-1$, which maps $H^{-1/2}$ into $H^{1/2}$ (see [7]). So the assumptions on $K$ are satisfied with $t = 1$.

For the maximal resolution level $J \leq 15$ we discretize $K$ by computing the matrix $K_J$ with entries

$$(K_J)_{m,n=0,\dots,2^J-1} = (\langle K_J\varphi_{J,m}, \varphi_{J,n}\rangle)_{m,n=0,\dots,2^J-1},$$

where the $\varphi_{J,m} = 2^{J/2}1_{[m2^{-J},(m+1)2^{-J})}$ are the Haar functions. Each

$$\langle K_J\varphi_{J,m}, \varphi_{J,n}\rangle = \int_0^1\int_0^1 k(x,y)\varphi_{J,m}(x)\varphi_{J,n}(y)dxdy$$

is computed by midpoint rule at scale $2^{-18}$. It is noteworthy that $k$ is a periodic convolution kernel. In turn the discretization $K_{15}$ of $K$ is a Toeplitz cyclic matrix, of the form $K_J(m,n) = K_J\big((m-n)[\text{mod } 2^J]\big)$. As a consequence, the fast Fourier transform diagonalizes the matrix $K_J$, which makes the computation of its singular values an easy numerical task. We take $K_{15}$ as a proxy for $K$ and let the level of analysis of our method vary for $j = 1, \dots, 15$. We consider the test function $f$, defined for $x \in [0,1]$ by

$$f(x) = \max\{1 - |30(x - \tfrac{1}{2})|, 0\}.$$

The piky function $f$ is badly approximated by the singular functions of $K$, but it has a sparse representation in a wavelet basis, so the Galerkin method shall be more effective for the estimation problem.

**Methodology.** We first pick the maximal resolution level $J := 12$, a noise level $\varepsilon := 2 \cdot 10^{-4}$, and a single typical sample of a white noise process $w_{12} = (w_{k,12})_{k=0,\dots,2^{12}-1}$. This means that the $w_{k,12}$ are outcomes of independent identically distributed standard Gaussian random variables that contaminate the action of $K_{12}$ on $f$, up to the noise level $\varepsilon = 2 \cdot 10^{-4}$. Figure 1 shows the true signal $f$ (dash-dotted) together with the data process.

Let us recall that given a family of estimators $\hat{f}_j$ depending on a tuning constant $j = 1, \dots, j_{max}$ (here, $\hat{f}_j$ is constructed with the SVD or the linear Galerkin method, and $j$ varies from level 1 to level 12), the oracle estimator $\hat{f}^*$ is defined as $\hat{f}^* = \hat{f}_{j_*}$, where

