MATHEMATICS OF COMPUTATION
Volume 81, Number 277, January 2012, Pages 419–446
S 0025-5718(2011)02507-X
Article electronically published on May 25, 2011

MULTILEVEL PRECONDITIONING AND ADAPTIVE SPARSE SOLUTION OF INVERSE PROBLEMS

STEPHAN DAHLKE, MASSIMO FORNASIER, AND THORSTEN RAASCH

Abstract. We are concerned with the efficient numerical solution of minimization problems in Hilbert spaces involving sparsity constraints. These optimizations arise, e.g., in the context of inverse problems. In this work we analyze an efficient variant of the well-known iterative soft-shrinkage algorithm for large or even infinite dimensional problems. This algorithm is modified in the following way. Instead of prescribing a fixed thresholding parameter, we use a decreasing thresholding strategy. Moreover, we use suitable variants of the adaptive schemes derived by Cohen, Dahmen and DeVore for the approximation of the infinite matrix-vector products. We derive a block multiscale preconditioning technique which allows for local well-conditioning of the underlying matrices and for extending the concept of *restricted isometry property* to infinitely labelled matrices. The combination of these ingredients gives rise to a numerical scheme that is guaranteed to converge with exponential rate, and which allows for a controlled inflation of the support size of the iterations. We also present numerical experiments that confirm the applicability of our approach which extends concepts from *compressed sensing* to large scale simulation.

1. Introduction

In this paper we are concerned with the efficient minimization of functionals of the type

$$J(\mathbf{u}) := J_\alpha(\mathbf{u}) := \left\|(K \circ F)\mathbf{u} - y\right\|_Y^2 + 2\|\mathbf{u}\|_{\ell_{1,\alpha}(\mathcal{I})}, \tag{1.1}$$

where $K : X \to Y$ is a bounded linear operator acting between two separable Hilbert spaces X and Y, $y \in Y$ is a given datum and $F : \ell_2(\mathcal{I}) \ni \mathbf{u} \mapsto \sum_{\lambda\in\mathcal{I}} u_\lambda \psi_\lambda$ is the synthesis operator of a prescribed countable basis or a tight frame $\Psi := (\psi_\lambda)_{\lambda\in\mathcal{I}}$ for X. For $1 \le p < \infty$, the sequence norm $\|\mathbf{u}\|_{\ell_{p,\alpha}(\mathcal{I})} := \left(\sum_{\lambda\in\mathcal{I}} |u_\lambda|^p \alpha_\lambda\right)^{1/p}$ is the

Received by the editor August 14, 2009 and, in revised form, November 18, 2010.
2010 *Mathematics Subject Classification.* Primary 41A25, 65F35, 65F50, 65N12, 65T60.
Key words and phrases. Operator equations, multiscale methods, adaptive algorithms, sparse optimization and compressed sensing, multilevel preconditioning, wavelets, iterative thresholding, restricted isometry property.
The work of Stephan Dahlke was supported by Deutsche Forschungsgemeinschaft (DFG), Grant DA 360/12–1.
The work of Massimo Fornasier was supported by the FWF project Y 432-N15 START-Preis "Sparse Approximation and Optimization in High Dimensions" and Deutsche Forschungsgemeinschaft (DFG), Grant DA 360/12–1.
The work of Thorsten Raasch was supported by Deutsche Forschungsgemeinschaft (DFG), Grants DA 360/7–1 and DA 360/11–1.

usual norm for weighted p-summable sequences, with weight $\alpha = (\alpha_\lambda)_{\lambda\in\mathcal{I}} \in \mathbb{R}_+^{\mathcal{I}}$. We denote $A := K \circ F$.

Functionals of the type (1.1) may arise in the numerical treatment of linear inverse problems

$$y = Ku + e, \tag{1.2}$$

where the data y may be corrupted by unknown noise e and u is expanded in the basis Ψ. It was shown in [17] that penalization by the ℓ_1-norm term in the functional J corresponds to a regularization scheme [20], and it promotes sparse minimal solutions.

Our approach to approximate a minimizer $\mathbf{u}^*$ of J is based on its characterization by means of the fixed point equation

$$\mathbf{u}^* = \mathbb{S}_\alpha\big[\mathbf{u}^* + A^*y - A^*A\mathbf{u}^*\big], \tag{1.3}$$

which involves the *soft thresholding* operator defined componentwise by $\mathbb{S}_\alpha(\mathbf{u})_\lambda = S_{\alpha_\lambda}(u_\lambda)$ and

$$S_\tau(x) = \begin{cases} x-\tau & x > \tau, \\ 0 & |x| \le \tau, \\ x+\tau & x < -\tau. \end{cases} \tag{1.4}$$

In fact, several authors have independently proposed to use the associated fixed point iteration

$$\mathbf{u}^{(n+1)} = \mathbb{S}_\alpha\big[\mathbf{u}^{(n)} + A^*y - A^*A\mathbf{u}^{(n)}\big] \tag{1.5}$$

to approximate minimizers $\mathbf{u}^*$ of J (see [22, 36, 37, 19]), which is called the *iterative soft-thresholding algorithm* or the *thresholded Landweber iteration* (ISTA). This is our starting point and the reference iteration on which we want to work out several improvements. Strong convergence of this algorithm was proved in [17], under the assumption that $\|A\| < 1$ (actually, convergence can be shown also for $\|A\| < \sqrt{2}$ [9]; nevertheless, the condition $\|A\| < \sqrt{2}$ is by no means a restriction, since it can always be met by a suitable rescaling of the functional J, in particular, of K, y, and α). However, in several concrete circumstances, ISTA may converge slowly in practice. Therefore, in order to end up with a fast recovery algorithm, it is helpful to apply certain modifications. In this paper, we propose to apply a combination of the following three key strategies:

(i) *multilevel preconditioning*;
(ii) *decreasing thresholding parameters*;
(iii) *adaptivity*.

1.1. **Rate of convergence and local preconditioning.** Let us start by addressing (i) *multilevel preconditioning*. Concerning the qualitative convergence properties of iterative soft-thresholding, note first that the necessary condition $\|A\| < \sqrt{2}$ does not guarantee contractivity of the iteration operator $I - A^*A$, since A^*A may not be boundedly invertible. The insertion of $\mathbb{S}_\alpha$ does not improve the situation since $\mathbb{S}_\alpha$ is only nonexpansive,

$$\big\|\mathbb{S}_\alpha(\mathbf{v}) - \mathbb{S}_\alpha(\mathbf{w})\big\|_{\ell_2(\mathcal{I})} \le \|\mathbf{v}-\mathbf{w}\|_{\ell_2(\mathcal{I})}, \quad \text{for all } \mathbf{v},\mathbf{w} \in \ell_2(\mathcal{I}). \tag{1.6}$$

Hence, for any minimizer $\mathbf{u}^*$, the estimate

$$\|\mathbf{u}^*-\mathbf{u}^{(n+1)}\|_{\ell_2(\mathcal{I})} \leq \left\|(I-A^*A)(\mathbf{u}^*-\mathbf{u}^{(n)})\right\|_{\ell_2(\mathcal{I})} \leq \|I-A^*A\|\|\mathbf{u}^*-\mathbf{u}^{(n)}\|_{\ell_2(\mathcal{I})} \tag{1.7}$$

does not give rise to linear error reduction. However, under additional assumptions on the operator A or on the minimizers $\mathbf{u}^*$, linear convergence of (1.5) can be ensured. In particular, if A fulfills the so-called *finite basis injectivity* (FBI) condition (see [3] where this terminology is introduced), i.e., for any finite set $\Lambda \subset \mathcal{I}$, the restriction A_Λ is injective, then (1.5) converges linearly to a minimizer $\mathbf{u}^*$ of J. Here A_Λ shall denote the submatrix extracted from A by retaining only the columns indexed in Λ. We also denote $A^*A|_{\Lambda\times\Lambda} := A_\Lambda^* A_\Lambda$, the square submatrix extracting from A only the entries indexed on $\Lambda \times \Lambda$.

However, depending on the active set Λ, the matrix $A^*A|_{\Lambda\times\Lambda}$ can be arbitrarily badly conditioned, which would imply contraction rates γ_Λ close to 1 in the error reduction $\|\mathbf{u}^* - \mathbf{u}^{(n+1)}\|_{\ell_2(\mathcal{I})} \leq \gamma_\Lambda \|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2(\mathcal{I})}$.

Our first improvement and contribution of this paper is to show that for several FBI operators K and for certain choices of Ψ, the matrix A^*A can be preconditioned by a matrix $D^{-1/2}$, resulting in the matrix $D^{-1/2}A^*AD^{-1/2}$, in such a way that any restriction $(D^{-1/2}A^*AD^{-1/2})_{\Lambda\times\Lambda}$ turns out to be well-conditioned as soon as $\Lambda \subset \mathcal{I}$ is a small set, but independently of its "location" within $\mathcal{I}$.

As a typical example, we consider injective (nonlocal) compact operators K with Schwartz kernel having certain polynomial decay properties of the derivatives, i.e.,

$$Ku(x) = \int_\Omega \Phi(x,\xi)u(\xi)d\xi, \quad x \in \tilde{\Omega},$$

for $\tilde{\Omega}, \Omega \subset \mathbb{R}^d$, $u \in X := H^t(\Omega)$, and

$$|\partial_x^\alpha \partial_\xi^\beta \Phi(x,\xi)| \leq c_{\alpha,\beta}|x-\xi|^{-(d+2t+|\alpha|+|\beta|)}, \quad t \in \mathbb{R}, \text{ and multi-indices } \alpha, \beta \in \mathbb{N}^d.$$

Moreover, for the proper definition of the discrete matrix $A^*A := F^*K^*KF$, we will show that multiscale bases Ψ, such as wavelets, do a good job for us [14, 13, 15].

1.2. **Iterative soft-thresholding with decreasing thresholding parameter.** We consider now (ii) *decreasing thresholding parameters*. In view of such local well-conditioning, it becomes obvious that iterating on small sets Λ will also improve the convergence rate. Unfortunately, the thresholded Landweber iteration (1.5) does not iterate in general on small sets, but rather it starts iterating on relatively large sets, slowly shrinking to the size of the support of the limit $\mathbf{u}^*$; see Figure 6 and Figure 7 below.

The second improvement and contribution of this paper is the proposal and the proof of convergence of an algorithm which starts with large thresholding parameters $\alpha^{(n)}$ and geometrically reduces them during the iterations to a target limit $\alpha > 0$,

$$\mathbf{u}^{(n+1)} = \mathbb{S}_{\alpha^{(n)}}\left[\mathbf{u}^{(n)} + A^*y - A^*A\mathbf{u}^{(n)}\right]. \tag{1.8}$$

This strategy promotes small supports of initial iterates, which are inflated during the process. For matrices A for which the restrictions $A^*A|_{\Lambda\times\Lambda}$ are uniformly well-conditioned with respect to Λ of small size, our analysis provides also a prescribed linear rate of convergence of the iteration (1.8). The strategy of choosing slowly decreasing thresholding parameters is not at all new. Several other contributions, for example [21], proposed similar approaches. However, none of them quantifies rigorously the convergence improvement by relating the rate of convergence with the local conditioning of A^*A.

1.3. **Adaptivity.** Finally, we address (iii) *adaptivity*. In the case that A maps into an infinite-dimensional space Y, e.g., when A is the forward solution operator from a partial differential equation, the iteration (1.8) will not be exactly implementable. Instead, one has to replace the exact applications of A and A^* by suitable numerical approximations. In particular, we assume that for any finitely supported vector $\mathbf{w} \in \ell_2(\mathcal{I})$, we are able to compute an approximation of $A^*A\mathbf{w}$. For this, we will rely on the existence of a numerical routine $\mathbf{APPLY}[A^*A, \mathbf{w}, \epsilon]$ which, for given input data $\mathbf{w}$ and $\epsilon > 0$, outputs a finitely supported vector $\mathbf{w}_\epsilon$ such that $\|A^*A\mathbf{w} - \mathbf{w}_\epsilon\|_{\ell_2(\mathcal{I})} \leq \epsilon$. We refer, e.g., to [7, 28, 38], for concrete examples of such a procedure, when the matrix A^*A can be assumed compressible [13, §9.5]; see Section 4. Moreover, we assume that there exists a routine $\mathbf{RHS}[A^*y, \epsilon]$ which, for given input data ϵ, outputs a finitely supported vector $\mathbf{r}_\epsilon$ such that $\|A^*y - \mathbf{r}_\epsilon\|_{\ell_2(\mathcal{I})} \leq \epsilon$.

Instead of (1.8), we will consider then the numerically implementable algorithm

$$\tilde{\mathbf{u}}^{(0)} = \mathbf{0}, \tilde{\mathbf{u}}^{(n+1)} = \mathbb{S}_{\alpha^{(n)}}\big(\tilde{\mathbf{u}}^{(n)} + \mathbf{RHS}[A^*y, \delta_n] - \mathbf{APPLY}[A^*A, \tilde{\mathbf{u}}^{(n)}, \xi_n]\big), \quad n = 0, 1, \ldots, \tag{1.9}$$

with certain tolerances δ_n and ξ_n. Such an *adaptive* scheme was first proposed in the context of sparse optimization and inverse problems in the paper [23], "borrowing a leaf" from the analysis of adaptive schemes for well-posed problems [7, 8, 10, 12, 38]; in the present work we would like to extend those preliminary results and make them more rigorous. More recent contributions, e.g., [2, 33], also adopted this idea and proposed similar approaches.

Our third contribution is the proof of linear convergence of (1.9) when the matrix A^*A has restrictions $A^*A|_{\Lambda\times\Lambda}$ which are uniformly well-conditioned with respect to small active sets Λ.

1.4. **Outline of the paper.** In Section 2 we collect a few technical results on sparse and compressible vectors as well as on properties of the soft-thresholding operator. Section 3 is dedicated to the analysis of the decreasing iterative soft-thresholding algorithm when both exact and adaptive matrix-vector multiplications are involved and local well-conditioning of the matrix A^*A is assumed. In Section 4 we address the issue of multiscale preconditioning and we show that for a large class of operators it is possible to choose a multiscale basis for which the resulting discretization matrices can be locally preconditioned by simple block-diagonal preconditioners. In Section 5 we clarify how preconditioning can be used in ill-posed problems and how we can take advantage of the sparsity of minimal solutions and of adaptive matrix-vector multiplications in order to avoid topological issues. Since our results on the efficiency of the proposed strategy depend on the sparsity of the minimal solution, in Section 6 we quantify how the support size of minimal solutions may increase as the thresholding parameter α is made smaller and smaller. We conclude the paper with Section 7 where we show numerical results which demonstrate and confirm the theoretical achievements of the previous sections.

This paper is significantly reduced with respect to its original preprint version [11], and we refer the interested reader to it for more details and a broader discussion.

