J Fourier Anal Appl (2008) 14: 764–792
DOI 10.1007/s00041-008-9039-8

Accelerated Projected Gradient Method for Linear Inverse Problems with Sparsity Constraints

Ingrid Daubechies · Massimo Fornasier · Ignace Loris

Received: 28 June 2007 / Revised: 8 October 2007 / Published online: 17 September 2008

Abstract Regularization of ill-posed linear inverse problems via ℓ_1 penalization has been proposed for cases where the solution is known to be (almost) sparse. One way to obtain the minimizer of such an ℓ_1 penalized functional is via an iterative soft-thresholding algorithm. We propose an alternative implementation to ℓ_1-constraints, using a gradient method, with projection on ℓ_1-balls. The corresponding algorithm uses again iterative soft-thresholding, now with a variable thresholding parameter. We also propose accelerated versions of this iterative method, using ingredients of the (linear) steepest descent method. We prove convergence in norm for one of these projected gradient methods, without and with acceleration.

Keywords Linear inverse problems · Sparse recovery · Projected gradient method

Mathematics Subject Classification (2000) 15A29 · 49M30 · 65F22 · 65K10 · 90C25 · 52A41

1 Introduction

Our main concern in this paper is the construction of iterative algorithms to solve inverse problems with an ℓ_1-penalization or an ℓ_1-constraint, and that converge faster

Communicated by Ronald A. DeVore.

I. Daubechies · M. Fornasier (✉)
Program in Computational and Applied Mathematics, Princeton University, Fine Hall, Washington Road 08544-1000 Princeton, NJ, USA
e-mail: mfornasi@math.princeton.edu

I. Daubechies
e-mail: ingrid@math.princeton.edu

I. Loris
Dienst Theoretische Natuurkunde, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium
e-mail: igloris@vub.ac.be

than the iterative algorithm proposed in [22] (see also formulas (2.5) and (2.6) below). Before we get into technical details, we introduce here the background, framework, and notations for our work.

In many practical problems, one cannot observe directly the quantities of most interest; instead their values have to be inferred from their effect on observable quantities. When this relationship between observable y and interesting quantity f is (approximately) linear, as it is in surprisingly many cases, the situation can be modeled mathematically by the equation

$$y = Af, \tag{1.1}$$

where A is a linear operator mapping a vector space $\mathcal{K}$ (which we assume to contain all possible "objects" f) to a vector space $\mathcal{H}$ (which contains all possible data y). The vector spaces $\mathcal{K}$ and $\mathcal{H}$ can be finite- or infinite-dimensional; in the latter case, we assume that $\mathcal{K}$ and $\mathcal{H}$ are (separable) Hilbert spaces, and that $A : \mathcal{K} \to \mathcal{H}$ is a bounded linear operator. Our main goal consists in reconstructing the (unknown) element $f \in \mathcal{K}$, when we are given y. If A is a "nice", easily invertible operator, and if the data y are free of noise, then this is a trivial task. Often, however, the mapping A is ill-conditioned or not invertible. Moreover, typically (1.1) is only an idealized version in which noise has been neglected; a more accurate model is

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

in which the data are corrupted by an (unknown) noise. In order to deal with this type of reconstruction problem a *regularization* mechanism is required [30]. Regularization techniques try, as much as possible, to take advantage of (often vague) prior knowledge one may have about the nature of f. The approach in this paper is tailored to the case when f can be represented by a *sparse* expansion, i.e., when f can be represented by a series expansion with respect to an orthonormal basis or a frame [11, 20] that has only a small number of large coefficients. In this paper, as in [22], we model the sparsity constraint by adding an ℓ_1-term to a functional to be minimized; it was shown in [22] that this assumption does indeed correspond to a regularization scheme.

Several types of signals appearing in nature admit sparse frame expansions and thus, sparsity is a realistic assumption for a very large class of problems. For instance, natural images are well approximated by sparse expansions with respect to wavelets or curvelets [4, 20].

Sparsity has had already a long history of successes. The design of frames for sparse representations of digital signals has led to extremely efficient compression methods, such as JPEG2000 and MP3 [39]. A new generation of optimal numerical schemes has been developed for the computation of sparse solutions of differential and integral equations, exploiting adaptive and greedy strategies [12–14, 18, 19]. The use of sparsity in inverse problems for data recovery is the most recent step of this concept's long career of "simplifying and understanding complexity", with an enormous potential in applications [2, 9, 15, 17, 21–25, 33–36, 38, 40, 44]. In particular, the observation that it is possible to reconstruct sparse signals from vastly incomplete information just seeking for the ℓ_1-minimal solutions [6, 8, 26, 41] has led to a new line of research called *sparse recovery* or *compressed sensing*, with very fruitful mathematical and applied results.

2 Framework and Notations

Before starting our discussion let us briefly introduce some of the notations we will need. For some countable index set Λ we denote by $\ell_p = \ell_p(\Lambda)$, $1 \le p \le \infty$, the space of real sequences $x = (x_\lambda)_{\lambda\in\Lambda}$ with norm

$$\|x\|_p := \left(\sum_{\lambda\in\Lambda} |x_\lambda|^p\right)^{1/p}, \quad 1 \le p < \infty$$

and $\|x\|_\infty := \sup_{\lambda\in\Lambda} |x_\lambda|$ as usual. For simplicity of notation, in the following $\|\cdot\|$ will denote the ℓ_2-norm $\|\cdot\|_2$.

As is customary for an index set, we assume we have a natural enumeration order for the elements of Λ, using (implicitly) a one-to-one map $\mathcal{N}$ from Λ to $\mathbb{N}$. In some convergence proofs, we shall use the shorthand notations $|\lambda|$ for $\mathcal{N}(\lambda)$, and (in the case where Λ is infinite) $\lambda \to \infty$ for $\mathcal{N}(\lambda) \to \infty$.

We also assume that we have a suitable frame $\{\psi_\lambda : \lambda \in \Lambda\} \subset \mathcal{K}$ indexed by the countable set Λ. This means that there exist constants $c_1, c_2 > 0$ such that

$$c_1\|f\|^2_{\mathcal{K}} \le \sum_{\lambda\in\Lambda} |\langle f, \psi_\lambda\rangle|^2 \le c_2\|f\|^2_{\mathcal{K}}, \quad \text{for all } f \in \mathcal{K}. \tag{2.1}$$

Orthonormal bases are particular examples of frames, but there also exist many interesting frames in which the ψ_λ are not linearly independent. Frames allow for a (stable) series expansion of any $f \in \mathcal{K}$ of the form

$$f = \sum_{\lambda\in\Lambda} x_\lambda \psi_\lambda =: \mathrm{F}x, \tag{2.2}$$

where $x = (x_\lambda)_{\lambda\in\Lambda} \in \ell_2(\Lambda)$. The linear operator $\mathrm{F} : \ell_2(\Lambda) \to \mathcal{K}$ (called the *synthesis map* in frame theory) is bounded because of (2.1). When $\{\psi_\lambda : \lambda \in \Lambda\}$ is a frame but not a basis, the coefficients x_λ need not be unique. For more details on frames and their differences from bases we refer to [11].

We shall assume that f is sparse, i.e., that f can be written by a series of the form (2.2) with only a small number of non-vanishing coefficients x_λ with respect to the frame $\{\psi_\lambda\}$, or that f is *compressible*, i.e., that f can be well-approximated by such a sparse expansion. This can be modeled by assuming that the sequence x is contained in a (weighted) $\ell_1(\Lambda)$-space. Indeed, the minimization of the $\ell_1(\Lambda)$ norm promotes such sparsity. (This has been known for many years, and put to use in a wide range of applications, most notably in statistics. David Donoho calls one form of it the Logan phenomenon in [28]—see also [27]—after its first observation by Ben Logan [37].) These considerations lead us to model the reconstruction of a sparse f as the minimization of the following functional:

$$F_\tau(x) = \|Kx - y\|^2_{\mathcal{H}} + 2\tau\|x\|_1, \tag{2.3}$$

where we will assume that the data y and the linear operator $K := A \circ \mathrm{F} : \ell_2(\Lambda) \to \mathcal{H}$ are given. The second term in (2.3) is often called the *penalization* or *regularizing* term; the first term goes by the name of *discrepancy*,

$$D(x) := \|Kx - y\|^2_{\mathcal{H}}. \tag{2.4}$$

In what follows we shall drop the subscript $\mathcal{H}$, because the space in which we work will always be clear from the context. We discuss the problem of finding (approximations to) $\bar{x}(\tau)$ in $\ell_2(\Lambda)$ that minimize the functional (2.3). (We adopt the usual convention that for $u \in \ell_2(\Lambda) \setminus \ell_1(\Lambda)$, the penalty term "equals" ∞, and that, for such u, $F_\tau(u) > F_\tau(x)$ for all $x \in \ell_1(\Lambda)$. Since we want to minimize F_τ, we shall consider, implicitly, only $x \in \ell_1(\Lambda)$.) The solutions $\bar{f}(\tau)$ to the original problem are then given by $\bar{f}(\tau) = \mathrm{F}\bar{x}(\tau)$.