$$j_* := \text{argmin}_{j=1,\dots,j_{max}}\|\hat{f}_j - f\|_{L^2}.$$

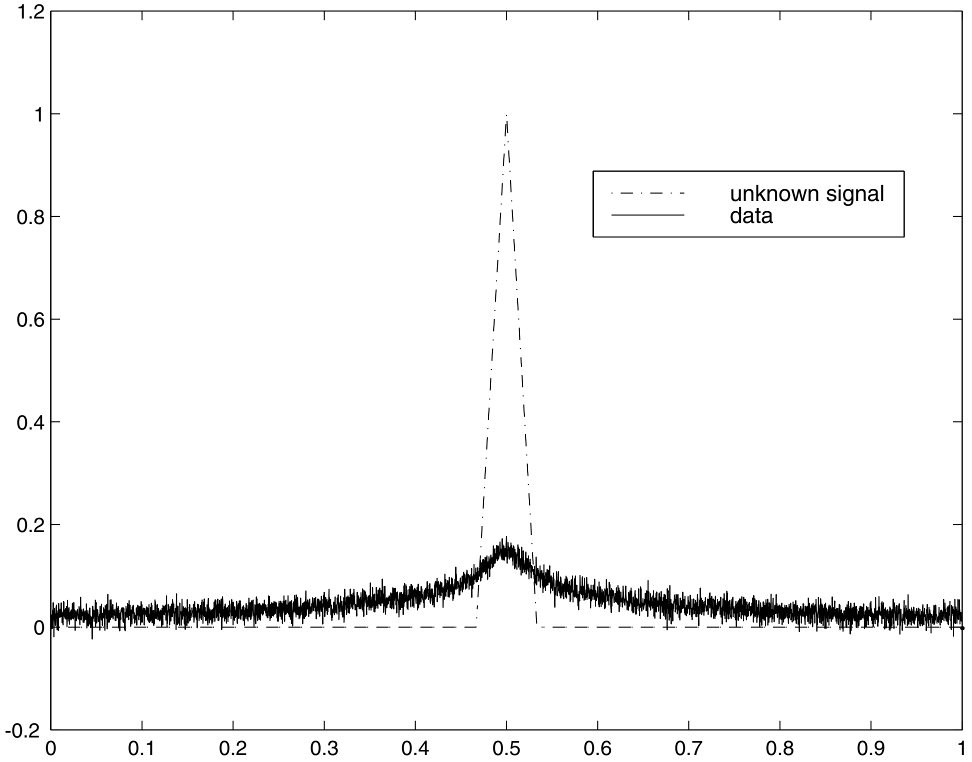

FIG. 1. *True function $f$ and noisy signal.*

TABLE 1

| | SVD oracle | | Linear Galerkin | | Nonlinear Galerkin |
|---|---|---|---|---|---|
| | $j_*$ | $L^2$-error | $j_*$ | $L^2$-error | $L^2$-error |
| $f$ | 5 | $5.66 \cdot 10^{-4}$ | 5 | $5.31 \cdot 10^{-4}$ | $3.75 \cdot 10^{-4}$ |

Note that, strictly speaking, $\hat{f}^*$ is not an estimator (the ideal level $j_*$ depends on the unknown) and appears as a benchmark for the method at hand. For technical reasons, we have replaced the $L^2$-risk by its empirical version.

**Numerical results.** The oracle estimator $\hat{f}^*$, for the SVD is displayed in Figure 2 and the oracle linear Galerkin estimator is displayed in Figure 3. In Figure 4, we show the performance of the nonlinear Galerkin estimator (Method 3), when applying a threshold $T_\varepsilon(\cdot)$ in the observation domain, specified with $t(\varepsilon) := 8 \cdot 10^{-4}$. We take a wavelet filter corresponding to compactly supported Daubechies wavelets of order 14. The numerical results of the three methods are summarized in Table 1.

**Compression rate and approximation results.** In the same context, we next investigate (see Figure 5) the performance of the nonlinear Galerkin estimator when applying further the level dependent thresholding operator $S_\varepsilon(\cdot)$, recall (4.4), with $t(\varepsilon) := 8 \cdot 10^{-4}$. We also indicate the number of wavelet coefficients put to zero divided by the total number of coefficients. The very high compression rate (see Table 2) that still ensures a small estimation error advocates in favor of the iterative adaptive scheme of section 4.