2. Technical lemmas

We are particularly interested in computing approximations with the smallest possible number of nonzero entries. As a benchmark, we recall that the most economical approximations of a given vector $\mathbf{v} \in \ell_2(\mathcal{I})$ are provided by the *best N-term approximations* $\mathbf{v}_N$, defined by discarding in $\mathbf{v}$ all but the $N \in \mathbb{N}_0$ largest coefficients in absolute value. The error of best N-term approximation is defined as

$$\sigma_N(\mathbf{v}) := \|\mathbf{v} - \mathbf{v}_N\|_{\ell_2(\mathcal{I})}. \tag{2.1}$$

The subspace of all ℓ_2 vectors with best N-term approximation rate $s > 0$, i.e., $\sigma_N(\mathbf{v}) \lesssim N^{-s}$ for some decay rate $s > 0$, is commonly referred to as the *weak ℓ_τ space* $\ell_\tau^w(\mathcal{I})$, for $\tau = (s + \frac{1}{2})^{-1}$, which, endowed with

$$\|\mathbf{v}\|_{\ell_\tau^w(\mathcal{I})} := \sup_{N \in \mathbb{N}_0} (N+1)^s \sigma_N(\mathbf{v}), \tag{2.2}$$

becomes the quasi-Banach space $(\ell_\tau^w(\mathcal{I}), \|\cdot\|_{\ell_\tau^w(\mathcal{I})})$. Moreover, for any $0 < \epsilon \leq 2-\tau$, we have the continuous embedding $\ell_\tau(\mathcal{I}) \hookrightarrow \ell_\tau^w(\mathcal{I}) \hookrightarrow \ell_{\tau+\epsilon}(\mathcal{I})$, justifying why $\ell_\tau^w(\mathcal{I})$ is called weak $\ell_\tau(\mathcal{I})$ (note that $\tau < 2$ by definition).

When it comes to the concrete computations of good approximations with a small number of active coefficients, one frequently utilizes certain thresholding procedures. Here small entries of a given vector are simply discarded, whereas the large entries may be slightly modified. In this paper, we will make use of *soft-thresholding* that we already introduced in (1.4). It is well known (see [17]), that $\mathbb{S}_\alpha$ is non-expansive for any $\alpha \in \mathbb{R}_+^{\mathcal{I}}$. Moreover, for any fixed $x \in \mathbb{R}$, the mapping $\tau \mapsto S_\tau(x)$ is Lipschitz continuous with

$$\big|S_\tau(x) - S_{\tau'}(x)\big| \leq |\tau - \tau'|, \quad \text{for all } \tau, \tau' \geq 0. \tag{2.3}$$

We readily infer the following technical estimate (for the detailed proof we refer the reader to [11], which is an extended preprint version of this paper).

Lemma 2.1. *Assume $\mathbf{v} \in \ell_2(\mathcal{I})$, $\alpha, \beta \in \mathbb{R}_+^{\mathcal{I}}$ such that $\bar{\alpha} = \inf_\lambda \alpha_\lambda = \inf_\lambda \beta_\lambda = \bar{\beta} > 0$, and define $\Lambda_{\bar{\alpha}}(\mathbf{v}) := \big\{\lambda \in \mathcal{I} : |v_\lambda| > \bar{\alpha}\big\}$. Then*

$$\big\|\mathbb{S}_\alpha(\mathbf{v}) - \mathbb{S}_\beta(\mathbf{v})\big\|_{\ell_2(\mathcal{I})} \leq \Big(\#\Lambda_{\bar{\alpha}}(\mathbf{v})\Big)^{1/2} \max_{\lambda \in \Lambda_{\bar{\alpha}}(\mathbf{v})} |\alpha_\lambda - \beta_\lambda|. \tag{2.4}$$

Let $\mathbf{v} \in \ell_\tau^w(\mathcal{I})$, it is well known [18, §7] that

$$\#\Lambda_{\bar{\alpha}}(\mathbf{v}) \leq C(\tau) \|\mathbf{v}\|_{\ell_\tau^w(\mathcal{I})}^\tau \bar{\alpha}^{-\tau}, \tag{2.5}$$

and, for $\alpha_\lambda = \bar{\alpha}$ for all $\lambda \in \mathcal{I}$, we have

$$\big\|\mathbf{v} - \mathbb{S}_\alpha(\mathbf{v})\big\|_{\ell_2(\mathcal{I})} \leq C(\tau) \|\mathbf{v}\|_{\ell_\tau^w(\mathcal{I})}^{\tau/2} \bar{\alpha}^{1-\tau/2}, \tag{2.6}$$

where $C(\tau) > 0$ may depend on τ. Let $\mathbf{v} \in \ell_0(\mathcal{I}) := \bigcap_{\tau > 0} \ell_\tau^w(\mathcal{I})$ the set of finitely supported vectors, and $\|\mathbf{v}\|_{\ell_0} := \#\operatorname{supp} \mathbf{v} < \infty$. Then we have the straightforward estimate

$$\#\Lambda_{\bar{\alpha}}(\mathbf{v}) \leq \|\mathbf{v}\|_{\ell_0} \tag{2.7}$$

and, for $\alpha_\lambda = \bar{\alpha}$ for all $\lambda \in \mathcal{I}$, we have

$$\big\|\mathbf{v} - \mathbb{S}_\alpha(\mathbf{v})\big\|_{\ell_2(\mathcal{I})} \leq \|\mathbf{v}\|_{\ell_0}^{1/2} \bar{\alpha}, \tag{2.8}$$

which is easily shown by a direct computation. In the sequel, we shall also use the following support size estimate, whose proof follows the lines of [7, Lemma 5.1], more details are provided in [11].

Lemma 2.2. *Let* $\mathbf{v} \in \ell_\tau^w(\mathcal{I})$ *and* $\mathbf{w} \in \ell_2(\mathcal{I})$ *with* $\|\mathbf{v}-\mathbf{w}\|_{\ell_2(\mathcal{I})} \leq \epsilon$. *Assume* $\alpha = (\alpha_\lambda)_{\lambda\in\mathcal{I}} \in \mathbb{R}_+^{\mathcal{I}}$ *and* $\inf_\lambda \alpha_\lambda = \bar{\alpha} > 0$. *Then it holds that*

$$\# \operatorname{supp} \mathbb{S}_\alpha(\mathbf{w}) \leq \#\Lambda_{\bar{\alpha}}(\mathbf{w}) \leq \frac{4\epsilon^2}{\bar{\alpha}^2} + 4C\|\mathbf{v}\|_{\ell_\tau^w(\mathcal{I})}^\tau \bar{\alpha}^{-\tau}, \tag{2.9}$$

where $C = C(\tau) > 0$. *In particular, if* $\mathbf{v} \in \ell_0(\mathcal{I})$, *then the estimate is refined:*

$$\# \operatorname{supp} \mathbb{S}_\alpha(\mathbf{w}) \leq \#\Lambda_{\bar{\alpha}}(\mathbf{w}) \leq \frac{4\epsilon^2}{\bar{\alpha}^2} + \|\mathbf{v}\|_{\ell_0(\mathcal{I})}. \tag{2.10}$$

3. Adaptive iterative soft-thresholding

In this section we address first the analysis of the iteration (1.8) where exact matrix-vector multiplications involving A^*A are assumed to be realizable. In order to ensure a linear rate of convergence of the algorithm to a minimizer $\mathbf{u}^*$, we will use spectral conditions on A^*A, the so-called *Restricted Isometry Property*, a well-known concept from *compressed sensing* problems [4, 5]. In the second part of the section, we address the problem of the use of such iteration when inexact applications of the matrix A^*A are necessary and we provide again a detailed analysis of the rate of convergence.

3.1. A variant of iterative soft-thresholding.

In the case of exact operator evaluations and threshold parameters $\alpha, \alpha^{(n)} \in \mathbb{R}_+^{\mathcal{I}}$, where $\alpha^{(n)} \geq \alpha$, i.e., $\alpha_\lambda^{(n)} \geq \alpha_\lambda$ for all $\lambda \in \mathcal{I}$, and $\bar{\alpha} = \inf_{\lambda\in\mathcal{I}} \alpha_\lambda > 0$, we consider the iteration

$$\mathbf{u}^{(0)} = \mathbf{0}, \quad \mathbf{u}^{(n+1)} = \mathbb{S}_{\alpha^{(n)}}\left[\mathbf{u}^{(n)} + A^*(y - A\mathbf{u}^{(n)})\right], \quad n = 0, 1, \ldots \tag{3.1}$$

which we call, as $\alpha_\lambda^{(n)} \geq \alpha_\lambda^{(n+1)}$ for all $\lambda \in \mathcal{I}$, the *decreasing iterative soft-thresholding algorithm* (D-ISTA).

For obtaining a controlled liner rate of convergence, we shall need the following spectral condition on a bounded linear operator $A : \ell_2(\mathcal{I}) \to Y$. We say that A has the *Restricted Isometry Property* (RIP) of order k if there is a $0 < \gamma_k < 1$ such that

$$(1-\gamma_k)\|\mathbf{u}_\Lambda\|_{\ell_2}^2 \leq \|A_\Lambda \mathbf{u}_\Lambda\|_Y^2 \leq (1+\gamma_k)\|\mathbf{u}_\Lambda\|_{\ell_2}^2 \tag{3.2}$$

for all $\mathbf{u}_\Lambda \in \ell_2(\Lambda)$ with $\Lambda \subset \mathcal{I}$, and $\#\Lambda \leq k$. This property essentially requires that every set of columns with cardinality less than k is approximately an orthonormal system.

For our purposes, the following characterization of the RIP is of particular importance; see [4, Definition 1.2].

Proposition 3.1. *The following conditions are equivalent for* $A \in \mathcal{L}(\ell_2(\mathcal{I}), Y)$:

(i) *A has the RIP property of order k and constant γ_k.*
(ii) *For all $\Lambda \subset \mathcal{I}$ with $\#\Lambda \leq k$, the symmetric matrix $A^*A|_{\Lambda\times\Lambda}$ is positive definite with eigenvalues in $[1-\gamma_k, 1+\gamma_k]$.*
(iii) *For all $\Lambda \subset \mathcal{I}$ with $\#\Lambda \leq k$, it holds that $\|(I - A^*A)|_{\Lambda\times\Lambda}\| \leq \gamma_k$.*

We are now in the position to state the convergence result for D-ISTA in the case of exact operator evaluations.

Theorem 3.2. *Let* $\|A\| < \sqrt{2}$ *and* $\bar{\mathbf{u}} := (I - A^*A)\mathbf{u}^* + A^*y \in \ell_\tau^w(\mathcal{I})$ *for some* $0 < \tau < 2$*. Moreover, let* $L = L(\alpha) := \frac{4\|\mathbf{u}^*\|^2_{\ell_2(\mathcal{I})}}{\bar{\alpha}^2} + 4C\|\bar{\mathbf{u}}\|^\tau_{\ell_\tau^w(\mathcal{I})}\bar{\alpha}^{-\tau}$*, and assume that* A *fulfills the RIP of order* $2L + \#\operatorname{supp}\mathbf{u}^*$ *with constant* $0 < \gamma_0 < 1$*. Then, for any* $\gamma_0 \leq \gamma < 1$*, the iterates* $\mathbf{u}^{(n)}$ *from* (3.1) *fulfill* $\#\operatorname{supp}\mathbf{u}^{(n)} \leq L$ *and they converge to* $\mathbf{u}^*$ *at a linear rate*

$$\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2(\mathcal{I})} \leq \gamma^n \|\mathbf{u}^*\|_{\ell_2(\mathcal{I})} =: \epsilon_n \tag{3.3}$$

whenever the $\alpha^{(n)}$ *are chosen according to*

$$\alpha_\lambda \leq \alpha_\lambda^{(n)} \leq \alpha_\lambda + (\gamma - \gamma_0)L^{-1/2}\epsilon_n, \text{ for all } \lambda \in \mathcal{I}. \tag{3.4}$$

Remark 3.3. The RIP may seem restrictive, nevertheless, in Section 4 we will show how to obtain such property, for the cases when it fails, by suitable preconditioning strategies.

Proof of Theorem 3.2. We develop the proof by induction. For the initial iterate, we have $\mathbf{u}^{(0)} = \mathbf{0}$, so that $\#\operatorname{supp}\mathbf{u}^{(0)} \leq L$ and (3.3) is trivially true. Assume as an induction hypothesis that $S^{(n)} := \operatorname{supp}\mathbf{u}^{(n)}$ is such that $\#S^{(n)} \leq L$, and $\|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2(\mathcal{I})} \leq \epsilon_n$. Abbreviating $\mathbf{w}^{(n)} := \mathbf{u}^{(n)} + A^*(y - A\mathbf{u}^{(n)})$, by $\|A^*A\| \leq 2$ and the induction hypothesis, it follows that

$$\|\bar{\mathbf{u}} - \mathbf{w}^{(n)}\|_{\ell_2(\mathcal{I})} = \big\|(I - A^*A)(\mathbf{u}^* - \mathbf{u}^{(n)})\big\|_{\ell_2(\mathcal{I})} \leq \|\mathbf{u}^* - \mathbf{u}^{(n)}\|_{\ell_2(\mathcal{I})} \leq \epsilon_n.$$

Hence, using (2.9), we obtain the estimate

$$\#S^{(n+1)} = \#\operatorname{supp}\mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)}) \leq \#\Lambda_{\bar{\alpha}}(\mathbf{w}^{(n)}) \leq \frac{4\epsilon_n^2}{\bar{\alpha}^2} + 4C\|\bar{\mathbf{u}}\|^\tau_{\ell_\tau^w(\mathcal{I})}\bar{\alpha}^{-\tau} \leq L. \tag{3.5}$$