Several authors have proposed independently an iterative soft-thresholding algorithm to approximate the solution $\bar{x}(\tau)$ [29, 31, 42, 43]. More precisely, $\bar{x}(\tau)$ is the limit of sequences $x^{(n)}$ defined recursively by

$$x^{(n+1)} = \mathbb{S}_\tau \left[x^{(n)} + K^* y - K^* K x^{(n)} \right], \tag{2.5}$$

starting from an arbitrary $x^{(0)}$, where $\mathbb{S}_\tau$ is the soft-thresholding operation defined by $\mathbb{S}_\tau(x)_\lambda = S_\tau(x_\lambda)$ with

$$S_\tau(x) = \begin{cases} x - \tau, & x > \tau, \\ 0, & |x| \le \tau, \\ x + \tau, & x < -\tau. \end{cases} \tag{2.6}$$

Convergence of this algorithm was proved in [22]. Soft-thresholding plays a role in this problem because it leads to the unique minimizer of a functional combining ℓ_2 and ℓ_1-norms, i.e., (see [10, 22])

$$\mathbb{S}_\tau(a) = \arg \min_{x \in \ell_2(\Lambda)} \left(\|x - a\|^2 + 2\tau \|x\|_1 \right). \tag{2.7}$$

We will call the iteration (2.5) the *iterative soft-thresholding algorithm* or the *thresholded Landweber iteration*.

3 Discussion of the Thresholded Landweber Iteration

The problem of finding the sparsest solution to the under-determined linear equation $Kx = y$ is a hard combinatorial problem, not tractable numerically except in relatively low dimensions. For some classes of K, however, one can prove that the problem reduces to the convex optimization problem of finding the solution with the smallest ℓ_1 norm [6–8, 26]. Even for K outside this class, ℓ_1-minimization seems to lead to very good approximations to the sparsest solutions. It is in this sense that an algorithm of type (2.5) could conceivably be called 'fast': it is fast compared to a brute-force exhaustive search for the sparsest x.

A more honest evaluation of the speed of convergence of algorithm (2.5) is a comparison with *linear* solvers that minimize the corresponding ℓ_2 penalized functional, such as, e.g., the conjugate gradient method. One finds, in practice, that the

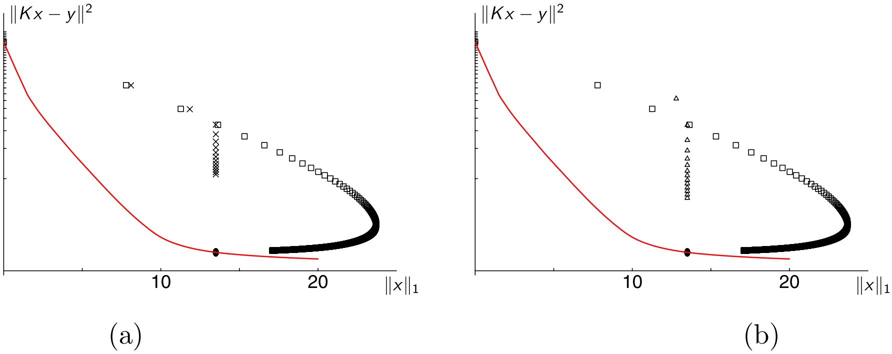

Fig. 1 The path, in the $\|x\|_1$ vs. $\|Kx - y\|^2$ plane, followed by the iterates $x^{(n)}$ of three different iterative algorithms. The operator K and the data y are taken from a seismic tomography problem [38] (see also Sect. 6). The boxes (in both (**a**) and (**b**)) correspond to the thresholded Landweber algorithm. In this example, iterative thresholded Landweber (2.5) first overshoots the ℓ_1 norm of the limit (represented by the *fat dot*), and then requires a large number of iterations to reduce $\|x^{(n)}\|_1$ again (500 are shown in this figure). In (**a**) the *crosses* correspond to the path followed by the iterates of the projected Landweber iteration (3.1); in (**b**) the *triangles* correspond to the projected steepest descent iteration (3.2); in both cases, only 15 iterates are shown. The discrepancy decreases more quickly for projected steepest descent than for the projected Landweber algorithm. How this translates into faster convergence (in norm) is discussed in Sect. 6. The *solid line* corresponds to the limit *trade-off curve*, generated by $\bar{x}(\tau)$ for decreasing values of $\tau > 0$. The *vertical axes* uses a logarithmic scale for clarity

thresholded Landweber iteration (2.5) is not competitive at all in this comparison. It is, after all, the composition of thresholding with the (linear) Landweber iteration $x^{(n+1)} = x^{(n)} + K^*y - K^*Kx^{(n)}$, which is a gradient descent algorithm with a fixed step size, known to converge usually quite slowly; interleaving it with the nonlinear thresholding operation does unfortunately not change this slow convergence. On the other hand, this nonlinearity did foil our attempts to "borrow a leaf" from standard linear steepest descent methods by using an adaptive step length—once we start taking larger steps, the algorithm seems to no longer converge in at least some numerical experiments.

We take a closer look at the characteristic dynamics of the thresholded Landweber iteration in Fig. 1. As this plot of the discrepancy $\mathcal{D}(x^{(n)}) = \|Kx^{(n)} - y\|^2$ versus $\|x^{(n)}\|_1$ shows, the algorithm converges initially relatively fast, then it overshoots the value $\|\bar{x}(\tau)\|_1$ (where $\bar{x}(\tau) := \lim_{n\to\infty} x^{(n)}$), and it takes very long to re-correct back. In other words, starting from $x^{(0)} = 0$, the algorithm generates a path $\{x^{(n)}; n \in \mathbb{N}\}$ that is initially fully contained in the ℓ_1-ball $B_R := \{x \in \ell_2(\Lambda); \|x\|_1 \le R\}$, with $R := \|\bar{x}(\tau)\|_1$. Then it gets out of the ball to slowly inch back to it in the limit. A first intuitive way to avoid this long "external" detour is to force the successive iterates to remain within the ball B_R. One method to achieve this is to substitute for the thresholding operations the projection $\mathbb{P}_{B_R}$, where, for any closed convex set C, and any x, we define $\mathbb{P}_C(x)$ to be the unique point in C for which the ℓ_2-distance to x is minimal. With a slight abuse of notation, we shall denote $\mathbb{P}_{B_R}$ by $\mathbb{P}_R$; this will not cause confusion, because it will be clear from the context whether the subscript of $\mathbb{P}$ is a set or a positive number. We thus obtain the following algorithm: Pick an arbitrary $x^{(0)} \in \ell_2(\Lambda)$, for example $x^{(0)} = 0$, and iterate

$$x^{(n+1)} = \mathbb{P}_R \left[x^{(n)} + K^* y - K^* K x^{(n)} \right]. \tag{3.1}$$

We will call this the *projected Landweber iteration*.

The typical dynamics of this projected Landweber algorithm are illustrated in Fig. 1a. The norm $\|x^{(n)}\|_1$ no longer overshoots R, but quickly takes on the limit value (i.e., $\|\bar{x}(\tau)\|_1$); the speed of convergence remains very slow, however. In this projected Landweber iteration case, modifying the iterations by introducing an adaptive "descent parameter" $\beta^{(n)} > 0$ in each iteration, defining $x^{(n+1)}$ by

$$x^{(n+1)} = \mathbb{P}_R \left[x^{(n)} + \beta^{(n)} K^* (y - K x^{(n)}) \right], \tag{3.2}$$

does lead, in numerical simulations, to promising, converging results (in which it differs from the soft-thresholded Landweber iteration, where introducing such a descent parameter did not lead to numerical convergence, as noted above).

The typical dynamics of this modified algorithm are illustrated in Fig. 1b, which clearly shows the larger steps and faster convergence (when compared with the projected Landweber iteration in Fig. 1a). We shall refer to this modified algorithm as the *projected gradient iteration* or the *projected steepest descent*; it will be the main topic of this paper.

The main issue is to determine how large we can choose the successive $\beta^{(n)}$, and still prove norm convergence of the algorithm in $\ell_2(\Lambda)$.

There exist results in the literature on convergence of projected gradient iterations, where the projections are (as they are here) onto convex sets, see, e.g., [1, 16] and references therein. These results treat iterative projected gradient methods in much greater generality than we need: they allow more general functionals than $\mathcal{D}$, and the convex set on which the iterative procedure projects need not be bounded. On the other hand, these general results typically have the following restrictions:

- The convergence in infinite-dimensional Hilbert spaces (i.e., Λ is countable but infinite) is proved only in the weak sense and often only for subsequences;
- In [1] the descent parameters are typically restricted to cases for which $\lim_{n\to\infty} \beta^{(n)} = 0$. In [16], it is shown that the algorithm converges weakly for any choice of $\beta^{(n)} \in [\varepsilon, \frac{2-\varepsilon}{\|K\|}]$, for $\varepsilon > 0$ arbitrarily small. Of most interest to us is the case where the $\beta^{(n)}$ are picked adaptively, can grow with n, and are not limited to values below $\frac{2}{\|K\|}$; this case is not covered by the methods of either [1] or [16].

To our knowledge there are no results in the literature for which the whole sequence $(x^{(n)})_{n\in\mathbb{N}}$ converges in the Hilbert space norm to a unique accumulation point, for "descent parameters" $\beta^{(n)} \geq 2$. It is worthwhile emphasizing that strong convergence is not automatic: in [16, Remark 5.12], the authors provide a counterexample in which strong convergence fails. (This question had been open for some time.) One of the main results of this paper is to prove a theorem that establishes exactly this type of convergence; see Theorem 1 below. Moreover, the result is achieved by imposing a choice of $\beta^{(n)} \geq 1$ which ensures a monotone decay of a suitable energy. This establishes a principle of *best descent* similar to the well-known steepest-descent in unconstrained minimization.