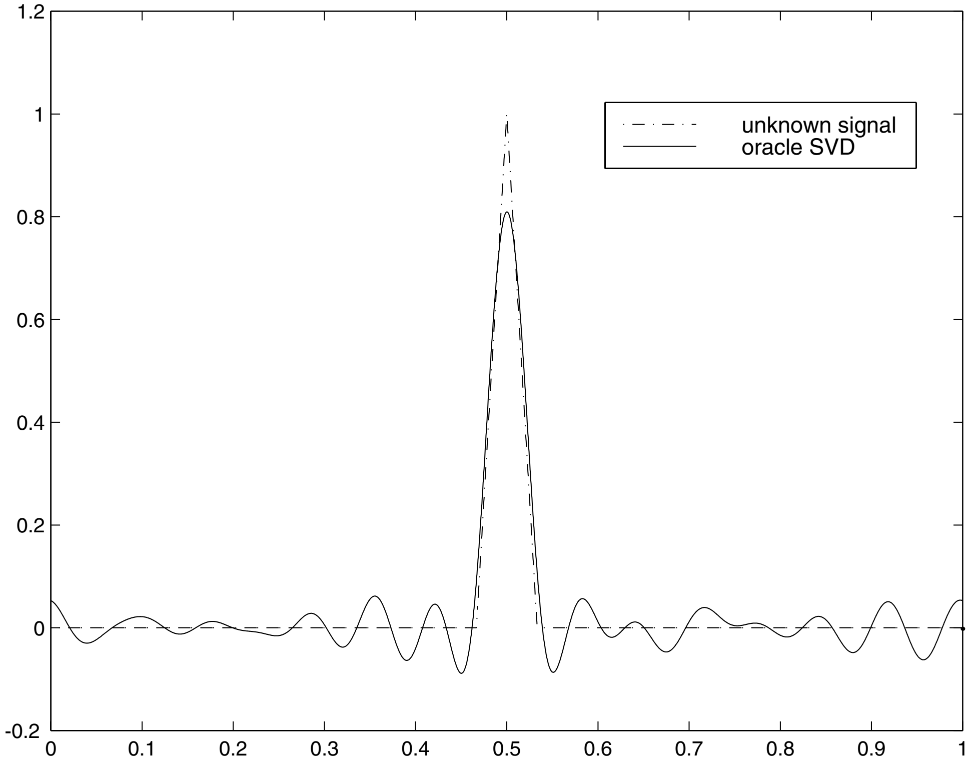

FIG. 2. *Oracle SVD.*

TABLE 2

| | $L^2$-error, $T_\varepsilon$ only | $L^2$-error, $T_\varepsilon$ and $S_\varepsilon$ | comp. rate |
|---|---|---|---|
| $f$ | $3.75 \cdot 10^{-4}$ | $4.12 \cdot 10^{-4}$ | 0.996 |

On a visual level, we observe that the nonlinear methods avoid the persistence of high oscillations far away from the singularity, in contrast to the linear methods. However, we still observe an artifact on the right side of the central peak. We hope to remedy this defect by (i) using biorthogonal spline wavelets instead of Daubechies orthogonal wavelets and (ii) apply a translation-invariant processing as introduced in [11] for denoising.

**Appendix A. Extension to nonelliptic operators.** In this appendix, we shall briefly explain how the methods and results that we have presented throughout can be extended by the mean-square approach to the case where $K$ is not an elliptic operator. Here, the smoothing property of order $t$ is expressed by the ellipticity property of the normal operator

$$\|Kf\|_{L^2}^2 = \langle K^*Kf, f\rangle \sim \|f\|_{H^{-t}}^2. \tag{A.1}$$

We discuss the adaptation of sections 3 and 4 to this more general context.

**Linear Galerkin estimation.** The method becomes the Galerkin projection

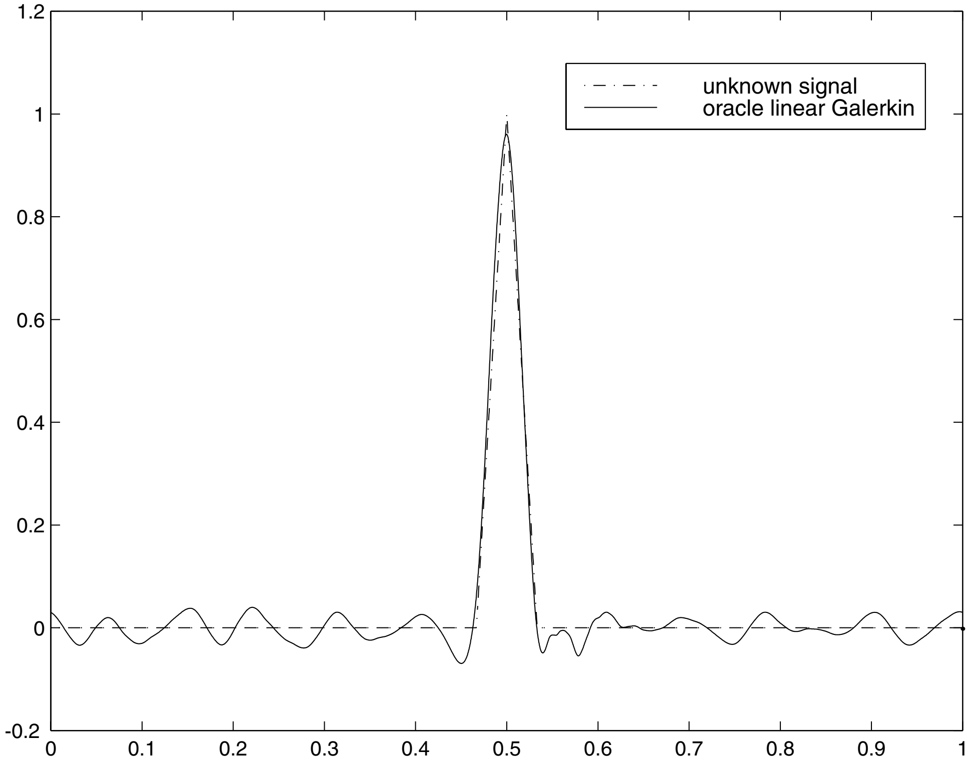

FIG. 3. *Oracle linear Galerkin method.*

method applied to the normal equation $K^*Kf = K^*g$. It therefore reads as follows:

$$\text{find } f_\varepsilon \in V_j \text{ such that } \langle Kf_\varepsilon, Kv\rangle = \langle g_\varepsilon, Kv\rangle \text{ for all } v \in V_j. \tag{A.2}$$

As in the elliptic case, we can use the decomposition $f_\varepsilon = f_j + h_\varepsilon$ in order to estimate the mean-square error according to

$$E(\|f - f_\varepsilon\|_{L^2}^2) \lesssim \|f - f_j\|_{L^2}^2 + E(\|h_\varepsilon\|_{L^2}^2). \tag{A.3}$$

For the variance term, we write

$$\begin{aligned} \|h_\varepsilon\|_{H^{-t}}^2 &\sim \langle K^*Kh_\varepsilon, h_\varepsilon\rangle = \varepsilon\langle \dot W, Kh_\varepsilon\rangle \\ &\le \varepsilon\|P_j^K \dot W\|_{L^2}\|Kh_\varepsilon\|_{L^2} \lesssim \varepsilon\|P_j^K \dot W\|_{L^2}\|h_\varepsilon\|_{H-t}, \end{aligned}$$

where $P_j^K$ is the orthogonal projector onto $KV_j$. Using the inverse inequality which states that $\|h_\varepsilon\|_{L^2} \lesssim 2^{tj}\|h_\varepsilon\|_{H^{-t}}$, we therefore obtain

$$\|h_\varepsilon\|_{L^2} \lesssim \varepsilon 2^{tj}\|P_j^K \dot W\|_{L^2}, \tag{A.4}$$

and therefore

$$E(\|h_\varepsilon\|_{L^2}^2) \lesssim \varepsilon^2 2^{2tj} \dim(KV_j) \lesssim \varepsilon^2 2^{(2t+d)j}. \tag{A.5}$$