Since also $\#S^{(n)} \leq L$ by induction hypothesis, the set $\Lambda^{(n)} := S^{(n)} \cup S^{(n+1)}$ has at most $2L$ elements. Let us abbreviate $S := \operatorname{supp}\mathbf{u}^*$. By assumption, $(I - A^*A)|_{S\cup\Lambda^{(n)}\times S\cup\Lambda^{(n)}}$ is contractive with constant γ_0. Using the identities

$$\mathbf{u}^*_{S\cup\Lambda^{(n)}} = \mathbb{S}_\alpha(\bar{\mathbf{u}}_{S\cup\Lambda^{(n)}}) = \mathbb{S}_\alpha\big(\mathbf{u}^*_{S\cup\Lambda^{(n)}} + A^*_{S\cup\Lambda^{(n)}}(y - A_{S\cup\Lambda^{(n)}}\mathbf{u}^*_{S\cup\Lambda^{(n)}})\big)$$

and

$$\mathbf{u}^{(n+1)}_{S\cup\Lambda^{(n)}} = \mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)}_{S\cup\Lambda^{(n)}}) = \mathbb{S}_{\alpha^{(n)}}\big(\mathbf{u}^{(n)}_{S\cup\Lambda^{(n)}} + A^*_{S\cup\Lambda^{(n)}}(y - A_{S\cup\Lambda^{(n)}}\mathbf{u}^{(n)}_{S\cup\Lambda^{(n)}})\big),$$

it follows from (1.6), (2.4), (1.7), and $\alpha^{(n)} \geq \alpha$ that

$$\begin{aligned}
\|\mathbf{u}^* - \mathbf{u}^{(n+1)}\|_{\ell_2(\mathcal{I})} &= \big\|\mathbb{S}_\alpha(\bar{\mathbf{u}}_{S\cup\Lambda^{(n)}}) - \mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)}_{S\cup\Lambda^{(n)}})\big\|_{\ell_2(S\cup\Lambda^{(n)})} \\
&\leq \big\|\mathbb{S}_\alpha(\bar{\mathbf{u}}_{S\cup\Lambda^{(n)}}) - \mathbb{S}_\alpha(\mathbf{w}^{(n)}_{S\cup\Lambda^{(n)}})\big\|_{\ell_2(S\cup\Lambda^{(n)})} \\
&\quad + \big\|\mathbb{S}_\alpha(\mathbf{w}^{(n)}_{S\cup\Lambda^{(n)}}) - \mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)}_{S\cup\Lambda^{(n)}})\big\|_{\ell_2(S\cup\Lambda^{(n)})} \\
&\leq \big\|(I - A^*A|_{S\cup\Lambda^{(n)}\times S\cup\Lambda^{(n)}})(\mathbf{u}^* - \mathbf{u}^{(n)})_{S\cup\Lambda^{(n)}}\big\|_{\ell_2(S\cup\Lambda^{(n)})} \\
&\quad + \Big(\#\Lambda_{\bar{\alpha}}(\mathbf{w}^{(n)})\Big)^{1/2} \max_{\lambda\in\Lambda_{\bar{\alpha}}(\mathbf{w}^{(n)})} |\alpha_\lambda - \alpha_\lambda^{(n)}| \\
&\leq \gamma_0\epsilon_n + \Big(\#\Lambda_{\bar{\alpha}}(\mathbf{w}^{(n)})\Big)^{1/2} \max_{\lambda\in\Lambda_{\bar{\alpha}}(\mathbf{w}^{(n)})} |\alpha_\lambda - \alpha_\lambda^{(n)}|.
\end{aligned}$$

Using (3.5) we obtain $\|\mathbf{u}^* - \mathbf{u}^{(n+1)}\|_{\ell_2(\mathcal{I})} \leq \gamma_0\epsilon_n + \sqrt{L}\max_{\lambda\in\Lambda_{\bar{\alpha}}(\mathbf{w}^{(n)})} |\alpha_\lambda^{(n)} - \alpha_\lambda|$, and, since the $\alpha^{(n)}$ are chosen according to (3.4), the claim follows. $\square$

3.2. **Iterative thresholding with inexact operator evaluations.** In the case of A being, e.g., the forward solution operator from a partial differential equation, iteration (3.1) will not be implementable as it is. Instead, one will have to replace the applications of A and A^* by suitable approximations. Instead of (3.1), we may then consider the implementable algorithm

$$\tilde{\mathbf{u}}^{(0)} = \mathbf{0}, \tilde{\mathbf{u}}^{(n+1)} = \mathbb{S}_{\alpha^{(n)}}\big(\tilde{\mathbf{u}}^{(n)} + \mathbf{RHS}[A^*y, \delta_n] - \mathbf{APPLY}[A^*A, \tilde{\mathbf{u}}^{(n)}, \xi_n]\big), \quad n = 0, 1, \ldots, \tag{3.6}$$

with certain tolerances δ_n and ξ_n, where the numerical routines **APPLY** and **RHS** are introduced in Section 1.3. We call this iteration the *adaptive iterative soft-thresholding algorithm* (A-ISTA). As we shall see, the proof of Theorem 3.2 still works also in the case of inexact operator evaluations.

Proposition 3.4. *Let $\|A\| < \sqrt{2}$ and let $\bar{\mathbf{u}} := (I - A^*A)\mathbf{u}^* + A^*y \in \ell_\tau^w(\mathcal{I})$ for some $0 < \tau < 2$. Assume $0 < \gamma_0 \le \gamma < \tilde{\gamma} < 1$ and $\rho \ge \frac{1}{1-\frac{\gamma}{\tilde{\gamma}}}$, i.e., $\frac{\gamma}{\tilde{\gamma}} + \frac{1}{\rho} \le 1$. Moreover, let $\tilde{L} = \tilde{L}(\alpha) := \frac{4(1+\tilde{\gamma}/\rho)\|\mathbf{u}^*\|^2_{\ell_2(\mathcal{I})}}{\bar{\alpha}^2} + 4C\|\bar{\mathbf{u}}\|^\tau_{\ell_\tau^w(\mathcal{I})}\bar{\alpha}^{-\tau}$, and assume that A fulfills the RIP of order $2\tilde{L} + \#\operatorname{supp}\mathbf{u}^*$ with constant γ_0. If $\delta_n = \xi_n = \frac{\tilde{\epsilon}_{n+1}}{2\rho}$, $\tilde{\epsilon}_n = \tilde{\gamma}^n\|\mathbf{u}^*\|_{\ell_2(\mathcal{I})}$, and the $\alpha^{(n)}$ are chosen according to*

$$\alpha_\lambda \le \alpha_\lambda^{(n)} \le \alpha_\lambda + (\gamma - \gamma_0)L^{-1/2}\tilde{\epsilon}_n, \ \textit{for all } \lambda \in \mathcal{I}, \tag{3.7}$$

then the iterates (3.6) *fulfill*

$$\#\operatorname{supp}\tilde{\mathbf{u}}^{(n)} \le \tilde{L} \ \textit{and} \ \|\mathbf{u}^* - \tilde{\mathbf{u}}^{(n)}\|_{\ell_2(\mathcal{I})} \le \tilde{\epsilon}_n. \tag{3.8}$$

Proof. Since $\tilde{\mathbf{u}}^{(0)} = \mathbf{0}$ and $\tilde{\epsilon}_0 = \|\mathbf{u}^*\|_{\ell_2(\mathcal{I})}$, we only discuss the n-th iteration step. Setting $\mathbf{w}^{(n)} := (I - A^*A)\tilde{\mathbf{u}}^{(n)} + A^*y$ and $\tilde{\mathbf{w}}^{(n)} := \tilde{\mathbf{u}}^{(n)} + \mathbf{RHS}[\frac{\tilde{\epsilon}_{n+1}}{2\rho}] - \mathbf{APPLY}[A^*A, \tilde{\mathbf{u}}^{(n)}, \frac{\tilde{\epsilon}_{n+1}}{2\rho}]$, from $\|A^*A\| \le 2$ and the induction hypothesis it follows that $\|\bar{\mathbf{u}} - \tilde{\mathbf{w}}^{(n)}\|_{\ell_2(\mathcal{I})} \le \big\|(I - A^*A)(\mathbf{u}^* - \tilde{\mathbf{u}}^{(n)})\big\|_{\ell_2(\mathcal{I})} + \frac{\tilde{\epsilon}_{n+1}}{\rho} \le (1 + \frac{\tilde{\gamma}}{\rho})\tilde{\epsilon}_n$, and $\|\bar{\mathbf{u}} - \mathbf{w}^{(n)}\|_{\ell_2(\mathcal{I})} \le \big\|(I - A^*A)(\mathbf{u}^* - \tilde{\mathbf{u}}^{(n)})\big\|_{\ell_2(\mathcal{I})} \le \tilde{\epsilon}_n$. By an application of (2.9) we obtain $\#\operatorname{supp}\tilde{\mathbf{u}}^{(n+1)} = \#\operatorname{supp}\mathbb{S}_{\alpha^{(n)}}(\tilde{\mathbf{w}}^{(n)}) \le \frac{4\left((1+\tilde{\gamma}/\rho)\tilde{\epsilon}_n\right)^2}{\bar{\alpha}^2} + 4C\|\bar{\mathbf{u}}\|^\tau_{\ell_\tau^w}\bar{\alpha}^{-\tau} \le \tilde{L}$, and similarly $\#\operatorname{supp}\mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)}) \le L \le \tilde{L}$. As in Theorem 3.2, we can restrict the iterations on small sets of entries, so that the local contractivity of $I - A^*A$ and (3.7) imply $\|\mathbf{u}^* - \mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)})\|_{\ell_2(\mathcal{I})} \le \gamma\tilde{\epsilon}_n$. The claim finally follows from

$$\begin{aligned}\|\mathbf{u}^* - \tilde{\mathbf{u}}^{(n+1)}\|_{\ell_2(\mathcal{I})} &\le \big\|\mathbf{u}^* - \mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)})\big\|_{\ell_2(\mathcal{I})} + \big\|\mathbb{S}_{\alpha^{(n)}}(\mathbf{w}^{(n)}) - \mathbb{S}_{\alpha^{(n)}}(\tilde{\mathbf{w}}^{(n)})\big\|_{\ell_2(\mathcal{I})} \\ &\le \gamma\tilde{\epsilon}_n + \frac{\tilde{\epsilon}_{n+1}}{\rho} = (\frac{\gamma}{\tilde{\gamma}} + \frac{1}{\rho})\tilde{\epsilon}_{n+1} \le \tilde{\epsilon}_{n+1}.\end{aligned}$$

□

4. A multilevel preconditioning

In Section 3, the convergence analysis of A-ISTA was done under the assumption of the Restricted Isometry Property. Unfortunately, for compact operators in Hilbert spaces, this assumption fails to hold in general. Indeed, anytime we represent such operators with respect to a quasi-diagonalizing basis (which is the one we would prefer in order to maximally sparsify the matrix), then the diagonal entries of the resulting matrix will decay according to the behavior of the spectrum of the operator. As a remedy, we suggest a rescaling strategy which essentially consists of a block-diagonal preconditioning of certain classes of such operators, which include several concrete and interesting cases (see §4.2). Note that in general a block-diagonal preconditioning is indeed usually quite effective when the matrix representation of an operator with respect to a given basis is diagonal dominant, with diagonal entries decreasing to zero.

4.1. **Block-diagonal preconditioning implies a Restricted Isometry Property for nfinite matrices.** Throughout this section, we have to be more specific concerning the operator K and the generating system $\Psi = (\psi_\lambda)_{\lambda\in\mathcal{I}}$. Typically, we shall be concerned with the following situation. Let us assume that the Hilbert space X and its dual X', together with $L_2(\Omega)$, $\Omega \subset \mathbb{R}^d$, form a Gelfand triple

$$X \subset L_2(\Omega) \subset X' \tag{4.1}$$

and that the operator K is a bounded linear operator from X' to $L_2(\Omega)$. Then, the operator K^*K is a well-defined bounded operator from X' to X. Moreover, we assume that the generating system is given by a *compactly supported* basis or a frame of *wavelet type* [6, §2.12] for $L_2(\Omega)$. In this case, the index $\lambda = (j, k, e)$ typically encodes several types of information, namely the *scale* $j \in \mathbb{Z}$, often denoted by $|\lambda|$, the *spatial location* k, and the *type* of the wavelet indexed by e. In the following we assume that we can label $k \in \mathbb{Z}^d$, exclusively for simplifying the analysis, which is in general a legitimate assumption for $\Omega = [0, 1]^d$, whereas for more complex domains one can reconduct the analysis to such latter situation, by using suitable decompositions and transformations; see, e.g., [6]. Within this setting, we shall make the following technical assumptions:

- The entries in the stiffness matrix of K^*K satisfy the following decay estimate,

$$\big|\langle K^*K\psi_\lambda, \psi_\mu\rangle\big| \le c_1 \frac{2^{-s||\lambda|-|\mu||}2^{-\sigma(|\lambda|+|\mu|)}}{\big(1 + 2^{\min(|\lambda|,|\mu|)}\mathrm{dist}(\Omega_\mu, \Omega_\lambda)\big)^r} \tag{4.2}$$

  where $c_1, s, \sigma, r \in \mathbb{R}_+, r > d$ and Ω_μ denotes the support of ψ_μ.
- For the diagonal entries, i.e., for $\mu = \lambda$, we require an additional estimate from below:

$$\big|\langle K^*K\psi_\lambda, \psi_\lambda\rangle\big| \ge c_2 2^{-2\sigma|\lambda|}. \tag{4.3}$$

- For the same scale, i.e., for $|\mu| = |\lambda|, \lambda = (|\lambda|, k, e), \mu = (|\lambda|, k', e')$, we assume that

$$\big|\langle K^*K\psi_\lambda, \psi_\mu\rangle\big| \le c_3 \frac{2^{-2\sigma|\lambda|}}{\big(1 + \|k - k'\|\big)^r}. \tag{4.4}$$

Remark 4.1. i) At first sight, the conditions (4.2), (4.3) and (4.4) seem to be rather restrictive. However, in §4.2 we present some concrete examples for which condition (4.2) is indeed satisfied, whereas the others may be checked case by case.

ii) For the special case $\sigma = 0$, the estimate (4.2) usually holds for zero order operators with Schwartz kernels. The parameters s and r depend on the smoothness of the wavelet basis, the mapping properties of the underlying operator, and on the number of vanishing moments of the wavelet basis. Typically, increasing the smoothness and the number of vanishing moments produces larger values of r and s. We refer to [13] for details.

With the assumptions above, we can now prove the following theorem.

Theorem 4.2. *Let $A^*A = F^*K^*KF = (\langle K^*K\psi_\lambda, \psi_\mu\rangle)_{\lambda,\mu\in\mathcal{I}}$ denote the stiffness matrix of K^*K. Let $D_j^b = (\langle K^*K\psi_\lambda, \psi_\mu\rangle)_{|\lambda|=|\mu|=j}$ denote the diagonal block of A^*A corresponding to the refinement level j, and let D^b denote the block diagonal matrix $D^b = (D_0^b, D_1^b, \ldots)$. Suppose that* (4.2)*,* (4.3)*, and* (4.4) *are satisfied with $c_2 > c_3\frac{d\tau_d}{r-d}$ and $\tau_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$. Then there exists a constant $C = C(c_1, c_2, c_3, r, d)$ such that for each finite set $\Lambda \subset \mathcal{I}$ with $|\Lambda| < 2^s C^{-1}$ the submatrix $((D^b)^{-1/2}A^*A(D^b)^{-1/2})|_{\Lambda\times\Lambda}$ satisfies*

$$\left\|(I - (D^b)^{-1/2}A^*A(D^b)^{-1/2})|_{\Lambda\times\Lambda}\right\| < C2^{-s}|\Lambda| < 1 \tag{4.5}$$

and

$$\mathcal{K}\big((D^b)^{-1/2}A^*A\,(D^b)^{-1/2}|_{\Lambda\times\Lambda}\big) \le \frac{1 + C\,2^{-s}|\Lambda|}{1 - C\,2^{-s}|\Lambda|}. \tag{4.6}$$

Remark 4.3. (1) Obviously, Theorem 4.2 implies that for increasing values of s (which can, e.g., be achieved by increasing the smoothness of the wavelet basis), larger index sets Λ can be used. We have been purposely not very precise concerning the concrete value of C, because it is difficult to evaluate, and Theorem 4.2 is based on Gerschgorin's theorem which is well known to give highly suboptimal results; see [34, Section 2.5]. In Figure 2 below, we show that the proposed preconditioning strategy works in practice also for quite large index sets Λ.