Before we get to this theorem, we need to build some more machinery first.

4 Projections onto ℓ_1-Balls via Thresholding Operators

In this section we discuss some properties of ℓ_2-projections onto ℓ_1-balls. In particular, we investigate their relations with thresholding operators and their explicit computation. We also estimate the time complexity of such projections in finite dimensions.

We first observe a useful property of the soft-thresholding operator.

Lemma 1 *For any fixed $a \in \ell_2(\Lambda)$ and for $\tau > 0$, $\|\mathbb{S}_\tau(a)\|_1$ is a piecewise linear, continuous, decreasing function of τ; moreover, if $a \in \ell_1(\Lambda)$ then $\|\mathbb{S}_0(a)\|_1 = \|a\|_1$ and $\|\mathbb{S}_\tau(a)\|_1 = 0$ for $\tau \geq \max_i |a_i|$.*

Proof $\|\mathbb{S}_\tau(a)\|_1 = \sum_\lambda |S_\tau(a_\lambda)| = \sum_\lambda S_\tau(|a_\lambda|) = \sum_{|a_\lambda|>\tau}(|a_\lambda| - \tau)$; the sum in the right hand side is finite for $\tau > 0$. □

A schematic illustration is given in Fig. 2.

The following lemma shows that the ℓ_2 projection $\mathbb{P}_R(a)$ can be obtained by a suitable thresholding of a.

Lemma 2 *If $\|a\|_1 > R$, then the ℓ_2 projection of a on the ℓ_1 ball with radius R is given by $\mathbb{P}_R(a) = \mathbb{S}_\mu(a)$ where μ (depending on a and R) is chosen such that $\|\mathbb{S}_\mu(a)\|_1 = R$. If $\|a\|_1 \leq R$ then $\mathbb{P}_R(a) = \mathbb{S}_0(a) = a$.*

Proof Suppose $\|a\|_1 > R$. Because, by Lemma 1, $\|\mathbb{S}_\mu(a)\|_1$ is continuous in μ and $\|\mathbb{S}_\mu(a)\|_1 = 0$ for sufficiently large μ, we can choose μ such that $\|\mathbb{S}_\mu(a)\|_1 = R$. (See Fig. 2.) On the other hand (see above, or [10, 22]), $b = \mathbb{S}_\mu(a)$ is the unique minimizer of $\|x - a\|^2 + 2\mu\|x\|_1$, i.e.,

$$\|b - a\|^2 + 2\mu\|b\|_1 < \|x - a\|^2 + 2\mu\|x\|_1$$

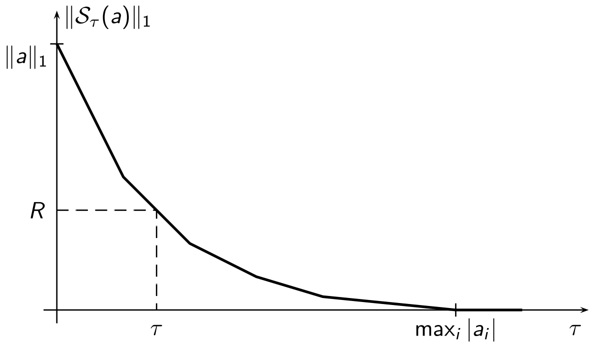

Fig. 2 For a given vector $a \in \ell_2$, $\|\mathbb{S}_\tau(a)\|_1$ is a piecewise linear continuous and decreasing function of τ (strictly decreasing for $\tau < \max_i |a_i|$). The knots are located at $\{|a_i|, i : 1, \ldots, m\}$ and 0. Finding τ such that $\|\mathbb{S}_\tau(a)\|_1 = R$ ultimately comes down to a linear interpolation. The figure is made for the finite dimensional case

for all $x \neq b$. Since $\|b\|_1 = R$, it follows that

$$\forall x \in B_R, \quad x \neq b : \|b-a\|^2 < \|x-a\|^2.$$

Hence b is closer to a than any other x in B_R. In other words, $\mathbb{P}_R(a) = b = \mathbb{S}_\mu(a)$. □

These two lemmas prescribe the following simple recipe for computing the projection $\mathbb{P}_R(a)$. In a first step, sort the absolute values of the components of a (an $\mathcal{O}(m \log m)$ operation if $\#\Lambda = m$ is finite), resulting in the rearranged sequence $(a^*_\ell)_{\ell=1,\dots,m}$, with $a^*_\ell \geq a^*_{\ell+1} \geq 0$ for all ℓ. Next, perform a search to find k such that

$$\|\mathbb{S}_{a^*_k}(a)\|_1 = \sum_{\ell=1}^{k-1} \left(a^*_\ell - a^*_k\right) \leq R < \sum_{\ell=1}^{k} \left(a^*_\ell - a^*_{k+1}\right) = \|\mathbb{S}_{a^*_{k+1}}(a)\|_1$$

or equivalently,

$$\|\mathbb{S}_{a^*_k}(a)\|_1 = \sum_{\ell=1}^{k-1} \ell \left(a^*_\ell - a^*_{\ell+1}\right) \leq R < \sum_{\ell=1}^{k} \ell \left(a^*_\ell - a^*_{\ell+1}\right) = \|\mathbb{S}_{a^*_{k+1}}(a)\|_1;$$

the complexity of this step is again $\mathcal{O}(m \log m)$. Finally, set $\nu := k^{-1}(R - \|\mathbb{S}_{a^*_k}(a)\|_1)$, and $\mu := a^*_k + \nu$. Then

$$\begin{aligned}\|\mathbb{S}_\mu(a)\|_1 &= \sum_{i \in \Lambda} \max(|a_i| - \mu, 0) = \sum_{\ell=1}^{k} \left(a^*_\ell - \mu\right) \\ &= \sum_{\ell=1}^{k-1} \left(a^*_\ell - a^*_k\right) + k\nu = \|\mathbb{S}_{a^*_k}(a)\|_1 + k\nu = R.\end{aligned}$$

These formulas were also derived in [36, Lemmas 4.1 and 4.2], by observing that $\mathbb{P}_R(a) = a - \mathbb{S}^\infty_R(a)$, where

$$\mathbb{S}^\infty_R(a) = \arg\min_{x \in \mathbb{R}^m} (\|x-a\|^2 + 2R\|x\|_\infty), \quad x \in \mathbb{R}^m. \tag{4.1}$$

The latter is again a thresholding operator, but it is related to an ℓ_∞ penalty term. Similar descriptions of the ℓ_2 projection onto ℓ_1 balls appear also in [5].

Finally, $\mathbb{P}_R$ has the following additional properties:

Lemma 3 *For any $x \in \ell_2(\Lambda)$, $\mathbb{P}_R(x)$ is characterized as the unique vector in B_R such that*

$$\langle w - \mathbb{P}_R(x), x - \mathbb{P}_R(x)\rangle \leq 0, \quad \textit{for all } w \in B_R. \tag{4.2}$$

Moreover the projection $\mathbb{P}_R$ is non-expansive:

$$\|\mathbb{P}_R(x) - \mathbb{P}_R(x')\| \leq \|x - x'\| \tag{4.3}$$

for all $x, y \in \ell_2(\Lambda)$.

The proof is standard for projection operators onto convex sets; we include it because its technique will be used often in this paper.

Proof Because B_R is convex, $(1-t)\mathbb{P}_R(x)+tw\in B_R$ for all $w\in B_R$ and $t\in[0,1]$. It follows that $\|x-\mathbb{P}_R(x)\|^2\le\|x-[(1-t)\mathbb{P}_R(x)+tw]\|^2$ for all $t\in[0,1]$. This implies

$$0\le -2t\langle w-\mathbb{P}_R(x),x-\mathbb{P}_R(x)\rangle+t^2\|w-\mathbb{P}_R(x)\|^2$$

for all $t\in[0,1]$. It follows that

$$\langle w-\mathbb{P}_R(x),x-\mathbb{P}_R(x)\rangle\le 0,$$

which proves (4.2).

Setting $w=\mathbb{P}_R(x')$ in (4.2), we get, for all x,x',

$$\langle\mathbb{P}_R(x')-\mathbb{P}_R(x),x-\mathbb{P}_R(x)\rangle\le 0.$$

Switching the role of x and x' one finds:

$$\langle\mathbb{P}_R(x')-\mathbb{P}_R(x),x'-\mathbb{P}_R(x')\rangle\ge 0.$$

By combining these last two inequalities, one finds:

$$\langle\mathbb{P}_R(x')-\mathbb{P}_R(x),x'-x-\mathbb{P}_R(x')+\mathbb{P}_R(x)\rangle\ge 0$$

or

$$\|\mathbb{P}_R(x')-\mathbb{P}_R(x)\|^2\le\langle\mathbb{P}_R(x')-\mathbb{P}_R(x),x'-x\rangle;$$

by Cauchy-Schwarz this gives

$$\|\mathbb{P}_R(x')-\mathbb{P}_R(x)\|^2\le\langle\mathbb{P}_R(x')-\mathbb{P}_R(x),x'-x\rangle\le\|\mathbb{P}_R(x')-\mathbb{P}_R(x)\|\,\|x'-x\|,$$

from which inequality (4.3) follows. □

5 The Projected Gradient Method

We have now collected all the terminology needed to identify some conditions on the $\beta^{(n)}$ that will ensure convergence of the $x^{(n)}$, defined by (3.2), to $\tilde{x}_R$, the minimizer in B_R of $\mathcal{D}(x)=\|Kx-y\|^2$. For notational simplicity we set $r^{(n)}=K^*(y-Kx^{(n)})$. With this notation, the thresholded Landweber iteration (2.5) can be written as

$$x^{(n+1)}=\mathbb{S}_\tau\big(x^{(n)}+r^{(n)}\big). \tag{5.1}$$

As explained above, we consider, instead of straightforward soft-thresholding with fixed τ, adapted soft-thresholding operations $\mathbb{S}_{\mu(R,x^{(n)}+r^{(n)})}$ that correspond to the projection operator $\mathbb{P}_R$:

$$x^{(n+1)}=\mathbb{P}_R\big(x^{(n)}+r^{(n)}\big). \tag{5.2}$$

The dependence of $\mu(R, x^{(n)}+r^{(n)})$ on R is described above; R is kept fixed throughout the iterations. If, for a given value of τ, R were picked such that $R = R_\tau := \|\bar{x}_\tau\|_1$ (where $\bar{x}_\tau$ is the minimizer of $\|Kx - y\|^2 + 2\tau\|x\|_1$), then Lemma 2 would ensure that $\bar{x}_\tau = \tilde{x}_R$. Of course, we don't know, in general, the exact value of $\|\bar{x}_\tau\|_1$, so that we can't use it as a guideline to pick R. In practice, however, it is customary to determine $\bar{x}_\tau$ for a range of τ-values; this then amounts to the same as determining $\tilde{x}_R$ for a range of R.