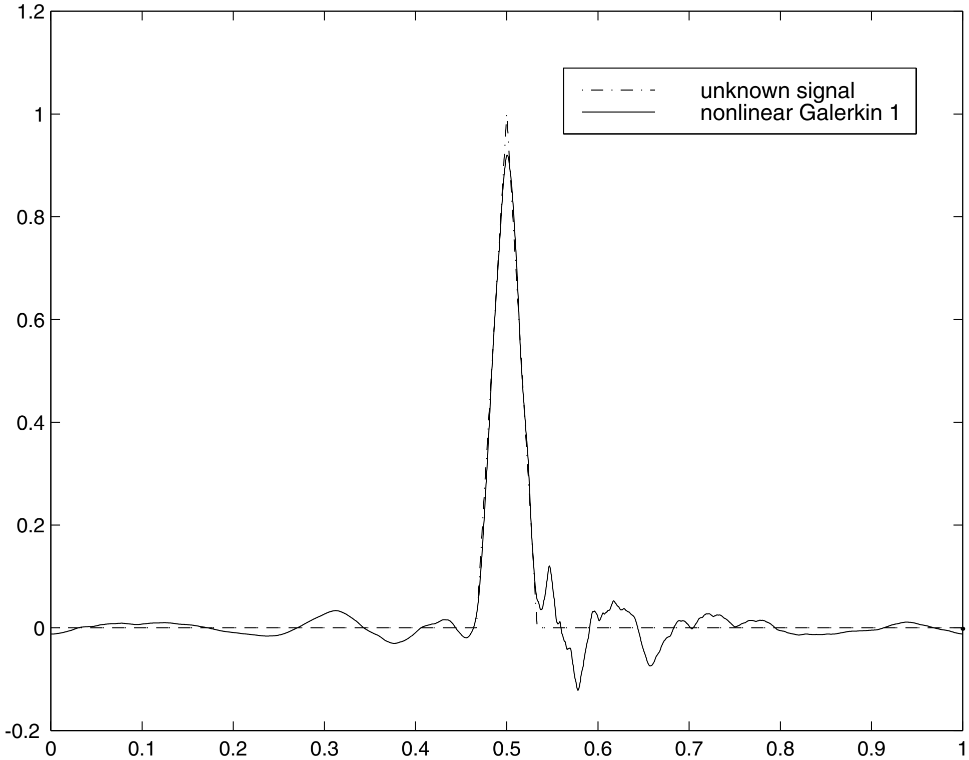

FIG. 4. *Nonlinear Galerkin estimator, thresholding in the image domain.*

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

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

where we have used the inverse inequality and Galerkin orthogonality. Assuming that $f$ belongs to a Sobolev ball $B := \{f \in H^s \ ; \ \|f\|_{H^s} \le M\}$, we obtain from approximation theory the direct estimate

$$\text{(A.6)} \qquad \|f - f_j\|_{L^2}^2 \lesssim \inf_{g_j \in V_j} [\|f - g_j\|_{L^2} + 2^{tj}\|f - g_j\|_{H^{-t}}] \lesssim 2^{-sj}.$$

Proceeding as in section 3, we therefore achieve the same estimate for $E(\|f - f_\varepsilon\|_{L^2}^2)$ under the same assumptions as in the elliptic case.

**Nonlinear estimation by linear Galerkin.** The method becomes the Galerkin projection applied to the normal equation after thresholding the observed data. It therefore reads as follows: find $f_\varepsilon = \sum_{|\lambda|<j} f_{\varepsilon,\lambda}\psi_\lambda \in V_j$ such that

$$\text{(A.7)} \qquad \langle K f_\varepsilon, K\psi_\lambda \rangle = \tilde{T}_\varepsilon(\langle g_\varepsilon, K\psi_\lambda \rangle)$$

for all $|\lambda| < j$. Here the thresholding operator $\tilde{T}_\varepsilon$ differs from $T_\varepsilon$ since it is applied to the wavelet coefficients of $K^* g_\varepsilon$. More precisely, we define

$$\text{(A.8)} \qquad \tilde{T}_\varepsilon(d_\lambda) = d_\lambda \chi(|d_\lambda| \ge 2^{-t|\lambda|} t(\varepsilon)) = 2^{-t|\lambda|} T_\varepsilon(2^{t|\lambda|} d_\lambda),$$