(2) In the proof of Theorem 4.2, we will ignore the dependence on the type of the wavelets, i.e., on the parameter e, since this dependence only produces an additional constant.

(3) The use of a diagonal preconditioner works as well in practice; see §7.0.1 and Figure 2. Unfortunately, Gerschgorin's theorem is too pessimistic in this case to allow for an estimate of the type (4.5).

Proof of Theorem 4.2. The proof is essentially based on Gerschgorin's theorem; see, e.g., [25, Theorem 7.2.1]. The first step is to estimate the decay of the entries of $(D_j^b)^{-1/2}$. To this end, we will use the following theorem which has been proved by Jaffard [29]; see [29, formula (7)] in particular.

Theorem 4.4. *Let $A = (a_{k,k'})_{k,k'\in\mathbb{Z}^d}$ be a symmetric matrix satisfying $\|A\| < 1$ and*

$$|a_{k,k'}| \le \frac{c_4}{\big(1 + \|k - k'\|\big)^r}. \tag{4.7}$$

Then the entries of $B = (b_{k,k'})_{k,k'\in\mathbb{Z}^d}, B = A^{-1}$ satisfy

$$|b_{k,k'}| \le \frac{c_5\,\lambda_{\min}(A)^{-1}}{\big(1 + \|k - k'\|\big)^r}. \tag{4.8}$$

Moreover, the entries of $B^{1/2} = (b_{k,k'}^{1/2})_{k,k'\in\mathbb{Z}^d}$ satisfy

$$|b_{k,k'}^{1/2}| \le \frac{c_6\,\lambda_{\min}(A)^{-1/2}}{\big(1 + \|k - k'\|\big)^r}. \tag{4.9}$$

To apply Theorem 4.4, we need to estimate the minimal eigenvalue of the j-th diagonal block $D_j^b =: (d_{k,k'}^j)_{k,k'\in\mathbb{Z}^d}$. Gerschgorin's theorem tells us that

$$\lambda_{\min}(D_j^b) \geq \min\Big\{d_{k,k}^j - \sum_{k'\neq k} |d_{k,k'}^j|\Big\}.$$

By using (4.3), (4.4), and $\sum_{k'\neq k}\frac{1}{(1+\|k-k'\|)^r} \leq \int_{\|x\|\geq 1}\frac{1}{\|x\|^r}\,\mathrm{d}x \leq \frac{d\tau_d}{r-d}$, we obtain $d_{kk}^j - \sum_{k'\neq k}|d_{k,k'}^j| \geq \Big(c_2 - c_3\sum_{k'\neq k}\frac{1}{\big(1+\|k-k'\|\big)^r}\Big)2^{-2\sigma j} \geq \Big(c_2 - c_3\frac{d\tau_d}{r-d}\Big)2^{-2j\sigma} =: c_7\,2^{-2\sigma j}$. Consequently, by (4.9), we get for $(D_j^b)^{-1/2} = (\tilde{d}_{k,k'}^j)$ the estimate

$$|\tilde{d}_{k,k'}^j| \leq \frac{c_8\,2^{\sigma j}}{\big(1+\|k-k'\|\big)^r}. \tag{4.10}$$

The next step is to estimate individual entries of $(D^b)^{-1/2}A^*A(D^b)^{-1/2}$. For simplicity, let us assume that $|\mu| > |\lambda|$. We get

$$\begin{aligned}\big|((D^b)^{-1/2}A^*A(D^b)^{-1/2})_{\lambda,\mu}\big| &= \Big|\sum_{|\lambda'|=|\lambda|}(D^b)^{-1/2}_{\lambda,\lambda'}\sum_{|\mu'|=|\mu|}(A^*A)_{\lambda',\mu'}(D^b)^{-1/2}_{\mu',\mu}\Big|\\ &\leq \sum_{|\mu'|=|\mu|}\sum_{|\lambda'|=|\lambda|}\big|(D^b)^{-1/2}_{\lambda,\lambda'}\big|\big|(A^*A)_{\lambda',\mu'}\big|\big|(D^b)^{-1/2}_{\mu',\mu}\big|.\end{aligned}$$

Recall that for compactly supported wavelets $\operatorname{supp}\psi_\lambda \subset 2^{-|\lambda|}k + 2^{-|\lambda|}Q$, where Q is a suitable cube centered at the origin [6, §2.12]. Combining this observation with (4.10), (4.2) and the fact that $|\mu| > |\lambda|$, we obtain by setting $\lambda' = (|\lambda|, k'), \mu' = (|\mu|, l')$,

$$\begin{aligned}&\big|((D^b)^{-1/2}\,A^*A\,(D^b)^{-1/2})_{\lambda,\mu}\big|\\ &\leq c_9\sum_{|\mu'|=|\mu|}\sum_{k'}\frac{2^{\sigma|\lambda|}}{\big(1+\|k-k'\|\big)^r}\frac{2^{-s||\lambda|-|\mu||}2^{-\sigma(|\lambda|+|\mu|)}}{\big(1+2^{|\lambda|}\|2^{-|\lambda|}k'-2^{-|\mu|}l'\|\big)^r}(D^b)^{-1/2}_{\mu',\mu}\\ &\leq c_9\sum_{|\mu'|=|\mu|}\sum_{k'}\frac{2^{-s||\lambda|-|\mu||}2^{-\sigma|\mu|}}{\big(1+\|k-k'\|\big)^r\big(1+\|k'-2^{-|\mu|+|\lambda|}l'\|\big)^r}(D^b)^{-1/2}_{\mu',\mu}.\end{aligned}$$

Invoking the fact that the matrices satisfying (4.7) form an algebra (see again [29]), and using (4.10) for another time yields

$$\begin{aligned}&\big|((D^b)^{-1/2}A^*A(D^b)^{-1/2})_{\lambda,\mu}\big|\\ &\qquad\leq \sum_{l'}\frac{c_{10}2^{-s||\lambda|-|\mu||}2^{-\sigma|\mu|}}{\big(1+\|k-2^{-|\mu|+|\lambda|}l'\|\big)^r}\frac{c_8\,2^{\sigma|\mu|}}{\big(1+||l'-l||\big)^r}\\ &\qquad\leq c_{10}c_8 2^{-s||\lambda|-|\mu||}\sum_{l'}\frac{1}{\big(1+\|k-2^{-|\mu|+|\lambda|}l'\|\big)^r\big(1+\|l'-l\|\big)^r}\\ &\qquad\leq c_{10}c_8 2^{-s||\lambda|-|\mu||}\sum_{l'}\frac{1}{\big(1+\|l'-l\|\big)^r} \leq C2^{-s||\lambda|-|\mu||}.\end{aligned} \tag{4.11}$$

The case $|\lambda| < |\mu|$ can be treated similarly, hence

$$\big|((D^b)^{-1/2}A^*A(D^b)^{-1/2})_{\lambda,\mu}\big| \leq C2^{-s}, \quad |\lambda| \neq |\mu|.$$

Now another application of Gerschgorin's theorem yields

$$\lambda_{\min}\big((D^b)^{-1/2}A^*A(D^b)^{-1/2}|_{\Lambda\times\Lambda}\big) \geq 1 - C2^{-s}|\Lambda|$$

and

$$\lambda_{\max}\big((D^b)^{-1/2}A^*A(D^b)^{-1/2}|_{\Lambda\times\Lambda}\big) \le 1 + C2^{-s}|\Lambda|.$$

Consequently,

$$\begin{aligned}&\big\|(I-(D^b)^{-1/2}A^*A(D^b)^{-1/2})|_{\Lambda\times\Lambda}\big\|\\ &\qquad = 1-\lambda_{\min}((I-(D^b)^{-1/2}A^*A(D^b)^{-1/2})|_{\Lambda\times\Lambda}) \le C2^{-s}|\Lambda| < 1,\end{aligned}$$

and

$$\begin{aligned}\mathcal{K}\big((D^b)^{-1/2}A^*A(D^b)^{-1/2}|_{\Lambda\times\Lambda}\big) &= \frac{\lambda_{\max}\big((D^b)^{-1/2}A^*A(D^b)^{-1/2}|_{\Lambda\times\Lambda}\big)}{\lambda_{\min}\big((D^b)^{-1/2}A^*A(D^b)^{-1/2}|_{\Lambda\times\Lambda}\big)}\\ &\le \frac{1+C2^{-s}|\Lambda|}{1-C2^{-s}|\Lambda|}.\end{aligned}$$

□

Remark 4.5. Note that from (4.11) we have also (for $|\lambda| < |\mu|$),

$$\begin{aligned}&\big|((D^b)^{-1/2}A^*A(D^b)^{-1/2})_{\lambda,\mu}\big|\\ &\quad\le C2^{-s||\lambda|-|\mu||}\sum_{l'}\frac{1}{\big(1+\|k-2^{-|\mu|+|\lambda|}l'\|\big)^r\big(1+\|l'-l\|\big)^r}\\ &\quad\le C2^{-s||\lambda|-|\mu||}\int_{\mathbb{R}^d}\frac{1}{\big(1+\|k-2^{-|\mu|+|\lambda|}x\|\big)^r\big(1+2^{-|\mu|+|\lambda|}\|x-l\|\big)^r}dx\\ &\quad= C2^{-s||\lambda|-|\mu||}2^{d||\lambda|-|\mu||}\int_{\mathbb{R}^d}\frac{1}{\big(1+\|k-\xi\|\big)^r\big(1+\|\xi-2^{-|\mu|+|\lambda|}l\|\big)^r}d\xi\\ &\quad\le C\frac{2^{(d-s)||\lambda|-|\mu||}}{\big(1+2^{\min\{|\lambda|,|\mu|\}}\|2^{-|\lambda|}k-2^{-|\mu|}l\|\big)^r}.\end{aligned}$$

This means that for $s > d$, the preconditioned matrix $((D^b)^{-1/2}A^*A(D^b)^{-1/2})$ is a zero-order operator; see the details in [13, formula (9.4.10)]. In particular, this matrix is compressible in the sense described in [13, §9.5]. For such matrices, an efficient **APPLY** routine, as required in §3.2, is provided; see [7, 38, 24].

4.2. **Operators which allow for an effective block-diagonal preconditioning.** In this section, we are concerned with a class of integral operators that fulfill (4.2) and therefore give rise to suitable preconditioning techniques.

In the Gelfand triple (4.1), let X be a Sobolev space on a domain or a closed manifold $\Omega \subset \mathbb{R}^d$, i.e., $X = H^t(\Omega), X' = H^{-t}(\Omega)$. We consider a bounded operator $S : H^{-t}(\Omega) \longrightarrow H^t(\Omega)$ of the form

$$(Su)(x) = \int_\Omega \Phi(x,\xi)u(\xi)\,d\xi, \quad x\in\Omega, \tag{4.12}$$

where we assume that the kernel Φ is of Schwartz type, i.e,

$$\big|\partial_x^\alpha\partial_\xi^\beta\Phi(x,\xi)\big| \le c_{\alpha,\beta}\|x-\xi\|^{-(d+2t+|\alpha|+|\beta|)}, \qquad t\in\mathbb{R},\ \text{and multi-indices } \alpha,\beta\in\mathbb{N}^d, \tag{4.13}$$

and smooth off the main diagonal. The classical layer potential operators, such as the single and double layer potentials, fall into this category. Further cases are discussed in more detail in [11]. We are interested in solving operator equations of the form $Su = f$. In practice, due to noisy data, it might happen that the right–hand side is not contained in $H^t(\Omega)$, but only in $L_2(\Omega)$. Therefore, we consider the problem

$$Ku = f \in L_2(\Omega), \quad Ku(x) = \int_\Omega \Phi(x,\xi)u(\xi)\,d\xi, \quad x \in \Omega. \tag{4.14}$$

We changed purposely the symbol $S \to K$ of the operator in order to point out its different action, in particular, its image space. We want to discretize (4.14) by means of a tight wavelet frame $(\psi_\lambda)_{\lambda\in\mathcal{I}}$ on Ω for which the following *cancellation property* [13, formula 9.3.4], [35] holds:

$$\left|\int_\Omega \phi(\xi)\psi_\lambda(\xi)\,d\xi\right| \le c_{0,\beta}2^{-|\lambda|(d/2+d^*+1)}\|\phi\|_{W^{\infty,d^*+1}(\Omega_\lambda)}.$$

It can be shown that if (4.13) holds, then the decay of the entries $\langle K\psi_{\lambda'},\psi_\lambda\rangle$ is typically governed by the following basic estimate,

$$2^{(|\lambda'|+|\lambda|)t}\big|\langle K\psi_{\lambda'},\psi_\lambda\rangle\big| \le C\frac{2^{-\eta||\lambda|-|\lambda'||}}{\big(1+2^{\min(|\lambda|,|\mu|)}\mathrm{dist}(\Omega_{\lambda'},\Omega_\lambda)\big)^{d+2m-2t}}, \tag{4.15}$$

where η depends on the mapping properties of S and the smoothness of the wavelet frame whereas m is related with the number of vanishing moments d^*; see, e.g. [15, 16, 35] for details. Equation (4.15) means that the matrix $(2^{(|\lambda'|+|\lambda|)t)}|\langle K\psi_{\lambda'}, \psi_\lambda\rangle|))_{\lambda,\lambda'\in\mathcal{I}}$ is contained in the *Lemarié algebra* $\mathcal{M}$ which is the class of all matrices $\mathbf{N} = (n_{\lambda,\lambda'})_{\lambda,\lambda'\in\mathcal{I}}$ such that

$$|n_{\lambda,\lambda'}| \le C\frac{2^{-\eta||\lambda|-|\lambda'||}}{\big(1+2^{\min(|\lambda|,|\mu|)}\mathrm{dist}(\Omega_{\lambda'},\Omega_\lambda)\big)^{r}}, \quad \text{for all } \lambda,\lambda' \in \mathcal{I}, \tag{4.16}$$

for some constant C and suitable parameters $r > d$ and $\eta > d/2$. We refer to [30] for further information concerning this class. Due to the usual norm equivalences of wavelet bases or frames, the discretization of (4.14) leads to the following biinfinite matrix equation on ℓ_2:

$$(F \circ K \circ F^* \circ \mathbf{D}^{-1})\mathbf{u} = F(f), \tag{4.17}$$

where $\mathbf{D} = \mathrm{diag}(2^{-t|\lambda|}, \lambda \in \mathcal{I})$ and F denotes the analysis operator associated with the tight frame, i.e.,

$$F(v) := \big(\langle v,\psi_\lambda\rangle\big)_{\lambda\in\mathcal{I}}.$$

We refer to [10] for details. Now the fundamental estimate (4.15) implies that

$$\mathbf{D}^{-1} \circ F \circ K \circ F^* \circ \mathbf{D}^{-1} = \mathbf{M} \in \mathcal{M}$$

and hence

$$F \circ K \circ F^* \circ \mathbf{D}^{-1} = \mathbf{DM}.$$

The Lemarie algebra is stable under taking adjoints; therefore,

$$\mathbf{D}^{-1} \circ F \circ K^* \circ F^*\mathbf{D}^{-1} = \mathbf{M}^* \in \mathcal{M}.$$