We now propose to change the step $r^{(n)}$ into a step $\beta^{(n)}r^{(n)}$ (in the spirit of the "classical" steepest descent method), and to define the algorithm: Pick an arbitrary $x^{(0)} \in \ell_2(\Lambda)$, for example $x^{(0)} = 0$, and iterate

$$x^{(n+1)} = \mathbb{P}_R\big(x^{(n)} + \beta^{(n)}r^{(n)}\big). \tag{5.3}$$

In this section we prove the norm convergence of this algorithm to a minimizer $\tilde{x}_R$ of $\|Kx - y\|^2$ in B_R, under some assumptions on the descent parameters $\beta^{(n)} \geq 1$.

5.1 General Properties

We begin with the following characterization of the minimizers of $\mathcal{D}$ on B_R.

Lemma 4 *The vector $\tilde{x}_R \in \ell_2(\Lambda)$ is a minimizer of $\mathcal{D}(x) = \|Kx - y\|^2$ on B_R if and only if*

$$\mathbb{P}_R(\tilde{x}_R + \beta K^*(y - K\tilde{x}_R)) = \tilde{x}_R, \tag{5.4}$$

for any $\beta > 0$, which in turn is equivalent to the requirement that

$$\langle K^*(y - K\tilde{x}_R), w - \tilde{x}_R\rangle \leq 0, \quad \textit{for all } w \in B_R. \tag{5.5}$$

To lighten notation, we shall drop the subscript R on $\tilde{x}_R$ whenever no confusion is possible.

Proof If $\tilde{x}$ minimizes $\mathcal{D}$ on B_R, then for all $w \in B_R$, and for all $t \in [0, 1]$,

$$\begin{aligned} \mathcal{D}(\tilde{x}) &\leq \mathcal{D}((1-t)\tilde{x} + tw), \quad \text{or} \\ \|K\tilde{x} - y\|^2 &\leq \|K\tilde{x} - y + tK(w - \tilde{x})\|^2, \quad \text{or} \\ 0 &\leq 2t\langle K\tilde{x} - y, K(w - \tilde{x})\rangle + t^2\|K(w - \tilde{x})\|^2. \end{aligned}$$

This implies

$$\langle K^*(y - K\tilde{x}), w - \tilde{x}\rangle \leq 0. \tag{5.6}$$

It follows from this that, for all $w \in B_R$ and for all $\beta > 0$,

$$\langle \tilde{x} + \beta K^*(y - K\tilde{x}) - \tilde{x}, w - \tilde{x}\rangle \leq 0. \tag{5.7}$$

By Lemma 3 this implies (5.4).

Conversely, if $\mathbb{P}_R(\tilde{x} + \beta K^*(y - K\tilde{x})) = \tilde{x}$, then for all $w \in B_R$ and for all $t \in [0, 1]$:

$$\|(\tilde{x} + \beta K^*(y - K\tilde{x})) - ((1-t)\tilde{x} + tw)\|^2 \geq \|(\tilde{x} + \beta K^*(y - K\tilde{x})) - \tilde{x}\|^2, \quad \text{or}$$
$$\|\beta K^*(y - K\tilde{x}) + t(\tilde{x} - w)\|^2 \geq \|\beta K^*(y - K\tilde{x})\|^2 \quad \Longrightarrow$$
$$2t\beta\langle K^*(y - K\tilde{x}), \tilde{x} - w\rangle + t^2\|\tilde{x} - w\|^2 \geq 0.$$

This implies

$$\langle K^*(y - K\tilde{x}), \tilde{x} - w\rangle \geq 0 \quad \text{or} \quad \langle y - K\tilde{x}, K(\tilde{x} - w)\rangle \geq 0.$$

In other words:

$$-\|y - K\tilde{x}\|^2 - \|K\tilde{x} - Kw\|^2 + \|(y - K\tilde{x}) + K(\tilde{x} - w)\|^2 \geq 0$$

or

$$\mathcal{D}(\tilde{x}) + \|K(\tilde{x} - w)\|^2 \leq \mathcal{D}(w).$$

This implies that $\tilde{x}$ minimizes $\mathcal{D}$ on B_R. □

The minimizer of $\mathcal{D}$ on B_R need not be unique. We have, however

Lemma 5 *If $\tilde{x}, \tilde{\tilde{x}}$ are two distinct minimizers of $\mathcal{D}(x) = \|Kx - y\|^2$ on B_R, then $K\tilde{x} = K\tilde{\tilde{x}}$, i.e., $\tilde{x} - \tilde{\tilde{x}} \in \ker K$.*

Conversely, if $\tilde{x}, \tilde{\tilde{x}} \in B_R$, if $\tilde{x}$ minimizes $\|Kx - y\|^2$ and if $\tilde{x} - \tilde{\tilde{x}} \in \ker K$ then $\tilde{\tilde{x}}$ minimizes $\|Kx - y\|^2$ as well.

Proof The converse is obvious; we prove only the direct statement. From the last inequality in the proof of Lemma 4 we obtain $\mathcal{D}(\tilde{x}) + \|K(\tilde{x} - \tilde{\tilde{x}})\|^2 \leq \mathcal{D}(\tilde{\tilde{x}}) = \mathcal{D}(\tilde{x})$, which implies $\|K(\tilde{\tilde{x}} - \tilde{x})\| = 0$. □

In what follows we shall assume that the minimizers of $\mathcal{D}$ in B_R are not global minimizers for $\mathcal{D}$, i.e., that $K^*(y - K\tilde{x}) \neq 0$. We know from Lemma 2 that $\mathbb{P}_R(a)$ can be computed for $\|a\|_1 > R$ simply by finding the value $\mu > 0$ such that $\|\mathbb{S}_\mu(a)\|_1 = R$; one has then $\mathbb{P}_R(a) = \mathbb{S}_\mu(a)$. Using this we prove

Lemma 6 *Let u be the common image under K of all minimizers of $\mathcal{D}$ on B_R, i.e., for all $\tilde{x}$ minimizing $\mathcal{D}$ in B_R, $K\tilde{x} = u$. Then there exists a unique value $\tau > 0$ such that, for all $\beta > 0$ and for all minimizing $\tilde{x}$*

$$\mathbb{P}_R(\tilde{x} + \beta K^*(y - u)) = \mathbb{S}_{\tau\beta}(\tilde{x} + \beta K^*(y - u)). \tag{5.8}$$

Moreover, for all $\lambda \in \Lambda$ we have that if there exists a minimizer $\tilde{x}$ such that $\tilde{x}_\lambda \neq 0$, then

$$|(K^*(y - u))_\lambda| = \tau. \tag{5.9}$$

Proof From Lemmas 2 and 4, we know that for each minimizing $\tilde{x}$, and each $\beta > 0$, there exists a unique $\mu(\tilde{x}, \beta)$ such that

$$\tilde{x} = \mathbb{P}_R(\tilde{x} + \beta K^*(y-u)) = \mathbb{S}_{\mu(\tilde{x},\beta)}(\tilde{x} + \beta K^*(y-u)). \tag{5.10}$$

For $\tilde{x}_\lambda \neq 0$ we have $\tilde{x}_\lambda = \tilde{x}_\lambda + \beta(K^*(y-u))_\lambda - \mu(\tilde{x},\beta)\mathrm{sgn}\,\tilde{x}_\lambda$; this implies $\mathrm{sgn}\,\tilde{x}_\lambda = \mathrm{sgn}\,(\tilde{x}_\lambda + \beta(K^*(y-u))_\lambda)$ and also that $|(K^*(y-u))_\lambda| = \frac{1}{\beta}\mu(\tilde{x},\beta)$. If $\tilde{x}_\lambda = 0$ then $|(K^*(y-u))_\lambda| \le \frac{1}{\beta}\mu(\tilde{x},\beta)$. It follows that $\tau := \mu(\tilde{x},\beta)/\beta = \|K^*(y-u)\|_\infty$ does not depend on the choice of $\tilde{x}$. Moreover, if there is a minimizer $\tilde{x}$ for which $\tilde{x}_\lambda \neq 0$, then $|(K^*(y-u))_\lambda| = \tau$. □

Lemma 7 *If, for some* $\lambda \in \Lambda$, *two minimizers* $\tilde{x}, \tilde{\tilde{x}}$ *satisfy* $\tilde{x}_\lambda \neq 0$ *and* $\tilde{\tilde{x}}_\lambda \neq 0$, *then* $\mathrm{sgn}\,\tilde{x}_\lambda = \mathrm{sgn}\,\tilde{\tilde{x}}_\lambda$.