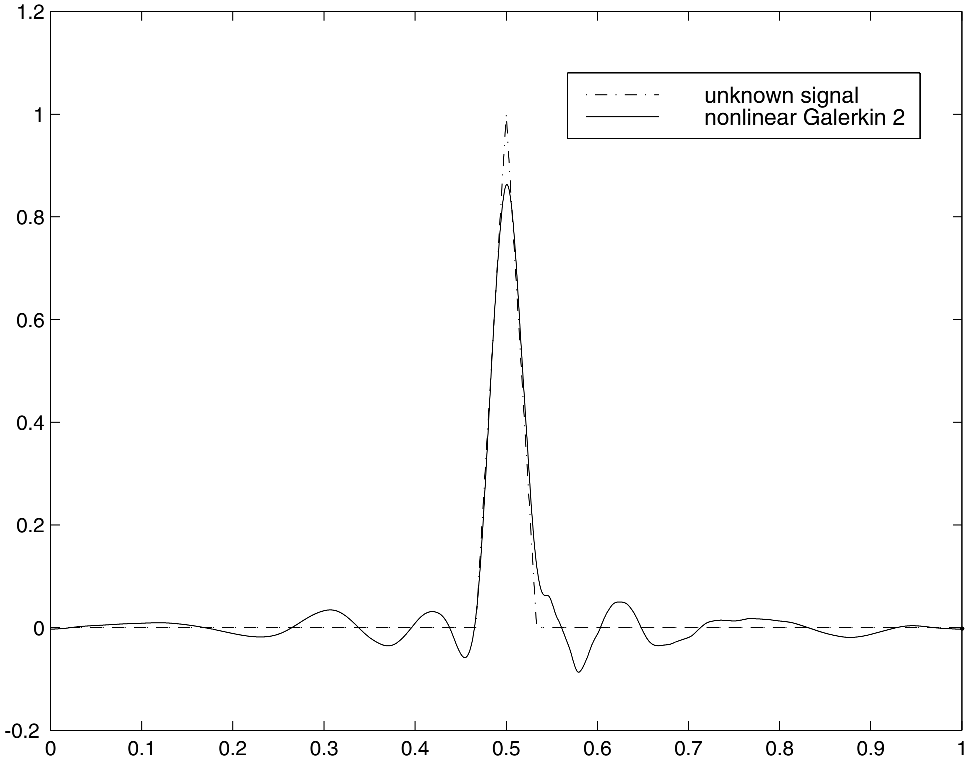

FIG. 5. *Nonlinear Galerkin estimator, thresholding in the image and solution domain.*

again with $t(\varepsilon) = C\varepsilon\sqrt{|\log \varepsilon|}$. With this method, Theorem 3.2 can be extended to the nonelliptic case, provided that we now use the assumption that $K^*K$ is an isomorphism from $L^2$ to $H^{2t}$ and has the smoothing property of order $2t$ with respect to the space $B^s_{p,p}$. For the proof of this result, we again write $f_\varepsilon = f_j + h_\varepsilon$, where the term $h_\varepsilon$ now represents the solution of the linear problem with the thresholding error as data. By the same analysis as in the proof of Theorem 3.2, we obtain

$$\|f - f_j\|^2_{L^2} \lesssim \varepsilon^{4s/(2s+2t+d)} \tag{A.9}$$

if $j$ is large enough so that $2^{-j} \leq \varepsilon^{1/t}$. For the stochastic term, if $H_\varepsilon$ is the coordinate vector of $h_\varepsilon$ in the basis $(\psi_\lambda)_{|\lambda|<j}$, we write

$$H_\varepsilon = L_j^{-1} D_j (T_\varepsilon(\tilde{G}_\varepsilon) - \tilde{G}_j), \tag{A.10}$$

where $L_j := (\langle K\psi_\lambda, K\psi_\mu\rangle)_{|\lambda|,|\mu|<j}$ is now the Galerkin matrix for the least-squares formulation, $\tilde{G}_j := (2^{t|\lambda|}\langle K^*g, \psi_\lambda\rangle)_{|\lambda|<j}$, $\tilde{G}_\varepsilon := (2^{t|\lambda|}\langle K^*g_\varepsilon, \psi_\lambda\rangle)_{|\lambda|<j}$, and again $D_j := \mathrm{Diag}(2^{-t|\lambda|})$. In this case, we invoke the stability property

$$\|L_j^{-1}U\|_{\ell^2} \lesssim \|U\|_{h^{2t}}, \tag{A.11}$$

which is proved by similar argument as in the proof of Theorem 3.2. Since $D_j$ is an i somorphism from $h^t$ to $h^{2t}$, we are therefore left to prove that

$$E(\|T_\varepsilon(\tilde{G}_\varepsilon) - \tilde{G}_j\|^2_{h^t}) \lesssim [\varepsilon\sqrt{|\log \varepsilon|}]^{4s/(2s+2t+d)}. \tag{A.12}$$

The components of $\tilde{G}_\varepsilon$ are related to those of $\tilde{G}_j$ by

$$\tilde{g}_{\varepsilon,\lambda} := 2^{t|\lambda|}\langle K^*Kf, \psi_\lambda\rangle + \varepsilon\langle \dot{W}, 2^{t|\lambda|}K\psi_\lambda\rangle = \tilde{g}_{j,\lambda} + \varepsilon\eta_\lambda, \tag{A.13}$$

where the $\eta_\lambda$ are normalized Gaussian variables since

$$\|2^{t|\lambda|}K\psi_\lambda\|_{L^2} \sim 2^{t|\lambda|}\|\psi_\lambda\|_{H^{-t}} \sim 1. \tag{A.14}$$

Therefore, the estimate (A.12) again follows from classical results on wavelet thresholding such as Theorem 4 in [10] using the fact that $K^*Kf$ belongs to a Besov ball $\tilde{B} = \{h \in B^{s+2t}_{p,p}\ ;\ \|h\|_{B^{s+2t}_{p,p}} \leq \tilde{M}\}$.