Since we are working with a tight frame, this yields:

$$\begin{aligned}
\mathbf{D}^{-1} \circ \big(\langle K^*K\psi_{\lambda'},\psi_\lambda\rangle\big)_{\lambda,\lambda'\in\mathcal{I}} \circ \mathbf{D}^{-1} &= \mathbf{D}^{-1} \circ F \circ K^*K \circ F^* \circ \mathbf{D}^{-1} \\
&= (\mathbf{D}^{-1} \circ F \circ K^* \circ F^*) \circ (F \circ K \circ F^* \circ \mathbf{D}^{-1}) \\
&= \mathbf{M}^*\mathbf{DDM}. \qquad (4.18)
\end{aligned}$$

Since $\mathbf{DD} = \mathrm{diag}(2^{2t|\lambda|}, \lambda \in \mathcal{I})$, we may estimate $|(\mathbf{M}^*\mathbf{DDM})_{\lambda,\lambda'}| \leq |(\mathbf{M}^*\mathbf{M})_{\lambda,\lambda'}|$, being $\mathbf{M}^*, \mathbf{M} \in \mathcal{M}$, also the product $\mathbf{M}^*\mathbf{M} \in \mathcal{M}$, maybe with slightly smaller parameters $r' < r$ and $\eta' < \eta$. Altogether, we arrive at

$$|\langle K^*K\psi_{\lambda'}\psi_\lambda\rangle| \leq C' \frac{2^{-\eta'||\lambda|-|\lambda'||}2^{-2t(|\lambda'|+|\lambda|)}}{\left(1+2^{\min(|\lambda|,|\mu|)}\mathrm{dist}(\Omega_{\lambda'},\Omega_\lambda)\right)^{r'}}, \quad \text{for all } \lambda, \lambda' \in \mathcal{I},$$

i.e., (4.2) holds.

5. Reduction to a finite dimensional problem

5.1. **Reformulation of the problem after preconditioning.** In this section we would like to show how preconditioned matrices can be indeed employed in the adaptive iteration (3.6). We start first by reformulating the functional J as follows:

$$\begin{aligned} J(\mathbf{u}) &= \|A\mathbf{u} - y\|_Y^2 + 2\alpha\|\mathbf{u}\|_{\ell_1(\mathcal{I})} \\ &= \|AD^{-1/2}\underbrace{D^{1/2}\mathbf{u}}_{:=\mathbf{z}} - y\|_Y^2 + 2\alpha\|D^{-1/2}\underbrace{D^{1/2}\mathbf{u}}_{:=\mathbf{z}}\|_{\ell_1(\mathcal{I})} \\ &= \|AD^{-1/2}\mathbf{z} - y\|_Y^2 + 2\alpha\|D^{-1/2}\mathbf{z}\|_{\ell_1(\mathcal{I})} := J^D(\mathbf{z}). \end{aligned}$$

Hence,

$$\operatorname*{argmin}_{\mathbf{u}\in\ell_2(\mathcal{I})} J(\mathbf{u}) = D^{-1/2}\left(\operatorname*{argmin}_{\mathbf{z}\in D^{1/2}\ell_2(\mathcal{I})} J^D(\mathbf{z})\right).$$

Here, we assume that $D^{1/2} : \ell_2(\mathcal{I}) \to D^{1/2}\ell_2(\mathcal{I})$ is a suitable preconditioning matrix which has a well-defined formal inverse $D^{-1/2}$ on its image $D^{1/2}\ell_2(\mathcal{I})$. Moreover, we also assume that this matrix $D^{1/2}$ is symmetric, block-diagonal, diagonal dominant, in the sense that $\|D^{-1/2}\mathbf{z}\|_{\ell_1(\mathcal{I})} \sim \|\,\mathrm{diag}(D^{-1/2})\mathbf{z}\|_{\ell_1(\mathcal{I})}$ for all $\mathbf{z} \in \ell_0(\mathcal{I})$ such that $|\mathbf{z}|_{\ell_0(\mathcal{I})} \leq \kappa$, and with positive diagonal elements decreasing to zero, so that the diagonal entries of its inverse $D^{-1/2}$ do increase to infinity. Moreover, since the treatment of the term $\|D^{-1/2}\mathbf{z}\|_{\ell_1(\mathcal{I})}$ can be difficult in practice, for the sake of simplicity we may want to exploit the approximation $\|D^{-1/2}\mathbf{z}\|_{\ell_1(\mathcal{I})} \sim \|\,\mathrm{diag}(D^{-1/2})\mathbf{z}\|_{\ell_1(\mathcal{I})}$, and, with a slight abuse, we redefine

$$J^D(\mathbf{z}) := \|AD^{-1/2}\mathbf{z} - y\|_Y^2 + 2\|\mathbf{z}\|_{\ell_{1,\alpha\,\mathrm{diag}(D^{-1/2})}(\mathcal{I})},$$

which is again in the form (1.1). The use of a diagonal preconditioner is also justified by the numerical evidences we report in §7.0.1 and Figure 2.

We provided examples of such matrices $D^{1/2}$ in §4, and we have also observed in Remark 4.5 that, when used, they produce preconditioned matrices $D^{-1/2}A^*AD^{-1/2}$ which represent zero-order operators. In particular, in these cases, $D^{-1/2}A^*AD^{-1/2}$ is a bounded operator on $\ell_2(\mathcal{I})$, so we will consider it throughout this section. However, in our discussion we will not be too concerned with the topology where the minimization $\mathrm{argmin}_{\mathbf{z}}\, J^D(\mathbf{z})$ should take place, because we will clarify below that, due to the thresholding action, this apparently *infinite* dimensional problem is actually a *finite* dimensional one. Hence no topological issues have to be taken into account.

5.2. **Equivalence of the infinite dimensional problem to a finite dimensional one.** Let us consider $\mathbf{z}^* = \operatorname{argmin}_{\mathbf{z}} J^D(\mathbf{z}) = D^{1/2}\mathbf{u}^*$. Since $\mathbf{u}^*$ is a finitely supported vector and $D^{1/2}$ is a block-diagonal matrix, then $\mathbf{z}^*$ is also a finitely supported vector. Let $S^* = \operatorname{supp}(\mathbf{z}^*)$ and $\Lambda^\circ \subset \mathcal{I}$ be any finite subset of indices sufficiently large, and $S^* \subset \Lambda^\circ$. We want to emphasize that we know the existence of the set Λ°, but we shall not define it precisely a priori. It is clear that the finite vector $\mathbf{z}^*|_{\Lambda^\circ}$ is also a minimizer of the functional

$$J^D|_{\Lambda^\circ}(\mathbf{z}) := \|(AD^{-1/2})_{\Lambda^\circ}\mathbf{z} - y\|_Y^2 + 2\alpha\|\operatorname{diag}(D^{-1/2})_{\Lambda^\circ}\mathbf{z}\|_{\ell_1(\Lambda^\circ)}, \tag{5.1}$$

defined for $\mathbf{z} \in \mathbb{R}^{\Lambda^\circ}$, which is a finite dimensional space.

5.3. **The adaptive numerical solution of the finite dimensional problem equals the one of the infinite dimensional problem.** Note that now the infinite dimensional optimization problem has been reformulated in this way into a finite dimensional one, for which the norm-topology on the solution space $\mathbb{R}^{\Lambda^\circ}$ is no longer a relevant issue. Let us assume, for instance, that the topology on $\mathbb{R}^{\Lambda^\circ}$ is simply the Euclidean. In such topology the adjoint operator of $AD^{-1/2}|_{\Lambda^\circ}$ is given simply by the transposed matrix. This implies that the minimizer $\mathbf{z}^*|_{\Lambda^\circ}$ may be computed via the following iterative thresholding algorithm:

$$\mathbf{z}^{(n+1)} = \mathbb{S}_{\alpha \operatorname{diag}(D^{-1/2})_{\Lambda^\circ}}(\mathbf{z}^{(n)} + (AD^{-1/2})^*_{\Lambda^\circ} y - (D^{-1/2}A^*AD^{-1/2})|_{\Lambda^\circ\times\Lambda^\circ}\mathbf{z}^{(n)}), \tag{5.2}$$

where the rate of convergence of the iterations is now governed by the conditioning of the iteration matrix $D^{-1/2}A^*AD^{-1/2}|_{\Lambda^\circ\times\Lambda^\circ}$, which is precisely the one we analyzed in Theorem 4.2.

Of course, the iteration (5.2) cannot be implemented, unless we know a priori precisely the set Λ°, which is, as previously noted, out of our capability. Nevertheless, in the following we want to show that the implementation of the *adaptive* algorithm (3.6), where the iteration matrix A^*A is substituted by $D^{-1/2}A^*AD^{-1/2}$ and the thresholding operations are suitably adapted to $\mathbb{S}_{\alpha^{(n)}\operatorname{diag}(D^{-1/2})}$, will turn out to be *equivalent* to applying *the same algorithm* to the resolution of the problem of minimizing the functional $J^D|_{\Lambda^\circ}$ in (5.1), as we were perfectly in possession of Λ°. In this preconditioned context we will denote by $\tilde{\mathbf{z}}^{(n)}$ the iterations of the adaptive algorithm and by $\mathbf{z}^*$ the minimizer, or fixed point of (5.2) in finite dimensions.

In particular, we will show that the use of the **APPLY** routine to approximate the matrix-vector multiplication for $D^{-1/2}A^*AD^{-1/2}$ or $D^{-1/2}A^*AD^{-1/2}|_{\Lambda^\circ\times\Lambda^\circ}$ is simply equivalent, despite the fact that Λ° is not given to us. Actually, it turns out that we can chose $\Lambda^\circ \supset \bigcup_{n=0}^\infty \Lambda^{(n)}$, where $\Lambda^{(n)} = \operatorname{supp}(\tilde{\mathbf{z}}^{(n)})$ is the support of the n^{th}-iteration. Therefore, the adaptive algorithm will construct for us the set Λ°.

5.4. **The procedure APPLY and its role in the reduction to finite dimensions.** In order to clarify our argument we need to have a more precise understanding of the functioning of the **APPLY** routine; in particular, for help with the explanation we refer to Figure 1. For the precise definition of **APPLY** we refer the reader to [7, 28, 38].

The underlying assumption for the efficient use of **APPLY** is the *compressibility* of the matrix $D^{-1/2}A^*AD^{-1/2}$, expressed by certain off-diagonal decay estimates. As shown in Remark 4.5 we can assume for several interesting cases that $D^{-1/2}A^*AD^{-1/2}$ is indeed a compressible matrix. Hence we are allowed to consider this procedure, and not only can we assume $D^{-1/2}A^*AD^{-1/2}$ bounded on ℓ_2, but

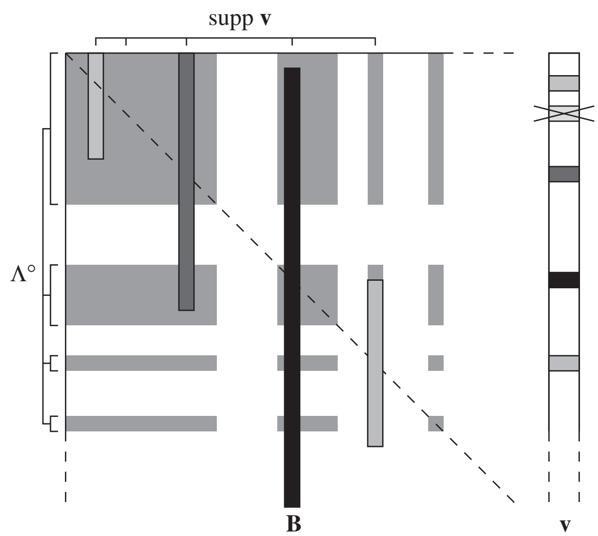


FIGURE 1. Scheme of approximate matrix-vector products $\mathbf{Bv}$, provided by the **APPLY** subroutine. The entries of $\mathbf{v}$ are multiplied with adaptively compressed columns of $\mathbf{B}$. Large entries of $\mathbf{v}$ meet only slightly compressed columns, whereas small entries are multiplied with strongly compressed columns or are even discarded. In the case that $\mathbf{B} = (D^{-1})^* A^* A D^{-1}$ and $\mathbf{v} = \mathbf{z}^{(n)}$, the support of the output may exceed Λ°.

also on ℓ_τ^w for $0 < \tau < 2$; see [7, 28, 38] for details. The procedure **APPLY** receives as an input a finite vector $\tilde{\mathbf{z}}^{(n)}$ and computes as output an approximation of the corresponding linear combination of columns of $D^{-1/2}A^*AD^{-1/2}$. Depending on the absolute value of the entries in $\tilde{\mathbf{z}}^{(n)}$ and on the requested accuracy, the associated matrix columns are adaptively compressed by discarding small off-diagonal matrix entries. The support size of the output $\mathbf{w}^{(n)}$ therefore hinges on the input $\tilde{\mathbf{z}}^{(n)}$ and on the accuracy requirements. We refer to [7, 38] for details on the particular compression rules.

Let us assume now that Λ° is sufficiently large so that, at the very beginning, none of the first, say, N^{th} iterations did exceed, after **APPLY**, the set Λ°. Hence, without loss of generality, we may assume that $\bigcup_{n=0}^N \Lambda^{(n)} \cup S^* \subset \Lambda^\circ$. We would like to show now that, thanks to the action of the thresholding operator, $\Lambda^{(n)} \subset \Lambda^\circ$ for all $n \geq N$ as soon as Λ° was chosen sufficiently large. Hence, regardless of the size of the support of the resulting vectors after an application of the **APPLY** routine on iterations $n \geq N$, the thresholding will produce the restriction of the output of **APPLY** to Λ°. Hence, it would be as if **APPLY** was outputting directly on Λ°.