Proof This follows from the arguments in the previous proof; $\tilde{x}_\lambda \neq 0$ implies $(K^*(y-u))_\lambda = \tau\,\mathrm{sgn}\,\tilde{x}_\lambda$. Similarly, $\tilde{\tilde{x}}_\lambda \neq 0$ implies $(K^*(y-u))_\lambda = \tau\,\mathrm{sgn}\,\tilde{\tilde{x}}_\lambda$, so that $\mathrm{sgn}\,\tilde{x}_\lambda = \mathrm{sgn}\,\tilde{\tilde{x}}_\lambda$. □

This immediately leads to

Lemma 8 *For all* $\tilde{x} \in B_R$ *that minimize* $\mathcal{D}$, *there are only finitely many* $\tilde{x}_\lambda \neq 0$. *More precisely,*

$$\{\lambda \in \Lambda : \tilde{x}_\lambda \neq 0\} \subset \Gamma := \{\lambda \in \Lambda : |(K^*(y-u))_\lambda| = \|(K^*(y-u))\|_\infty\}. \tag{5.11}$$

Moreover, if the vector e is defined by

$$e_\lambda = \begin{cases} 0, & \lambda \notin \Gamma, \\ \mathrm{sgn}((K^*(y-u))_\lambda), & \lambda \in \Gamma, \end{cases} \tag{5.12}$$

then $\langle \tilde{x}, e\rangle = R$ *for each minimizer* $\tilde{x}$ *of* $\mathcal{D}$ *in* B_R.

Proof We have already proved the set inclusion. Note that, since $K^*(y-u) \in \ell_2(\Lambda)$, the set Γ is necessarily a finite set. We also have, for each minimizer $\tilde{x}$,

$$\begin{aligned}\langle \tilde{x}, e\rangle &= \sum_{\lambda\in\Gamma} \tilde{x}_\lambda e_\lambda \\ &= \sum_{\lambda\in\Gamma,\tilde{x}_\lambda\neq 0} \tilde{x}_\lambda \mathrm{sgn}((K^*(y-u))_\lambda) \\ &= \sum_{\lambda\in\Gamma,\tilde{x}_\lambda\neq 0} \tilde{x}_\lambda \mathrm{sgn}(\tilde{x}_\lambda) = \|\tilde{x}\|_1 = R.\end{aligned}$$

□

Remark 1 By changing, if necessary, signs of the canonical basis vectors, we can assume, without loss of generality, that $e_\lambda = +1$ for all $\lambda \in \Gamma$. We shall do so from now on.

5.2 Weak Convergence to Minimizing Accumulation Points

We shall now impose some conditions on the $\beta^{(n)}$. We shall see examples in Sect. 6 where these conditions are verified.

Definition 1 We say that the sequence $(\beta^{(n)})_{n\in\mathbb{N}}$ *satisfies Condition (B) with respect to the sequence* $(x^{(n)})_{n\in\mathbb{N}}$ if there exists n_0 so that:

$$\text{(B1)} \quad \bar{\beta} := \sup\{\beta^{(n)}; n \in \mathbb{N}\} < \infty \quad \text{and} \quad \inf\{\beta^{(n)}; n \in \mathbb{N}\} \geq 1,$$

$$\text{(B2)} \quad \beta^{(n)} \|K(x^{(n+1)} - x^{(n)})\|^2 \leq r \|x^{(n+1)} - x^{(n)}\|^2 \quad \forall n \geq n_0.$$

We shall often abbreviate this by saying that 'the $\beta^{(n)}$ satisfy Condition (B)'. The constant r used in this definition is $r := \|K^*K\|_{\ell_2\to\ell_2} < 1$. (We can always assume, without loss of generality, that $\|K\|_{\ell_2\to\mathcal{H}} < 1$; if necessary, this can be achieved by a suitable rescaling of K and y.)

Note that the choice $\beta^{(n)} = 1$ for all n, which corresponds to the projected Landweber iteration, automatically satisfies Condition (B); since we shall show below that we obtain convergence when the $\beta^{(n)}$ satisfy Condition (B), this will then establish, as a corollary, convergence of the projected Landweber iteration algorithm (3.1) as well. We shall be interested in choosing, adaptively, larger values of $\beta^{(n)}$; in particular, we like to choose $\beta^{(n)}$ as large as possible.

Remark 2

- Condition (B) is inspired by the standard length-step in the steepest descent algorithm for the (unconstrained, unpenalized) functional $\|Kx - y\|^2$. In this case, one can speed up the standard Landweber iteration $x^{(n+1)} = x^{(n)} + K^*(y - Kx^{(n)})$ by defining instead $x^{(n+1)} = x^{(n)} + \alpha K^*(y - Kx^{(n)})$, where α is picked so that it gives the largest decrease of $\|Kx - y\|^2$ in this direction. This gives

$$\alpha = \left[\|K^*(y - Kx^{(n)})\|^2\right]\left[\|KK^*(y - Kx^{(n)})\|^2\right]^{-1}. \tag{5.13}$$

  In this linear case, one easily checks that α also equals

$$\alpha = \left[\|x^{(n+1)} - x^{(n)}\|^2\right]\left[\|K(x^{(n+1)} - x^{(n)})\|^2\right]^{-1}; \tag{5.14}$$

  in fact, it is this latter expression for α (which inspired the formulation of Condition (B)) that is most useful in proving convergence of the steepest descent algorithm.
- Because the definition of $x^{(n+1)}$ involves $\beta^{(n)}$, the inequality (B2), which uses $x^{(n+1)}$ to impose a limitation on $\beta^{(n)}$, has an "implicit" quality. In practice, it may not be straightforward to pick $\beta^{(n)}$ appropriately; one could conceive of trying first a "greedy" choice, such as e.g. $\frac{\|r^{(n)}\|^2}{\|Kr^{(n)}\|^2}$; if this value works, it is retained; if it doesn't, it can be gradually decreased (by multiplying it with a factor slightly smaller than 1) until (B2) is satisfied. (A similar way of testing appropriate step lengths is adopted in [32].)

In this section we prove that if the sequence $(x^{(n)})_{n\in\mathbb{N}}$ is defined iteratively by (3.2), and if the $\beta^{(n)}$ used in the iteration satisfy Condition (B) (with respect to the $x^{(n)}$), then the (weak) limit of any weakly convergent subsequence of $(x^{(n)})_{n\in\mathbb{N}}$ is necessarily a minimizer of $\mathcal{D}$ in B_R.

Lemma 9 *Assume* $\|K\|_{\ell_2\to\mathcal{H}} < 1$ *and* $\beta \ge 1$. *For arbitrary fixed* x *in* B_R, *define the functional* $F_\beta(\cdot; x)$ *by*

$$F_\beta(w; x) := \|Kw - y\|^2 - \|K(w - x)\|^2 + \frac{1}{\beta}\|w - x\|^2. \tag{5.15}$$

Then there is a unique choice for w *in* B_R *that minimizes the restriction to* B_R *of* $F_\beta(w; x)$. *We denote this minimizer by* $T_R(\beta; x)$; *it is given by* $T_R(\beta; x) = \mathbb{P}_R(x + \beta K^*(y - Kx))$.

Proof First of all, observe that the functional $F_\beta(\cdot, x)$ is strictly convex, so that it has a unique minimizer on B_R; let $\hat{x}$ be this minimizer. Then for all $w \in B_R$ and for all $t \in [0, 1]$

$$\begin{aligned}
&F_\beta(\hat{x}; x) \le F_\beta((1-t)\hat{x} + tw; x)\\
&\quad\Longrightarrow\quad 2t\left[\langle K\hat{x} - y, K(w - \hat{x})\rangle - \langle K\hat{x} - Kx, K(w - \hat{x})\rangle + \frac{1}{\beta}\langle \hat{x} - x, w - \hat{x}\rangle\right]\\
&\qquad\qquad + \frac{t^2}{\beta}\|w - \hat{x}\|^2 \ge 0\\
&\quad\Longrightarrow\quad \left[\beta\langle Kx - y, K(w - \hat{x})\rangle + \langle \hat{x} - x, w - \hat{x}\rangle\right] + \frac{t}{2}\|w - \hat{x}\|^2 \ge 0\\
&\quad\Longrightarrow\quad \langle \hat{x} - x + \beta K^*(Kx - y), w - \hat{x}\rangle \ge 0\\
&\quad\Longrightarrow\quad \langle x + \beta K^*(y - Kx) - \hat{x}, w - \hat{x}\rangle \le 0.
\end{aligned}$$

The latter implication is equivalent to $\hat{x} = \mathbb{P}_R(x + \beta K^*(y - Kx))$ by Lemma 3. □