**Nonlinear estimation by adaptive Galerkin.** The iterative method becomes

$$F^n_\varepsilon = S_\varepsilon[F^{n-1}_\varepsilon + \tau D_j^{-1}(\tilde{T}_\varepsilon(G_\varepsilon) - L_jF^{n-1}_\varepsilon)], \tag{A.15}$$

and the statements of Theorems 4.1 and 4.2 remain valid with the same proof.

**Appendix B. Direct and inverse inequalities for multiresolution spaces.** Direct and inverse inequalities are a key ingredient in multiresolution approximation theory. In their simplest form, the direct inequality reads as follows:

$$\inf_{g_j\in V_j}\|f - g_j\|_{L^2} \lesssim 2^{-sj}|f|_{H^s}, \tag{B.1}$$

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

$$|g_j|_{H^s} \lesssim 2^{sj}\|g_j\|_{L^2}. \tag{B.2}$$

The proof of such estimates is quite classical and we refer the reader to Chapter 3 in [7]. Basically, the validity of the direct inequality requires that the spaces $V_j$ have enough approximation power, in the sense that polynomials of degree $m$ are contained in $V_j$ for all $m < s$. On the other hand, the validity of the inverse estimate requires that the functions of $V_j$ have enough smoothness in the sense that they are contained in $H^s$. The direct and inverse estimate which have been used in section 3 are less standard since they involve the Sobolev space of negative order $H^{-t/2}$, and we shall therefore briefly discuss their validity. The inverse estimate states that for all $g_j \in V_j$,

$$\|g_j\|_{L^2} \lesssim 2^{tj/2}\|g_j\|_{H^{-t/2}}. \tag{B.3}$$

We prove it by a duality argument:

$$\begin{aligned}\|g_j\|_{L^2} &= \sup\nolimits_{f_j\in V_j, \|f_j\|_{L^2}=1} |\langle g_j, f_j\rangle| \\ &\lesssim \sup\nolimits_{f_j\in V_j, \|f_j\|_{L^2}=1} \|g_j\|_{H^{-t/2}}\|f_j\|_{H^{t/2}} \\ &\lesssim 2^{tj/2}\|g_j\|_{H^{-t/2}},\end{aligned}$$

where we have used the standard inverse estimate (B.2) with $s = t/2$. The direct estimate states that

$$\inf_{g_j\in V_j}[\|f - g_j\|_{L^2} + 2^{tj/2}\|f - g_j\|_{H^{-t/2}}] \lesssim 2^{-sj}\|f\|_{H^s}. \tag{B.4}$$

In order to prove it, we take $g_j = P_jf$ where $P_j$ is the $L^2$-orthogonal projector onto $V_j$. Clearly the first part $\|f - P_jf\|_{L^2} \lesssim 2^{-sj}\|f\|_{H^s}$ is simply the standard direct estimate (B.1). For the second part, we write

$$\begin{aligned}\|f-P_jf\|_{H^{-t/2}} &= \sup_{\|g\|_{H^{t/2}}=1} |\langle f-P_jf,g\rangle| \\ &= \sup_{\|g\|_{H^{t/2}}=1} |\langle f-P_jf,g-P_jg\rangle| \\ &= \|f-P_jf\|_{L^2}\sup_{\|g\|_{H^{t/2}}=1}\|g-P_jg\|_{L^2} \\ &\sim 2^{-sj}\|f\|_{H^s}2^{-tj/2},\end{aligned}$$

where we have used the fact that $(I-P_j)^2=(I-P_j)^*(I-P_j)=I-P_j$ and the standard direct estimate (B.1) both for $H^s$ and $H^{t/2}$.

*Remark.* The type of duality argument that we have used in order to prove both (B.3) and (B.4) can be generalized in such a way that the standard direct and inverse estimate between $L^2$ and $H^{t/2}$ are invoked for a dual space $\tilde{V}_j$ which might differ from $V_j$. For the direct estimate, this means that we take for $P_j$ a more general biorthogonal projector, such that $P_j^*$ is a projector onto $\tilde{V}_j$ (see [7] for examples of dual spaces and biorthogonal projectors), so that we are led to apply a standard direct inequality of the type

$$\|g-P_j^*g\|_{L^2} \lesssim 2^{-tj/2}\|g\|_{H^{t/2}} \tag{B.5}$$

which only requires polynomial exactness up to order $t/2$ for $\tilde{V}_j$. For the inverse estimate, we can also use the space $\tilde{V}_j$ in order to evaluate the $L^2$ norm according to