Theorem 5.1. *For convenience of notation let us fix $M = D^{-1/2}A^*AD^{-1/2}$ and $P = D^{-1/2}A^*$. We can also assume $\|M\| < 2$, without loss of generality (recall that M is a bounded operator on $\ell_2(\mathcal{I})$ as well as on $\ell_\tau^w(\mathcal{I})$). Moreover, as a technical assumption we require that $Py \in \ell_2(\mathcal{I}) \cap \ell_\tau^w(\mathcal{I})$, for $0 < \tau < 2$. Let us assume that we are also in the conditions of applicability of Proposition* 3.4 *for the convergence of the algorithm* (3.6) *applied for the minimization of the finite dimensional functional*

(5.1)*, but actually without restricting yet neither the matrices nor the vectors to any set* Λ° *a priori fixed.*

We define $\bar{\mathbf{z}} := \mathbf{z}^* + (Py - M\mathbf{z}^*) \in \ell_2(\mathcal{I})$ *and*

$$\Lambda^\circ = \bigcup_{n=0}^{N} \Lambda^{(n)} \cup S^* \cup \Lambda_\delta(\bar{\mathbf{z}}),$$

where $N \in \mathbb{N}$ *is such that*

$$\|\mathbf{z}^* - \tilde{\mathbf{z}}^{(N)}\| \leq \tilde{\epsilon}_N \leq \varepsilon, \tag{5.3}$$

for

$$\left(1 + \frac{1}{\rho}\right)\varepsilon + \delta \leq \alpha \left(\inf_{\lambda \in \mathcal{I}} \operatorname{diag}(D^{-1/2})_\lambda\right), \quad \delta > 0. \tag{5.4}$$

(We recall the notations $\Lambda^{(n)} = \operatorname{supp}(\tilde{\mathbf{z}}^{(n)})$*,* $S^* = \operatorname{supp}(\mathbf{z}^*)$*, and* $\Lambda_\delta(\bar{\mathbf{z}}) = \{\lambda \in \mathcal{I} : \delta < \mathbf{z}^*_\lambda + (Py - M\mathbf{z}^*)_\lambda\}$*. The constant* ρ *is as in Proposition* 3.4*.)*
Then the supports of the iterations never exceed Λ°*, i.e.,*

$$\Lambda^{(n)} \subset \Lambda^\circ, \text{ for all } n \geq 0.$$

Proof. Note that $\Lambda^{(n)} \subset \Lambda^\circ$, for all $n \leq N$ by definition of Λ°. Let us assume that $n > N$. Let $\lambda \notin \Lambda^\circ$, hence $\lambda \notin S^* = \operatorname{supp}(\mathbf{z}^*)$, and $|\tilde{\mathbf{z}}^{(n)}_\lambda| \leq \varepsilon$. Moreover, we have

$$\begin{aligned}
& \|\mathbf{z}^* + Py - M\mathbf{z}^* - \left(\tilde{\mathbf{z}}^{(n)} + \mathbf{RHS}[\delta_n] - \mathbf{APPLY}[M, \tilde{\mathbf{z}}^{(n)}, \gamma_n]\right)\| \\
\leq \quad & \|\mathbf{z}^* + Py - M\mathbf{z}^* - \left(\tilde{\mathbf{z}}^{(n)} + Py - M\tilde{\mathbf{z}}^{(n)}\right)\| \\
+ \quad & \|\left(\tilde{\mathbf{z}}^{(n)} + Py - M\tilde{\mathbf{z}}^{(n)}\right) - \left(\tilde{\mathbf{z}}^{(n)} + \mathbf{RHS}[\delta_n] - \mathbf{APPLY}[M, \tilde{\mathbf{z}}^{(n)}, \gamma_n]\right)\| \\
\leq \quad & \varepsilon + \frac{\varepsilon}{\rho} = \left(1 + \frac{1}{\rho}\right)\varepsilon. \qquad (5.5)
\end{aligned}$$

Since $\mathbf{z}^*$ is a fixed point of (5.2) (in the finite dimensional environment) we have $|\mathbf{z}^*_\lambda + (Py - M\mathbf{z}^*)_\lambda| \leq \alpha \operatorname{diag}(D^{-1/2})_\lambda$. If $|\mathbf{z}^*_\lambda + (Py - M\mathbf{z}^*)_\lambda| \leq \delta$, then from (5.5) and (5.4) we obtain also $|\left(\tilde{\mathbf{z}}^{(n)} + \mathbf{RHS}[\delta_n] - \mathbf{APPLY}[M, \tilde{\mathbf{z}}^{(n)}, \gamma_n]\right)_\lambda| \leq \left(1 + \frac{1}{\rho}\right)\varepsilon + \delta \leq \alpha \left(\inf_{\lambda \in \mathcal{I}} \operatorname{diag}(D^{-1/2})_\lambda\right) \leq \alpha^{(n)} \operatorname{diag}(D^{-1/2})_\lambda$. Hence $\tilde{\mathbf{z}}^{(n+1)}_\lambda = 0$. Moreover, since $\mathbf{z}^*$ is a finite dimensional vector, M is a bounded operator on $\ell_2(\mathcal{I})$, and $Py \in \ell_2(\mathcal{I})$, then $\bar{\mathbf{z}} := \mathbf{z}^* + (Py - M\mathbf{z}^*) \in \ell_2(\mathcal{I})$ and therefore $\Lambda_\delta(\bar{\mathbf{z}}) = \{\lambda \in \mathcal{I} : \delta < \mathbf{z}^*_\lambda + (Py - M\mathbf{z}^*)_\lambda\}$ has finite cardinality. Hence, if eventually we define $\Lambda^\circ = \bigcup_{n=0}^{N} \Lambda^{(n)} \cup S^* \cup \Lambda_\delta(\bar{\mathbf{z}})$, then we have that $\Lambda^{(n)} \subset \Lambda^\circ$ for all $n \geq 0$. $\square$

Remark 5.2. 1. We remark that in the previous theorem the boundedness properties of the matrix $M = D^{-1/2} A^* A D^{-1/2}$ on $\ell_2(\mathcal{I})$ and $\ell^w_\tau(\mathcal{I})$ play a crucial role. These properties are ensured by the compressibility of M as mentioned in Remark 4.5. Therefore, neither for every operator K nor for any preconditioning matrices $D^{1/2}$ or different bases Ψ too, can we expect Theorem 5.1 to hold.

2. Preconditioning might create some topological troubles and complicate significantly the setting where the algorithm (3.6) can work. In simple words, Theorem 5.1 establishes that for operators and block-diagonal preconditioning, as in §4, the use of the adaptive algorithm (3.6) to compute a sparse solution $\mathbf{u}^*$ is allowed and it will converge as expected, because in practice it will behave the same as if being used for a finite dimensional problem, although the finite dimensional reference space is built on the fly.

6. Convergence to compressible and sparse solutions

In this section, for the sake of simplicity, we assume $\alpha_\lambda = \bar{\alpha}$ for all $\lambda \in \mathcal{I}$. Hence, we may assume that $\alpha = \bar{\alpha}$ is a scalar. The analysis we developed so far is valid when the parameter α is not too small. Indeed, we may expect that for α becoming smaller, the support of the minimizer $\mathbf{u}_\alpha^*$ of J_α (as in (1.1)) is becoming also larger and larger. Therefore, we have to expect that the constants $\gamma_0 = \gamma_0(\alpha)$ in Proposition 3.4 tend to 1, as $\alpha \to 0$. In this section we would like to explore the main features of such limit behavior.

Theorem 6.1. *Assume $\|A\| < \sqrt{2}$ and that there exists a compressible solution $\mathbf{u}^\circ \in \ell_\tau^w(\mathcal{I})$, for $0 < \tau < 2$, such that $A\mathbf{u}^\circ = y$. Let us define the function*

$$\Gamma(\alpha) := \frac{\|(I - A^*A)(\mathbf{u}_\alpha^* - \mathbf{u}^\circ)\|_{\ell_2(\operatorname{supp}(\mathbf{u}_\alpha^*) \cup \operatorname{supp}(\mathbf{u}^\circ))}}{\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})}}. \tag{6.1}$$

The function Γ is certainly bounded above by 1; here we assume that $\Gamma(\alpha) < 1$. Then we have the following estimates:

(i) *there exists a constant $C = C(\tau)$ such that*

$$\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \le \frac{C}{1 - \Gamma(\alpha)} |\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau/2} \alpha^{1-\tau/2};$$

(ii) *for another constant $C' = C'(\tau) > 0$ we have*

$$\#\operatorname{supp}(\mathbf{u}_\alpha^*) \le \left(\frac{4C^2}{(1 - \Gamma(\alpha))^2} + 4C' \right) |\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau} \alpha^{-\tau}.$$

Proof. Let us write

$$\mathbf{u}_\alpha^* - \mathbf{u}^\circ = \mathbb{S}_\alpha(\mathbf{u}_\alpha^* - A^*A\mathbf{u}_\alpha^* + A^*A\mathbf{u}^\circ) - \mathbb{S}_\alpha(\mathbf{u}^\circ) + \mathbb{S}_\alpha(\mathbf{u}^\circ) - \mathbf{u}^\circ.$$

Then, from (1.6) and (2.6), we have the estimates

$$\begin{aligned}
\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} &\le \|\mathbb{S}_\alpha(\mathbf{u}_\alpha^* - A^*A\mathbf{u}_\alpha^* + A^*A\mathbf{u}^\circ) - \mathbb{S}_\alpha(\mathbf{u}^\circ)\|_{\operatorname{supp}(\mathbf{u}_\alpha^*) \cup \operatorname{supp}(\mathbf{u}^\circ)} \\
&\quad + \|\mathbb{S}_\alpha(\mathbf{u}^\circ) - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \\
&\le \|(I - A^*A)(\mathbf{u}_\alpha^* - \mathbf{u}^\circ)\|_{\ell_2(\operatorname{supp}(\mathbf{u}_\alpha^*) \cup \operatorname{supp}(\mathbf{u}^\circ))} + C|\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau/2} \alpha^{1-\tau/2} \\
&\le \Gamma(\alpha)\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} + C|\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau/2} \alpha^{1-\tau/2}.
\end{aligned}$$

The latter estimate immediately shows (i). Now note that

$$\begin{aligned}
\|(\mathbf{u}_\alpha^* - A^*A\mathbf{u}_\alpha^* + A^*y) - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} &= \|(I - A^*A)(\mathbf{u}_\alpha^* - \mathbf{u}^\circ)\|_{\ell_2(\mathcal{I})} \\
&\le \frac{C}{1 - \Gamma(\alpha)} |\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau/2} \alpha^{1-\tau/2}.
\end{aligned}$$

A straightforward application of (2.9) yields

$$\begin{aligned}
\#\operatorname{supp}(\mathbf{u}_\alpha^*) &\le \frac{4C^2|\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau} \alpha^{2-\tau}}{(1 - \Gamma(\alpha))^2 \alpha^2} + 4C'|\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau} \alpha^{-\tau} \\
&\le \left(\frac{4C^2}{(1 - \Gamma(\alpha))^2} + 4C' \right) |\mathbf{u}^\circ|_{\ell_\tau^w}^{\tau} \alpha^{-\tau}.
\end{aligned}$$

This concludes the proof of (ii). $\square$

Corollary 6.2. *Under the hypothesis and notation of Theorem* 6.1, *assume now that* $\mathbf{u}^\circ$ *is a sparse solution, i.e.,* $\mathbf{u}^\circ \in \ell_0(\mathcal{I}) := \bigcap_{\tau>0} \ell_\tau^w(\mathcal{I})$, *and* $|\mathbf{u}^\circ|_{\ell_0} = \#\operatorname{supp}(\mathbf{u}^\circ) < \infty$. *Let us denote* $S^\circ = \operatorname{supp}(\mathbf{u}^\circ)$ *and* $|\mathbf{u}^\circ|_{\ell_0(\mathcal{I})} := \#S^\circ$. *Then we have the following estimates:*

(i) $\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \leq \frac{1}{1-\Gamma(\alpha)} |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}^{1/2} \alpha$;

(ii) $\#\operatorname{supp}(\mathbf{u}_\alpha^*) \leq \left(\frac{4}{(1-\Gamma(\alpha))^2} + 1\right) |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}$.

Proof. The estimates (i) and (ii) follow immediately from Theorem 6.1 by taking the limit $\tau \to 0$ (see formulas (2.7) and (2.8); note the modification of the constants for $\tau = 0$ in Lemma 2.2). ☐

We conclude this section with a result which establishes conditions for a uniform behavior of algorithms (3.1) and (3.6) as $\alpha \to 0$.

Corollary 6.3. *Assume that there exists* $\mathbf{u}^\circ \in \ell_0(\mathcal{I})$ *such that* $A\mathbf{u}^\circ = y$, *and*

$$\|I - A^*A|_{\Lambda\times\Lambda}\| \leq \gamma_0 < 1 \tag{6.2}$$

for all $\Lambda \subset \mathcal{I}$ *such that* $\#\Lambda \leq \left(\frac{4}{(1-\gamma_0)^2} + 2\right) |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}$. *Then the following conditions are equivalent:*

(a) $\sup_{\alpha>0} \Gamma(\alpha) \leq \gamma_0 < 1$;

(b) $\mathbf{u}_\alpha^*$ *converges linearly to* $\mathbf{u}^\circ$, *i.e.,* $\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \leq \frac{1}{1-\gamma_0} |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}^{1/2} \alpha$;

(c) $\sup_{\alpha>0} \#\operatorname{supp}(\mathbf{u}_\alpha^*) \leq (\frac{4}{(1-\gamma_0)^2} + 1) |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}$.

Proof. Assume that (a) holds. Then by Corollary 6.2 (i) we have

$$\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \leq \frac{1}{1-\gamma_0} |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}^{1/2} \alpha,$$

which implies (b). Let us assume (b) now. Then, again an application of Lemma 2.2 (for the constants adjusted to $\tau = 0$) yields (c):

$$\#\operatorname{supp}(\mathbf{u}_\alpha^*) \leq \frac{4|\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}\alpha^2}{(1-\gamma_0)^2\alpha^2} + |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})} = \left(\frac{4}{(1-\gamma_0)^2} + 1\right) |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}.$$

If (c) holds, then from (6.2) we have

$$\Gamma(\alpha) = \frac{\|(I - A^*A)(\mathbf{u}_\alpha^* - \mathbf{u}^\circ)\|_{\ell_2(\operatorname{supp}(\mathbf{u}_\alpha^*)\cup\operatorname{supp}(\mathbf{u}^\circ))}}{\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})}} \leq \gamma_0.$$ ☐

Remark 6.4. In the recent papers [26, 27] the authors investigated the conditions:

(1) (source or range condition) $\operatorname{range}(A^*) \cap \partial\|\cdot\|_{\ell_1(\mathcal{I})}(\mathbf{u}^\circ) \neq \emptyset$, where $\partial\|\cdot\|_{\ell_1(\mathcal{I})}(\mathbf{u})$ is the subdifferential of the ℓ_1-norm at the point $\mathbf{u}$, given by

$$\partial\|\cdot\|_{\ell_1(\mathcal{I})}(\mathbf{u}) = \{\xi \in \ell_2(\mathcal{I}) : \ \xi_\lambda \in \partial|\cdot|(\mathbf{u}_\lambda), \lambda \in \mathcal{I}\}$$

where $\partial|\cdot|(z) = \{\operatorname{sign}(z)\}$ if $z \neq 0$ and $\partial|\cdot|(0) = [-1, 1]$; this condition implies that

$$\mathbf{u}^\circ = \arg\min_{A\mathbf{u}=y} \|\mathbf{u}\|_{\ell_1(\mathcal{I})};$$

(2) (local injectivity condition) A_{Λ° is injective where $\Lambda^\circ = \operatorname{supp}(\mathbf{u}^\circ)$.