An immediate consequence is

Lemma 10 *If the* $x^{(n)}$ *are defined by* (3.2), *and the* $\beta^{(n)}$ *satisfy* Condition (B) *with respect to the* $x^{(n)}$, *then the sequence* $(\mathcal{D}(x^{(n)}))_{n\in\mathbb{N}}$ *is decreasing, and*

$$\lim_{n\to\infty} \|x^{(n+1)} - x^{(n)}\| = 0. \tag{5.16}$$

Proof Comparing the definition of $x^{(n+1)}$ in (3.2) with the statement of Lemma 9, we see that $x^{(n+1)} = T_R(\beta^{(n)}; x^{(n)})$, so that $x^{(n+1)}$ is the minimizer, for $x \in B_R$, of $F_{\beta^{(n)}}(x; x^{(n)})$. Setting $\gamma = \frac{1}{r} - 1 > 0$, we have

$$\begin{aligned}
\mathcal{D}(x^{(n+1)}) &\le \mathcal{D}(x^{(n+1)}) + \gamma\|K(x^{(n+1)} - x^{(n)})\|^2\\
&= \|Kx^{(n+1)} - y\|^2 + (1+\gamma)\|K(x^{(n+1)} - x^{(n)})\|^2
\end{aligned}$$

$$
\begin{aligned}
&- \|K(x^{(n+1)} - x^{(n)})\|^2 \\
&\le \|Kx^{(n+1)} - y\|^2 - \|K(x^{(n+1)} - x^{(n)})\|^2 + \frac{1}{\beta^{(n)}}\|x^{(n+1)} - x^{(n)}\|^2 \\
&= F_{\beta^{(n)}}(x^{(n+1)}; x^{(n)}) \le F_{\beta^{(n)}}(x^{(n)}; x^{(n)}) = \mathcal{D}(x^{(n)}).
\end{aligned}
$$

We also have

$$
\begin{aligned}
&-F_{\beta^{(n+1)}}(x^{(n+1)}; x^{(n+1)}) + F_{\beta^{(n)}}(x^{(n+1)}; x^{(n)}) \\
&\quad = \frac{1}{\beta^{(n)}}\|x^{(n+1)} - x^{(n)}\|^2 - \|K(x^{(n+1)} - x^{(n)})\|^2 \\
&\quad \ge \frac{1-r}{\beta^{(n)}}\|x^{(n+1)} - x^{(n)}\|^2 \ge \frac{1-r}{\bar{\beta}}\|x^{(n+1)} - x^{(n)}\|^2.
\end{aligned}
$$

This implies

$$
\begin{aligned}
\sum_{n=0}^{N}\|x^{(n+1)} - x^{(n)}\|^2 &\le \frac{\bar{\beta}}{1-r}\sum_{n=0}^{N}\Big(F_{\beta^{(n)}}(x^{(n+1)}; x^{(n)}) - F_{\beta^{(n+1)}}(x^{(n+1)}; x^{(n+1)})\Big) \\
&\le \frac{\bar{\beta}}{1-r}\sum_{n=0}^{N}\Big(F_{\beta^{(n)}}(x^{(n)}; x^{(n)}) - F_{\beta^{(n+1)}}(x^{(n+1)}; x^{(n+1)})\Big) \\
&= \frac{\bar{\beta}}{1-r}\Big(F_{\beta^{(0)}}(x^{(0)}; x^{(0)}) - F_{\beta^{(N+1)}}(x^{(N+1)}; x^{(N+1)})\Big) \\
&\le \frac{\bar{\beta}}{1-r}F_{\beta^{(0)}}(x^{(0)}; x^{(0)}).
\end{aligned}
$$

Therefore, the series $\sum_{n=0}^{\infty}\|x^{(n+1)} - x^{(n)}\|^2$ converges and $\lim_{n\to\infty}\|x^{(n+1)} - x^{(n)}\| = 0$. □

Because the set $\{x^{(n)}; n \in \mathbb{N}\}$ is bounded in $\ell_1(\Lambda)$ ($x^{(n)}$ are all in B_R), it is bounded in $\ell_2(\Lambda)$ as well (since $\|a\|_2 \le \|a\|_1$). Because bounded closed sets in $\ell_2(\Lambda)$ are weakly compact, the sequence $(x^{(n)})_{n\in\mathbb{N}}$ must have weak accumulation points. We now have

Proposition 1 (Weak convergence to minimizing accumulation points) *If $x^{\#}$ is a weak accumulation point of $(x^{(n)})_{n\in\mathbb{N}}$ then $x^{\#}$ minimizes $\mathcal{D}$ in B_R.*

Proof Let $(x^{(n_j)})_{j\in\mathbb{N}}$ be a subsequence converging weakly to $x^{\#}$. Then for all $a \in \ell_2(\Lambda)$

$$
\langle Kx^{(n_j)}, a\rangle = \langle x^{(n_j)}, K^*a\rangle \underset{j\to\infty}{\longrightarrow} \langle x^{\#}, K^*a\rangle = \langle Kx^{\#}, a\rangle. \tag{5.17}
$$

Therefore w-$\lim_{j\to\infty} Kx^{(n_j)} = Kx^{\#}$. From Lemma 10 we have $\|x^{(n+1)} - x^{(n)}\| \underset{n\to\infty}{\longrightarrow} 0$, so that we also have w-$\lim_{j\to\infty} x^{(n_j+1)} = x^{\#}$. By the definition of $x^{(n+1)}$

$(x^{(n+1)} = \mathbb{P}_R(x^{(n)} + \beta^{(n)} K^*(y - Kx^{(n)})))$, and by Lemma 3, we have, for all $w \in B_R$,

$$\langle x^{(n)} + \beta^{(n)} K^*(y - Kx^{(n)}) - x^{(n+1)}, w - x^{(n+1)} \rangle \le 0. \tag{5.18}$$

In particular, specializing to our subsequence and taking the lim sup, we have

$$\limsup_{j\to\infty} \langle x^{(n_j)} - x^{(n_j+1)} + \beta^{(n_j)} K^*(y - Kx^{(n_j)}), w - x^{(n_j+1)} \rangle \le 0. \tag{5.19}$$

Because $\|x^{(n_j)} - x^{(n_j+1)}\| \to 0$, for $j \to \infty$, and $w - x^{(n_j+1)}$ is uniformly bounded, we have

$$\lim_{j\to\infty} |\langle x^{(n_j)} - x^{(n_j+1)}, w - x^{(n_j+1)} \rangle| = 0, \tag{5.20}$$

so that our inequality reduces to

$$\limsup_{j\to\infty} \beta^{(n_j)} \langle K^*(y - Kx^{(n_j)}), w - x^{(n_j+1)} \rangle \le 0. \tag{5.21}$$

By adding $\beta^{(n_j)} \langle K^*(y - Kx^{(n_j+1)}), x^{(n_j+1)} - x^{(n_j)} \rangle$, which also tends to zero as $j \to \infty$, we transform this into

$$\limsup_{j\to\infty} \beta^{(n_j)} \langle K^*(y - Kx^{(n_j)}), w - x^{(n_j)} \rangle \le 0. \tag{5.22}$$

Since the $\beta^{(n_j)}$ are all in $[1, \bar{\beta}]$, it follows that

$$\limsup_{j\to\infty} \langle K^*(y - Kx^{(n_j)}), w - x^{(n_j)} \rangle \le 0, \tag{5.23}$$

or

$$\limsup_{j\to\infty} \left[\langle K^* y, w - x^\# \rangle - \langle K^* K x^\#, w \rangle + \|Kx^{(n_j)}\|^2 \right] \le 0, \tag{5.24}$$

where we have used the weak convergence of $x^{(n_j)}$. This can be rewritten as

$$\langle K^*(y - Kx^\#), w - x^\# \rangle + \limsup_{j\to\infty} \left[\|Kx^{(n_j)}\|^2 - \|Kx^\#\|^2 \right] \le 0. \tag{5.25}$$

Since w-$\lim_{j\in\mathbb{N}} Kx^{(n_j)} = Kx^\#$, we have

$$\limsup_{j\to\infty} \left[\|Kx^{(n_j)}\|^2 - \|Kx^\#\|^2 \right] \ge 0.$$

We conclude thus that

$$\langle K^*(y - Kx^\#), w - x^\# \rangle \le 0, \quad \text{for all } w \in B_R, \tag{5.26}$$

so that $x^\#$ is a minimizer of $\mathcal{D}$ on B_R, by Lemma 4. □

5.3 Strong Convergence to Minimizing Accumulation Points

In this subsection we show how the weak convergence established in the preceding subsection can be strengthened into norm convergence, again by a series of lemmas. Since the distinction between weak and strong convergence makes sense only when the index set Λ is infinite, we shall implicitly assume this is the case throughout this section.