$$\|g_j\|_{L^2} \lesssim \sup_{\tilde{f}_j\in\tilde{V}_j,\|\tilde{f}_j\|_{L^2}=1} |\langle g_j,\tilde{f}_j\rangle| \lesssim \sup_{\tilde{f}_j\in\tilde{V}_j,\|\tilde{f}_j\|_{L^2}=1} \|g_j\|_{H^{-t/2}}\|\tilde{f}_j\|_{H^{t/2}}, \tag{B.6}$$

so that we are led to apply a standard inverse inequality of the type

$$\|\tilde{f}_j\|_{H^{t/2}} \lesssim 2^{tj/2}\|\tilde{f}_j\|_{L^2}, \tag{B.7}$$

which only requires that the space $\tilde{V}_j$ has $H^{t/2}$ smoothness. This last point is practically important, since it means that we are not enforced to use multiresolution spaces $V_j$ consisting of smooth functions, neither are we forced to use smooth wavelets in the nonlinear methods. In contrast, it is crucial that the spaces $V_j$ have enough polynomial reproduction ($\Pi_m \subset V_j$ for all $m<s$) in order to apply the direct estimate for $H^s$, and it is crucial for the nonlinear method that the wavelets $\psi_\lambda$ have enough vanishing moments ($\int x^m\psi_\lambda=0$ for all $m<s+t$) in order to apply the results on wavelet thresholding such as (3.30).

## REFERENCES

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

[2] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms*, I, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[3] J.H. Bramble, *Multigrid Methods*, Longman Scientific and Technical, Harlow, UK, 1993.

[4] L. Brown, T. Cai, M. Low, and C.H. Zhang, *Asymptotic equivalence theory for nonparametric regression with random design*, Ann. Statist., 30 (2002), pp. 688–707.

[5] S. Champier and L. Grammont, *A wavelet-vaguelet method for unfolding sphere size distributions*, Inverse Problems, 18 (2002), pp. 79–94.

[6] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, New York, Oxford, 1978.

[7] A. Cohen, *Wavelet methods in numerical analysis*, in Handbook of Numerical Analysis, Vol. VII, P.G. Ciarlet and J.L. Lions, eds., Elsevier, Amsterdam, 2000, pp. 417–711.

[8] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations: Convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

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

[10] A. Cohen, R. DeVore, G. Kerkyacharian, and D. Picard, *Maximal spaces with given rate of convergence for thresholding algorithms*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 167–191.

[11] R.R. Coifman and D.L. Donoho, *Translation-invariant de-noising*, in Wavelets and Statistics, Lecture Notes in Statist. 103, A. Antoniadis and G. Oppenheim, eds., Springer, New York, 1995, pp. 125–150.

[12] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[13] I. Daubechies, M. Defrise, and C. De Mol, *An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint*, preprint, Princeton University, Princeton, NJ, 2003.

[14] R. DeVore, *Nonlinear approximation*, Acta Numerica, 7 (1998), pp. 51–150.

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

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

[17] D.L. Donoho and I.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.

[18] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard, *Wavelet shrinkage: Asymptopia?*, J. Roy. Statist. Soc. Ser. B, 57 (1995), pp. 301–369.

[19] H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Press, Dordrecht, The Netherlands, 1996.

[20] T. Hida, *Brownian Motion*, Springer-Verlag, New York, 1980.

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

[22] I.M. Johnstone and B.W. Silverman, *Discretization effects in statistical inverse problems*, J. Complexity, 7 (1991), pp. 1–34.

[23] A. Korostelev and A. Tsybakov, *Minimax Theory of Image Reconstruction*, Lecture Notes in Statist. 82, Springer-Verlag, New York, 1993.

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

[25] F. Natterer, *The Mathematics of Computerized Tomography*, Classics Appl. Math. 32, SIAM, Philadelphia, 2001.

[26] M. Nussbaum and S. Pereverzev, *The Degree of Ill-Posedness in Stochastic and Deterministic Models*, Preprint 509, Weierstraß-Institut, Berlin, 1999.

[27] M. Reiß, *Minimax rates for nonparametric drift estimation in affine stochastic delay differential equations*, Stat. Inference Stoch. Process., 5 (2002), pp. 131–152.