According to [27, Proposition 3.12] these conditions imply the linear convergence

$$\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \leq C\alpha,$$

for a constant $C = C(\mathbf{u}^\circ, \|A\|, \|A_{\Lambda^\circ}^{-1}\|)$. This result is not uniform in the sense that the constant C depends on $\mathbf{u}^\circ$, and not only on its support size, and it reflects the nonuniformity of the conditions (1) and (2). Clearly in applications, one would prefer to have uniform guarantees. In particular, if the Restricted Isometry Property holds, more precisely, if

$$\|I - A^*A|_{\Lambda\times\Lambda}\| \leq \gamma_0' < 1 \tag{6.3}$$

for all $\Lambda \subset \mathcal{I}$ such that $\#\Lambda \leq |\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}$, then [27, Theorem 5.6] ensures

$$\|\mathbf{u}_\alpha^* - \mathbf{u}^\circ\|_{\ell_2(\mathcal{I})} \leq C'\alpha,$$

for a constant $C' = C'(|\mathbf{u}^\circ|_{\ell_0}^{1/2}, \gamma_0')$. The precise expression of this constant can be found in [27, formula (35)]. This constant becomes smaller for smaller γ_0'. Note further that this result is now (more) uniform because it depends exclusively on the support size of $\mathbf{u}^\circ$. When we compare conditions (6.2) and (6.3), we see immediately that

$$\gamma_0' \leq \gamma_0.$$

Hence there might be situations for which the values of γ_0' and γ_0 are such that

$$C'(|\mathbf{u}^\circ|_{\ell_0}^{1/2}, \gamma_0') \leq \left(\frac{1}{1-\gamma_0}|\mathbf{u}^\circ|_{\ell_0(\mathcal{I})}^{1/2}\right).$$

In this case, condition (b) of Corollary 6.3 would be satisfied, hence also the other equivalent conditions (a) and (b) would hold.

7. Numerical experiments

For instructive and simple numerical experiments in an infinite-dimensional setting, we will consider the Volterra integral operator $K : L_2(0,1) \to L_2(0,1)$,

$$Ku(t) = \int_0^t u(s)\,\mathrm{d}s, \quad K^*Ku(t) = \int_0^1 \big(1 - \max(s,t)\big)u(s)\,\mathrm{d}s. \tag{7.1}$$

The integration operator can be regarded as a model case for more general Fredholm-type integral operators. K is injective and bounded with norm $\|K\| = 2/\pi \approx 0.64$. The nonzero eigenvalues of K^*K are explicitly available as $\lambda_n = 1/(\pi(n+\frac{1}{2}))^2$; see [31]. In the following, the discretization of K will be performed using a biorthogonal, piecewise linear spline wavelet basis for $L_2(0,1)$ with 2 vanishing moments from [32]. Due to the piecewise linear kernel of K^*K and the compression properties of the wavelet basis, the system matrix A^*A is quasi-sparse. In the experiments, A^*A is scaled to have $\mathcal{L}(\ell_2)$-norm 1.

7.0.1. *Local well-conditioning.* As a first numerical test, we investigate the effect of preconditioning strategies on the spectral properties of small submatrices of A^*A. For each $N \in \mathbb{N}$, we randomly select 10 times a support set $\Lambda \subset \mathcal{I}$ of size $\#\Lambda \leq N$. The arithmetic mean over the spectral condition numbers of these submatrices is plotted against the support size N in Figure 2.

As we clearly see, both preconditioning strategies drastically improve the spectral properties of small submatrices. In particular, for square submatrices with less

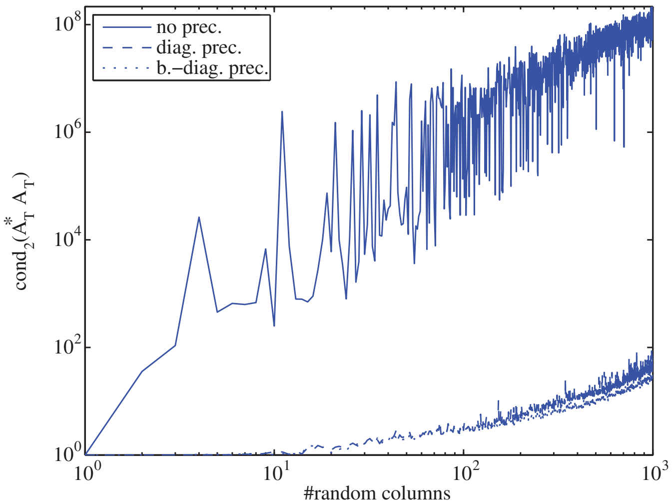


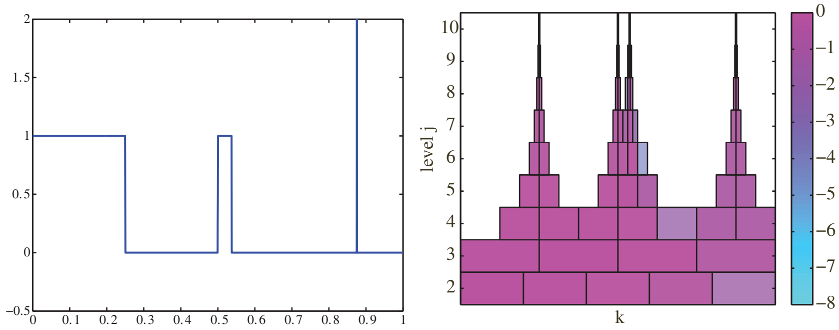
FIGURE 2. Average spectral condition numbers $\mathrm{cond}_2(A^*A|_{\Lambda\times\Lambda})$ of small $N \times N$-submatrices of A^*A, without preconditioning (*solid line*), with diagonal preconditioning (*dashed line*), and with block-diagonal preconditioning (*dotted line*)

than 1000 columns, their condition numbers stay below 10^2 for diagonal preconditioning, whereas they exceed 10^8 in the case of the original matrix. Note that the spectral improvement via diagonal preconditioning is already significant, whereas switching to block-diagonal preconditioning does not further improve the conditioning considerably. Therefore, in practice, the additional computational work of block-diagonal preconditioning can be avoided at no noticeable loss. Of course, due to the clustering of the singular values at 0, larger submatrices of A^*A will fail to have good spectral properties in both preconditioned cases as well.

7.0.2. *Recovery of sparse solutions.* In order to validate the numerical performance of the proposed recovery algorithms, we will fix in the sequel a piecewise linear function $u \in L_2(0,1)$ as the solution of $Ku = y$. The function u is chosen to have nodes at dyadic points, and therefore its expansion in a piecewise linear spline wavelet basis has only finitely many nonzero coefficients; see also Figure 3. In particular, we have $\|\mathbf{u}\|_{\ell_0} = 72$, $\|\mathbf{u}\|_{\ell_1} \approx 3.8525$ and $\|\mathbf{u}\|_{\ell_2} \approx 0.8711$.

7.0.3. *Choice of the regularization parameter.* The exact minimizer $\mathbf{u}^* = \mathbf{u}^*_\alpha$ depends on the regularization parameter α in a nonlinear way. Due to the presence of roundoff errors in the computation of the system matrix and of the right-hand side, and due to the truncation of the thresholding process after finitely many steps, we may not expect that the numerical reconstructions $\tilde{\mathbf{u}} \approx \mathbf{u}^*$ perfectly match the unknown solution $\mathbf{u}$ as α gets arbitrarily small. Experimentally, it can be observed that the residual errors begin to stagnate for values of α significantly smaller than 10^{-6}. Therefore, in the numerical experiments, we choose parameters $\alpha \geq 10^{-7}$.

7.0.4. *Linear convergence to the minimizer.* By the injectivity of the forward operator K and by the finite support of the minimizer $\mathbf{u}^*$ for each given $\alpha > 0$, any of the considered thresholding algorithms should converge linearly. However, the



(a) graph (b) wavelet coefficient distribution

FIGURE 3. Unknown function u to be recovered

particular error reduction constants will most likely depend on the magnitude of the index set in which the particular iterates are supported. Due to the spectral properties of A^*A, the choice of a large support set within a given iterative scheme will inhibit a significant error reduction per iteration step, while smaller support sets will be beneficial. Therefore, we hope that an iteration with decreasing threshold parameters converges faster than an algorithm with fixed parameters. Moreover, based on the observation from Figure 2, preconditioning strategies should further reduce the error reduction constants considerably.

In order to quantify the linear convergence to the minimizer, we first compute a very good approximation to the exact minimizer $\mathbf{u}^*$ by a combination of the well-known FISTA algorithm of Beck and Teboulle [1] and some semismooth Newton steps. Then we restart from $\mathbf{u}^{(0)} = \mathbf{0}$, and we measure the ℓ_2 error of the iterands with respect to that limit. In the experiments, the threshold parameters are chosen according to the rule $\alpha^{(n)} = \alpha + \eta\gamma^n\epsilon_0$, for certain parameters $0 < \gamma, \eta < 1$ and the choice $\epsilon_0 := 5\|\mathbf{u}\|_{\ell_1}$. The iteration is allowed to use all wavelet indices λ with level $|\lambda| \leq 12$. Note that the neglected entries $(A^*A)_{\lambda,\mu}$ are already smaller than 10^{-8} in absolute value, so this is not really a limitation.

In Figures 4 and 5, we plot the ℓ_2 error histories for different iterative thresholding algorithms. We clearly observe linear convergence of all iterative schemes. The original iterative thresholding method (ISTA) and its variant with decreasing threshold parameters (D-ISTA) show a similar asymptotic behavior. Both preconditioned variants P-ISTA and PD-ISTA have a significantly better error reduction per iteration for any choice of the regularization parameter α.

A decreasing choice of the threshold parameters only pays off for medium values of $\alpha \approx 10^{-6}$. An explanation of the varying slopes in Figure 4 can be found when considering the support sizes of the iterands during the iteration; see Figure 7 from the next subsection. For $\alpha = 10^{-6}$, e.g., PD-ISTA asymptotically works on a slightly larger index set than P-ISTA, resulting in a marginally worse error reduction per iteration.

7.0.5. *Support dynamics.* One of the core ideas towards the acceleration of iterative thresholding methods is to keep the number of active degrees of freedom as small as possible. The rationale is to take advantage of the spectral bounds of the associated submatrices of A^*A. In Figures 6 and 7, we track the support size histories for

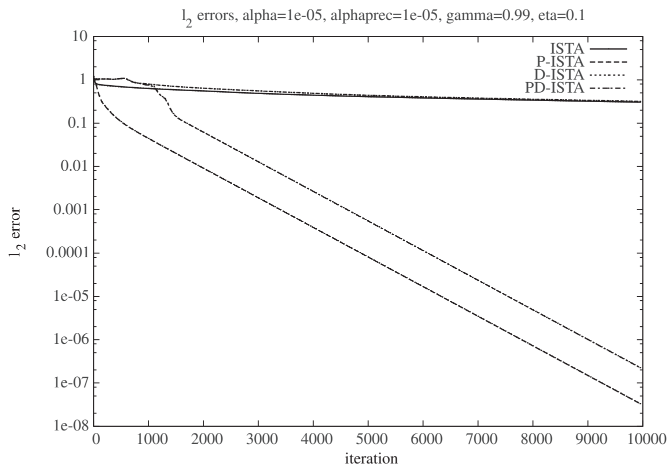


FIGURE 4. ℓ_2 error histories for different iterative thresholding algorithms, $\alpha = 10^{-5}$, $\gamma = 0.99$ and $\eta = 0.1$: ISTA (*solid line*), D-ISTA (*dotted line*), P-ISTA (*dashed line*), and PD-ISTA (*dash-dotted line*)

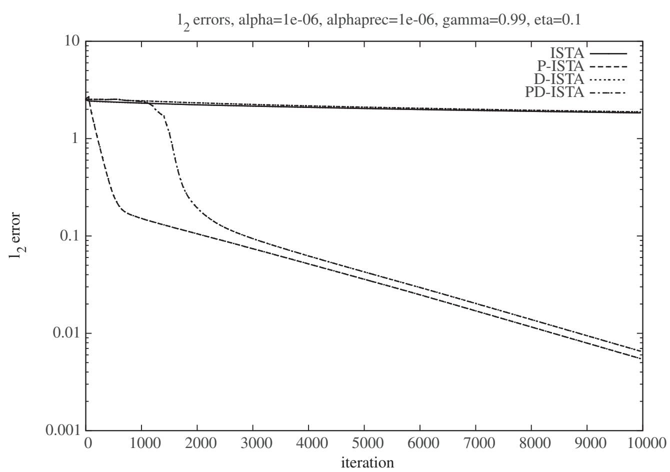


FIGURE 5. ℓ_2 error histories for different iterative thresholding algorithms, $\alpha = 10^{-6}$, $\gamma = 0.99$ and $\eta = 0.1$: ISTA (*solid line*), D-ISTA (*dotted line*), P-ISTA (*dashed line*), and PD-ISTA (*dash-dotted line*)

different iterative thresholding algorithms and different regularization parameters α. Apparently, in the course of the iteration, all algorithms gradually reduce the support size of the respective iterands. A decreasing choice of the threshold parameters α_n can efficiently reduce the active degrees of freedom in the transient phase of the first iterations. What is more important, the considered thresholding algorithms behave in a quantitatively different way in regard to the number of active coefficients. As soon as the regularization parameter α is sufficiently small,