Lemma 11 *For the subsequence $(x^{(n_j)})_{j\in\mathbb{N}}$ defined in the proof of Proposition 1,*

$$\lim_{j\to\infty} K(x^{(n_j)}) = Kx^{\#}.$$

Proof Specializing the inequality (5.25) to $w = x^{\#}$, we obtain

$$\limsup_{j\to\infty}\left[\|Kx^{(n_j)}\|^2 - \|Kx^{\#}\|^2\right] \le 0;$$

together with $\|Kx^{\#}\|^2 \le \liminf_{j\to\infty}\|Kx^{(n_j)}\|^2$ (a consequence of the weak convergence of $Kx^{(n_j)}$ to $Kx^{\#}$), this implies $\lim_{j\to\infty}\|K(x^{(n_j)})\|^2 = \|Kx^{\#}\|^2$, and thus $\lim_{j\to\infty} K(x^{(n_j)}) = Kx^{\#}$. □

Lemma 12 *Under the same assumptions as in* Proposition 1, *there exists a subsequence $(x^{(n'_\ell)})_{\ell\in\mathbb{N}}$ of $(x^{(n)})_{n\in\mathbb{N}}$ such that*

$$\lim_{\ell\to\infty}\|x^{(n'_\ell)} - x^{\#}\| = 0. \tag{5.27}$$

Proof Let $(x^{(n_j)})_{j\in\mathbb{N}}$ be the subsequence defined in the proof of Proposition 1. Define now $u^{(j)} := x^{(n_j)} - x^{\#}$ and $v^{(j)} := x^{(n_j+1)} - x^{\#}$. Since, by Lemma 10, $\|x^{(n+1)} - x^{(n)}\|\xrightarrow[n\to\infty]{}0$, we have $\|u^{(j)} - v^{(j)}\|\xrightarrow[j\to\infty]{}0$. On the other hand,

$$\begin{aligned} u^{(j)} - v^{(j)} &= u^{(j)} + x^{\#} - \mathbb{P}_R\left(u^{(j)} + x^{\#} + \beta^{(n_j)}K^*(y - K(u^{(j)} + x^{\#}))\right)\\ &= u^{(j)} + \mathbb{P}_R\left(x^{\#} + \beta^{(n_j)}K^*(y - Kx^{\#})\right)\\ &\quad - \mathbb{P}_R\left(x^{\#} + \beta^{(n_j)}K^*(y - Kx^{\#}) + u^{(j)} - \beta^{(n_j)}K^*Ku^{(j)}\right), \end{aligned}$$

where we have used Proposition 1 ($x^{\#}$ is a minimizer) and Lemma 4 (so that $x^{\#} = \mathbb{P}_R(x^{\#} + \beta^{(n_j)}K^*(y - Kx^{\#}))$). By Lemma 11, $\lim_{j\to\infty}\|Ku^{(j)}\| = 0$. Since the $\beta^{(n_j)}$ are uniformly bounded, we have, by formula (4.3),

$$\begin{aligned} &\Big\|\mathbb{P}_R\left(x^{\#} + \beta^{(n_j)}K^*(y - Kx^{\#}) + u^{(j)} - \beta^{(n_j)}K^*Ku^{(j)}\right)\\ &\quad\times \mathbb{P}_R\left(x^{\#} + \beta^{(n_j)}K^*(y - Kx^{\#}) + u^{(j)}\right)\Big\|\\ &\le \beta^{(n_j)}\|K^*Ku^{(j)}\|\xrightarrow[j\to\infty]{}0. \end{aligned}$$

Combining this with $\|u^{(j)} - v^{(j)}\| \xrightarrow[j\to\infty]{} 0$, we obtain

$$\lim_{j\to\infty} \Big\| \mathbb{P}_R \left(x^\# + \beta^{(n_j)} K^*(y - Kx^\#) + u^{(j)} \right) \\ - \mathbb{P}_R \left(x^\# + \beta^{(n_j)} K^*(y - Kx^\#) \right) - u^{(j)} \Big\| = 0. \tag{5.28}$$

Since the $\beta^{(n_j)}$ are uniformly bounded, they must have at least one accumulation point. Let $\beta^{(\infty)}$ be such an accumulation point, and choose a subsequence $(j_\ell)_{\ell\in\mathbb{N}}$ such that $\lim_{\ell\to\infty} \beta^{(n_{j_\ell})} = \beta^{(\infty)}$. To simplify notation, we write $n'_\ell := n_{j_\ell}$, $u'^{(\ell)} := u^{(j_\ell)}$, $v'^{(\ell)} := v^{(j_\ell)}$. We have thus

$$\lim_{\ell\to\infty} \beta^{(n'_\ell)} = \beta^{(\infty)}, \quad \text{and} \\ \lim_{\ell\to\infty} \Big\| \mathbb{P}_R \left(x^\# + \beta^{(n'_\ell)} K^*(y - Kx^\#) + u'^{(\ell)} \right) \\ - \mathbb{P}_R \left(x^\# + \beta^{(n'_\ell)} K^*(y - Kx^\#) \right) - u'^{(\ell)} \Big\| = 0. \tag{5.29}$$

Denote $h^\# := x^\# + \beta^{(\infty)} K^*(y - Kx^\#)$ and $h'^{(\ell)} := x^\# + \beta^{(n'_\ell)} K^*(y - Kx^\#)$. We have now

$$\begin{aligned} &\|\mathbb{P}_R(h^\# + u'^{(\ell)}) - \mathbb{P}_R(h^\#) - u'^{(\ell)}\| \\ &\quad\le \|\mathbb{P}_R(h'^{(\ell)} + u'^{(\ell)}) - \mathbb{P}_R(h'^{(\ell)}) - u'^{(\ell)}\| \\ &\qquad + \|\mathbb{P}_R(h'^{(\ell)} + u'^{(\ell)}) - \mathbb{P}_R(h^\# + u'^{(\ell)})\| + \|\mathbb{P}_R(h'^{(\ell)}) - \mathbb{P}_R(h^\#)\| \\ &\quad\le \|\mathbb{P}_R(h'^{(\ell)} + u'^{(\ell)}) - \mathbb{P}_R(h'^{(\ell)}) - u'^{(\ell)}\| + 2\|h'^{(\ell)} - h^\#\|. \end{aligned}$$

Since both terms on the right hand side converge to zero for $\ell \to \infty$ (see (5.29)), we have

$$\lim_{\ell\to\infty} \|\mathbb{P}_R(h^\# + u'^{(\ell)}) - \mathbb{P}_R(h^\#) - u'^{(\ell)}\| = 0. \tag{5.30}$$

Without loss of generality we can assume $\|h^\#\|_1 > R$. By Lemma 2 there exists $\mu > 0$ such that $\mathbb{P}_R(h^\#) = \mathbb{S}_\mu(h^\#)$. Because $|h^\#_\lambda| \to 0$ as $|\lambda| \to \infty$, this implies that, for some finite $K_1 > 0$, $(\mathbb{P}_R(h^\#))_\lambda = 0$ for $|\lambda| > K_1$. Pick now any $\epsilon > 0$ that satisfies $\epsilon < \mu/5$. There exists a finite $K_2 > 0$ so that $\sum_{|\lambda|>K_2} |h^\#_\lambda|^2 < \epsilon^2$. Set $K_0 := \max(K_1, K_2)$, and define the vector $\tilde{h}^\#$ by $\tilde{h}^\#_\lambda = h^\#_\lambda$ if $|\lambda| \le K_0$, $\tilde{h}^\#_\lambda = 0$ if $|\lambda| > K_0$.

By the weak convergence of the $u'^{(\ell)}$, we can, for this same K_0, determine $L_1 > 0$ such that, for all $\ell \ge L_1$, $\sum_{|\lambda|\le K_0} |u'^{(\ell)}_\lambda|^2 \le \epsilon^2$. Define new vectors $\tilde{u}'^{(\ell)}$ by $\tilde{u}'^{(\ell)}_\lambda = 0$ if $|\lambda| \le K_0$, $\tilde{u}'^{(\ell)}_\lambda = u'^{(\ell)}_\lambda$ if $|\lambda| > K_0$.

Because of (5.30), there exists $L_2 > 0$ such that $\|\mathbb{P}_R(h^\# + u'^{(\ell)}) - \mathbb{P}_R(h^\#) - u'^{(\ell)}\| \le \epsilon$ for $\ell \ge L_2$. Consider now $\ell \ge L := \max(L_1, L_2)$. We have

$$\|\mathbb{P}_R(\tilde{h}^\# + \tilde{u}'^{(\ell)}) - \mathbb{P}_R(\tilde{h}^\#) - \tilde{u}'^{(\ell)}\|$$

$$\begin{aligned}&\le \|\mathbb{P}_R(\tilde h^\# + \tilde u'^{(\ell)}) - \mathbb{P}_R(h^\# + \tilde u'^{(\ell)})\| + \|\mathbb{P}_R(h^\# + \tilde u'^{(\ell)}) - \mathbb{P}_R(h^\# + u'^{(\ell)})\|\\&\quad + \|\mathbb{P}_R(h^\# + u'^{(\ell)}) - \mathbb{P}_R(h^\#) - u'^{(\ell)}\| + \|\mathbb{P}_R(h^\#) - \mathbb{P}_R(\tilde h^\#)\| + \|u'^{(\ell)} - \tilde u'^{(\ell)}\|\\&\le 5\epsilon.\end{aligned}$$