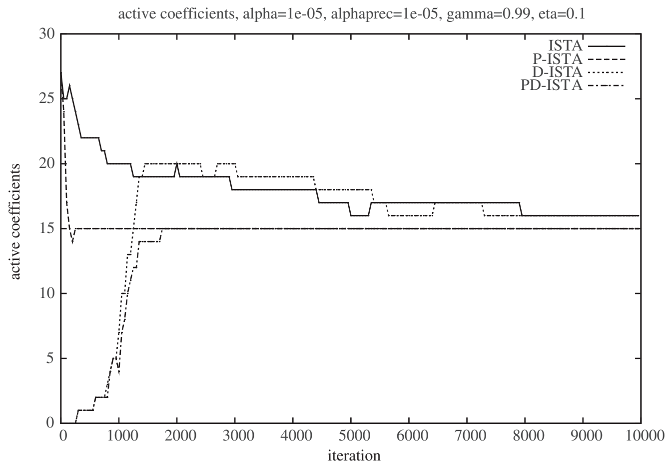


FIGURE 6. Support size histories for different iterative thresholding algorithms, $\alpha = 10^{-5}$, $\gamma = 0.99$ and $\eta = 0.1$: ISTA (*solid line*), D-ISTA (*dotted line*), P-ISTA (*dashed line*), and PD-ISTA (*dash-dotted line*)

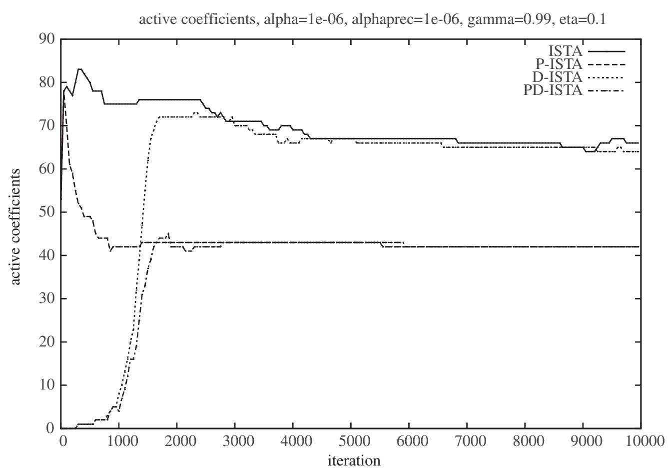
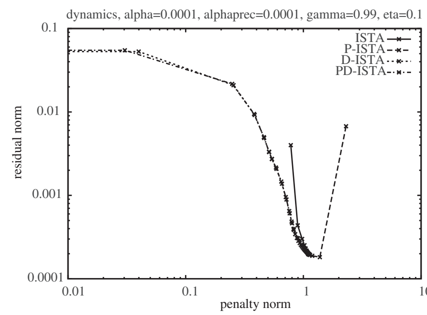


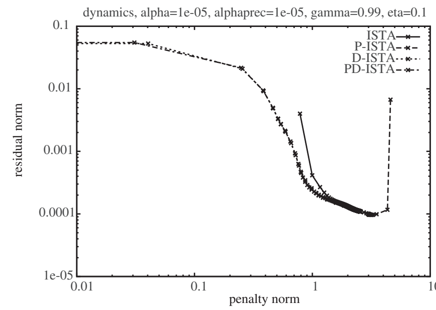
FIGURE 7. Support size histories for different iterative thresholding algorithms, $\alpha = 10^{-6}$, $\gamma = 0.99$ and $\eta = 0.1$: ISTA (*solid line*), D-ISTA (*dotted line*), P-ISTA (*dashed line*), and PD-ISTA (*dash-dotted line*)

the unpreconditioned algorithms iterate on significantly larger index sets than their preconditioned counterparts; see Figure 7.

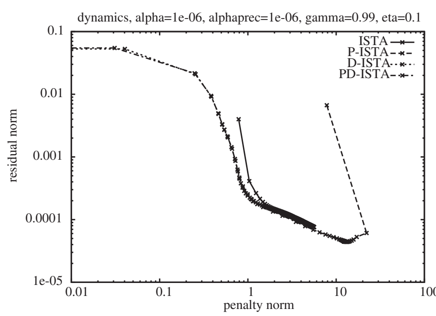
7.0.6. *Dynamics.* In order to further analyze the response of the considered thresholding algorithms, we investigate the orbit of the iterates towards the minimizer $\mathbf{u}^*$ in the so-called *dynamics plane*, i.e., the relationship between the residuals $\|A\mathbf{u}^{(n)} - y\|_{L_2(0,1)}$ and the respective penalty terms $\|\mathbf{u}^{(n)}\|_{\ell_1}$. We refer to Figure 8



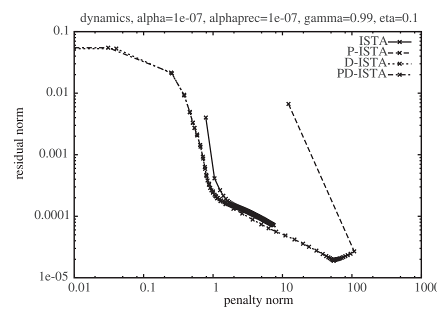
(a) $\alpha = 10^{-4}$, $\gamma = 0.99$, $\eta = 0.1$



(b) $\alpha = 10^{-5}$, $\gamma = 0.99$, $\eta = 0.1$



(c) $\alpha = 10^{-6}$, $\gamma = 0.99$, $\eta = 0.1$



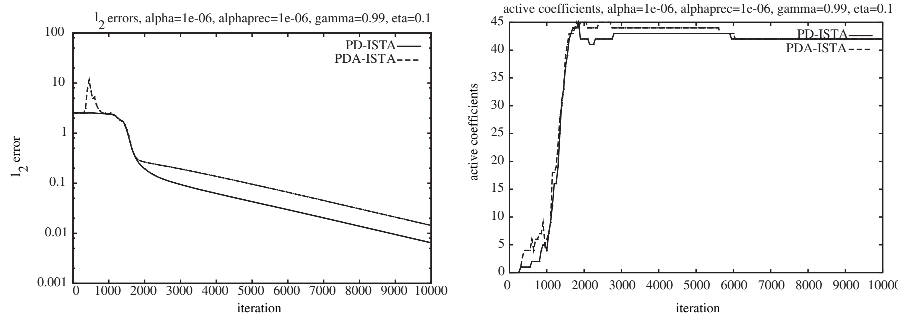
(d) $\alpha = 10^{-7}$, $\gamma = 0.99$, $\eta = 0.1$

FIGURE 8. Dynamics of the first 10000 iterations of ISTA (*solid line*), D-ISTA (*dotted line*), P-ISTA (*dashed line*), and PD-ISTA (*dash-dotted line*)

for a dynamics plot. All algorithms do have the same limit sequence $\mathbf{u}^*$, however, the associated paths in the dynamics plane look quite different. Apart from the preconditioned iterative thresholding algorithm with constant parameters (P-ISTA), which approaches the minimizer from the right, all other algorithms yield iterates with increasing penalty norms. However, since the plot only shows the first 10000 iterations, it becomes clear again that both preconditioned iterations converge much faster to the limit coefficient array than ISTA.

7.0.7. *Adaptivity.* As shown in Theorem 5.1, we have theoretical guarantees of convergence of the preconditioned algorithm only if the adaptive applications of the forward operator A^*A are implemented by the procedure **APPLY**. Moreover, the efficiency in terms of complexity of the considered iterative thresholding method can be further improved by these inexact operations. In particular, we can exploit that the biinfinite matrix A^*A is compressible in the aforementioned sense. More precisely, due to the piecewise smooth kernel of K^*K, the *compressibility exponent* s in formula (4.2) does not exceed $s^* = 1.5$, even for smooth wavelet bases.

We compare iterative thresholding algorithms with adaptive and nonadaptive operator evaluations in Figure 9, where the accuracies within the **APPLY** routine are chosen like $\tilde{\epsilon}_n = 500\eta\gamma^n\epsilon_0$.



(a) ℓ_2 error history (b) degrees of freedom

FIGURE 9. Convergence and complexity of iterative thresholding algorithms with exact and inexact operator applications for summable tolerances: PD-ISTA (*solid line*) and PDA-ISTA (*dashed line*)

Acknowledgement

The authors would like to thank Christine De Mol for a useful discussion on issues concerning the convergence to compressible solutions §6, and Manuel Werner for technical assistance.

References

1. A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. **2** (2009), no. 1, 183–202. MR2486527 (2010d:35390)
2. T. Bonesky, S. Dahlke, P. Maass, and T. Raasch, *Adaptive wavelet methods and sparsity reconstruction for inverse heat conduction problems*, Adv. Comput. Math. **33** (2010), no. 4, 385–411.
3. K. Bredies and D. A. Lorenz, *Linear convergence of iterative soft-thresholding*, J. Fourier Anal. Appl. **14** (2008), no. 5-6, 813–837. MR2461608 (2010i:65109)
4. E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), no. 12, 4203–4215. MR2243152 (2007b:94313)
5. ———, *Near optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory **52** (2006), no. 12, 5406–5425. MR2300700 (2008c:94009)
6. A. Cohen, *Numerical Analysis of Wavelet Methods.*, Studies in Mathematics and its Applications 32. Amsterdam: North-Holland., 2003. MR1990555 (2004c:65178)
7. A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations — Convergence rates*, Math. Comp. **70** (2001), 27–75. MR1803124 (2002h:65201)
8. ———, *Adaptive wavelet methods II: Beyond the elliptic case*, Found. Comput. Math. **2** (2002), no. 3, 203–245. MR1907380 (2003f:65212)
9. P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul. **4** (2005), no. 4, 1168–1200. MR2203849 (2007g:94016)
10. S. Dahlke, M. Fornasier, and T. Raasch, *Adaptive frame methods for elliptic operator equations*, Adv. Comput. Math. **27** (2007), no. 1, 27–63. MR2317920 (2008c:65366)
11. ———, *Multilevel preconditioning for adaptive sparse optimization*, Preprint 25, DFG-SPP 1324, August 2009.
12. S. Dahlke, M. Fornasier, T. Raasch, R. Stevenson, and M. Werner, *Adaptive frame methods for elliptic operator equations: The steepest descent approach*, IMA J. Numer. Anal. **27** (2007), no. 4, 717–740. MR2371829 (2008i:65239)
13. W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numerica **6** (1997), 55–228. MR1489256 (98m:65102)

14. W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math **63** (1992), 315–344. MR1186345 (93j:65065)
15. W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for pseudodifferential operators II: Matrix compression*, Adv. Comput. Math. (1993), 259–335. MR1242378 (95g:65149)
16. ______, *Multiscale methods for pseudo-differential equations on smooth manifolds*, Proceedings of the International Conference on Wavelets: Theory, Applications, and Algorithms, Academic Press, 1994, pp. 385–424. MR1321437 (96c:65208)
17. I. Daubechies, M. Defrise, and C. DeMol, *An iterative thresholding algorithm for linear inverse problems*, Comm. Pure Appl. Math. **57** (2004), no. 11, 1413–1457. MR2077704 (2005k:65106)
18. R. A. DeVore, *Nonlinear approximation*, Acta Numerica **7** (1998), 51–150. MR1689432 (2001a:41034)
19. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), no. 2, 407–499. MR2060166 (2005d:62116)
20. H.W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems.*, Mathematics and its Applications (Dordrecht). 375. Dordrecht: Kluwer Academic Publishers., 1996. MR1408680 (97k:65145)
21. M. Figueiredo, R. Nowak, and S. J. Wright, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Proc. (2009). MR2650165 (2010m:94030)
22. M. A. T. Figueiredo and R. D. Nowak, *An EM algorithm for wavelet-based image restoration.*, IEEE Trans. Image Proc. **12** (2003), no. 8, 906–916. MR2008658
23. M. Fornasier and F. Pitolli, *Adaptive iterative thresholding algorithms for magnetoencceophalography (MEG)*, J. Comput. Appl. Math. **221** (2008), no. 2, 386–395. MR2457671 (2010a:65089)
24. T. Gantumur, H. Harbrecht, and R. Stevenson, *An optimal adaptive wavelet method without coarsening of the iterands*, Math. Comp. **76** (2007), no. 258, 615–629. MR2291830 (2008i:65310)
25. G. H. Golub and C. F. van Loan, *Matrix Computations*, John Hopkins University Press, 1989.
26. M. Grasmair, M. Haltmeier, and O. Scherzer, *Sparse regularization with ℓ^q penalty term*, Inverse Problems **24** (2008), 055020 MR2438955 (2010a:65086)
27. ______, *Necessary and sufficient conditions for linear convergence of ℓ_1-regularization*, preprint (2009).
28. H. Harbrecht, T. Gantumur, and R. Stevenson, *An optimal adaptive wavelet method without coarsening of the iterands*, Mathematics of Computation **76** (2007), 615–629. MR2291830 (2008i:65310)
29. S. Jaffard, *Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications*, Ann. Inst. H. Poincaré Anal. Non Linéaire **7** (1990), no. 5, 461–476. MR1138533 (93h:47035)
30. P.-G. Lemarié, *Bases de'ondelettes sur les groupes de Lie stratifié*, Bulletin de la Société Mathématique de France **117** (1989), no. 2, 213–232.
31. A. K. Louis, *Inverse und schlechtgestellte Probleme*, Teubner, Stuttgart, 1989.
32. M. Primbs, *Stabile biorthogonale Spline-Waveletbasen auf dem Intervall*, Dissertation, Universität Duisburg-Essen, 2006.
33. R. Ramlau, G. Teschke, and M. Zhariy, *A compressive Landweber iteration for solving ill-posed inverse problems*, Inverse Problems **24** (2008), no. 6, 065013. MR2456960 (2010a:65218)
34. H. Rauhut, *Compressive sensing and structured random matrices*, Theoretical Foundations and Numerical Methods for Sparse Recovery (M. Fornasier, ed.), Radon Series Comp. Appl. Math., deGruyter, 2010.
35. R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur effizienten Lösung grosser vollbesetzter Gleichungssysteme*, Advances in Numerical Mathematics, Teubner Stuttgart, 1998. MR1623209 (99f:65067)
36. J.-L. Starck, E. J. Candès, and D. L. Donoho, *Astronomical image representation by curvelet transform*, Astronomy and Astrophysics **298** (2003), 785–800.
37. J.-L. Starck, M. K. Nguyen, and F. Murtagh, *Wavelets and curvelets for image deconvolution: a combined approach*, Signal Proc. **83** (2003), 2279–2283.
38. R. Stevenson, *Adaptive solution of operator equations using wavelet frames*, SIAM J. Numer. Anal **41** (2003), no. 3, 1074–1100. MR2005196 (2004e:42062)

Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans Meerwein Strasse, Lahnberge, 35032 Marburg, Germany
E-mail address: dahlke@mathematik.uni-marburg.de

Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040, Linz, Austria
Current address: Faculty of Mathematics, Technical University of Munich, Boltzmannstrasse 3, D-85748 Garching, Germany
E-mail address: massimo.fornasier@ma.tum.de

Johannes Gutenberg-Universität, Institut für Mathematik, Staudingerweg 9, 55099 Mainz, Germany
E-mail address: raasch@uni-mainz.de