On the other hand, Lemma 2 tells us that there exists $\sigma_\ell > 0$ such that $\mathbb{P}_R(\tilde h^\# + \tilde u'^{(\ell)}) = \mathbb{S}_{\sigma_\ell}(\tilde h^\# + \tilde u'^{(\ell)}) = \mathbb{S}_{\sigma_\ell}(\tilde h^\#) + \mathbb{S}_{\sigma_\ell}(\tilde u'^{(\ell)})$, where we used in the last equality that $\tilde h^\#_\lambda = 0$ for $|\lambda| > K_0$ and $\tilde u'^{(\ell)}_\lambda = 0$ for $|\lambda| \le K_0$. From $\|\mathbb{S}_\mu(\tilde h^\#)\|_1 = R = \|\mathbb{S}_{\sigma_\ell}(\tilde h^\#)\|_1 + \|\mathbb{S}_{\sigma_\ell}(\tilde u'^{(\ell)})\|_1$ we conclude that $\sigma_\ell \ge \mu$ for all $\ell \ge L$. We then deduce

$$\begin{aligned}(5\epsilon)^2 &\ge \|\mathbb{P}_R(\tilde h^\# + \tilde u'^{(\ell)}) - \mathbb{P}_R(\tilde h^\#) - \tilde u'^{(\ell)}\|^2\\&= \sum_{|\lambda|\le K_0} |S_{\sigma_\ell}(\tilde h^\#_\lambda) - S_\mu(\tilde h^\#_\lambda)|^2 + \sum_{|\lambda|>K_0} |S_{\sigma_\ell}(\tilde u'^{(\ell)}_\lambda) - \tilde u'^{(\ell)}_\lambda|^2\\&\ge \sum_{|\lambda|>K_0} \Big[\max\big(|\tilde u'^{(\ell)}_\lambda| - \sigma_\ell, 0\big) - |\tilde u'^{(\ell)}_\lambda|\Big]^2\\&= \sum_{|\lambda|>K_0} \min\big(|\tilde u'^{(\ell)}_\lambda|, \sigma_\ell\big)^2 \ge \sum_{|\lambda|>K_0} \min\big(|\tilde u'^{(\ell)}_\lambda|, \mu\big)^2.\end{aligned}$$

Because we picked $\epsilon < \mu/5$, this is possible only if $|\tilde u'^{(\ell)}_\lambda| \le \mu$ for all $|\lambda| > K_0$, $\ell \ge L$, and if, in addition,

$$\Bigg[\sum_{|\lambda|>K_0} |\tilde u'^{(\ell)}_\lambda|^2\Bigg]^{1/2} \le 5\epsilon, \quad \text{i.e., } \|\tilde u'^{(\ell)}\| \le 5\epsilon. \tag{5.31}$$

It then follows that $\|u'^{(\ell)}\| \le \|\tilde u'^{(\ell)}\| + [\sum_{|\lambda|\le K_0} |u'^{(\ell)}_\lambda|^2]^{1/2} \le 6\epsilon$.

We have thus obtained what we set out to prove: the subsequence $(x^{n_{j_\ell}})_{\ell\in\mathbb{N}}$ of $(x^{(n)})_{n\in\mathbb{N}}$ satisfies that, given arbitrary $\epsilon > 0$, there exists L so that, for $\ell > L$, $\|x^{n_{j_\ell}} - x^\#\| \le 6\epsilon$. □

Remark 3 In this proof we have implicitly assumed that $\|h^\# + u^{(j)}\|_1 > R$. Given that $\|h^\#\|_1 > R$, this assumption can be made without loss of generality, because it is not possible to have $\|h^\#\|_1 > R$ and $\|h^\# + u^{(j)}\|_1 < R$ infinitely often, as the following argument shows. Find K_0, L_0 such that $\sum_{|\lambda|<K_0} |h^\#_\lambda| \ge (\|h^\#\|_1 + R)/2$ and, $\forall \ell \ge L_0$ and $\forall |\lambda| < K_0$: $|u'^{(\ell)}_\lambda| < (K_0^{-1}(\|h^\#\|_1 - R)/4$. Then $\sum_{|\lambda|<K_0} |h^\#_\lambda + u'^{(\ell)}_\lambda| \ge \sum_{|\lambda|<K_0} |h^\#_\lambda| - |u'^{(\ell)}_\lambda| \ge (\|h^\#\|_1 + R)/2 - (\|h^\#\|_1 - R)/4 = R + (\|h^\#\|_1 - R)/4 > R$. Hence, $\forall \ell > L_0$, $\|h^\# + u'^{(\ell)}\|_1 \ge R$.

Remark 4 At the cost of more technicalities it is possible to show that the whole subsequence $(x^{(n_j)})_{j\in\mathbb{N}}$ defined in the proof of Proposition 1 converges in norm to $x^\#$, i.e., that $\lim_{j\to\infty} \|x^{(n_j)} - x^\#\| = 0$, without going to a subsequence $(x^{n_{j_\ell}})_{\ell\in\mathbb{N}}$.

The following proposition summarizes in one statement all the findings of the last two subsections.

Proposition 2 (Norm convergence to minimizing accumulation points) *Every weak accumulation point $x^\#$ of the sequence $(x^{(n)})_{n\in\mathbb{N}}$ defined by (3.2) is a minimizer of $\mathcal{D}$ in B_R. Moreover, there exists a subsequence $(x^{(n_\ell)})_{\ell\in\mathbb{N}}$ of $(x^{(n)})_{n\in\mathbb{N}}$ that converges to $x^\#$ in norm.*

5.4 Uniqueness of the Accumulation Point

In this subsection we prove that the accumulation point $x^\#$ of $(x^{(n)})_{n\in\mathbb{N}}$ is unique, so that the entire sequence $(x^{(n)})_{n\in\mathbb{N}}$ converges to $x^\#$ in norm. (Note that two sequences $(x^{(n)})_{n\in\mathbb{N}}$ and $(x'^{(n)})_{n\in\mathbb{N}}$, both defined by the same recursion, but starting from different initial points $x^{(0)} \neq x'^{(0)}$, can still converge to different limits $x^\#$ and $x'^\#$.)

We start again from the inequality

$$\langle x^{(n)} + \beta^{(n)} K^*(y - Kx^{(n)}) - x^{(n+1)}, w - x^{(n+1)}\rangle \leq 0, \tag{5.32}$$

for all $w \in B_R$ and for all $n \in \mathbb{N}$, and its many consequences. Define M_R to be the set of minimizers of $\mathcal{D}$ on B_R. By Lemma 5, $M_R = B_R \cap (\tilde{x} + \ker K)$, where $\tilde{x}$ is an arbitrary minimizer of $\mathcal{D}$ in B_R. By the convention adopted in Remark 1,

$$M_R \subset B_R^+ := \left\{x \in \ell_1(\Lambda);\, x_\lambda \geq 0 \text{ for all } \lambda \in \Lambda, \text{ and } \sum_{\lambda\in\Lambda} x_\lambda \leq R\right\}. \tag{5.33}$$

Moreover, for each element $z \in M_R$, $z_\lambda = 0$ if $\lambda \notin \Gamma$ (see Lemma 8). The set M_R is both closed and convex. We define the corresponding (nonlinear) projection operator $\mathbb{P}_{M_R}$ as usual,

$$\mathbb{P}_{M_R}(v) := \arg\min\{\|v - z\|^2;\, z \in M_R\}. \tag{5.34}$$

Because M_R is convex, this projection operator has the following property:

$$\forall \tilde{x} \in M_R : \langle z - \mathbb{P}_{M_R}(z), \tilde{x} - \mathbb{P}_{M_R}(z)\rangle \leq 0. \tag{5.35}$$

(The proof is standard, and is essentially given in the proof Lemma 3, where in fact only the convexity of B_R was used.) For each $n \in \mathbb{N}$, we introduce now $a^{(n)}$ and $b^{(n)}$ defined by

$$a^{(n)} := \mathbb{P}_{M_R}(x^{(n)}), \qquad b^{(n)} = x^{(n)} - a^{(n)}. \tag{5.36}$$

Specializing equation (5.35) to $x^{(n)}$, we obtain, for all $\tilde{x} \in M_R$ and for all $n \in \mathbb{N}$:

$$\langle x^{(n)} - a^{(n)}, \tilde{x} - a^{(n)}\rangle \leq 0 \tag{5.37}$$

or

$$\langle b^{(n)}, \tilde{x} - a^{(n)}\rangle \leq 0. \tag{5.38}$$

Because $a^{(n)}$ is a minimizer, we can also apply Lemma 4 to $a^{(n)}$ and conclude

$$\langle K^*(y - Ka^{(n)}), w - a^{(n)}\rangle \leq 0, \quad \text{for all } w \in B_R. \tag{5.39}$$

With these inequalities, we can prove the following crucial result.

Lemma 13 *For any $\tilde{x} \in M_R$, and for any $n \in \mathbb{N}$,*

$$\|x^{(n+1)} - \tilde{x}\| \le \|x^{(n)} - \tilde{x}\|. \tag{5.40}$$

Proof We set $w = \tilde{x}$ in (5.32), leading to

$$\langle x^{(n)} - x^{(n+1)}, \tilde{x} - x^{(n+1)} \rangle + \beta^{(n)} \langle K^*(y - Kx^{(n)}), -b^{(n+1)} \rangle \le 0, \tag{5.41}$$

where we have used that $K\tilde{x} = Ka^{(n+1)}$. We also have, setting $w = x^{(n+1)}$ in the $(n+1)$-version of (5.39),

$$\langle K^*(y - Ka^{(n+1)}), x^{(n+1)} - a^{(n+1)} \rangle \le 0, \tag{5.42}$$

or

$$\langle K^*(y - Ka^{(n)}), b^{(n+1)} \rangle \le 0, \tag{5.43}$$

where we have used $Ka^{(n)} = Ka^{(n+1)}$. It follows that

$$\langle x^{(n)} - x^{(n+1)}, \tilde{x} - x^{(n+1)} \rangle + \beta^{(n)} \langle -K^*Kb^{(n)}, -b^{(n+1)} \rangle \le 0, \tag{5.44}$$

or

$$\langle x^{(n)} - x^{(n+1)}, \tilde{x} - x^{(n+1)} \rangle + \beta^{(n)} \langle Kb^{(n)}, Kb^{(n+1)} \rangle \le 0, \tag{5.45}$$

which is also equivalent to

$$\begin{aligned} &\langle x^{(n)} - \tilde{x}, \tilde{x} - x^{(n+1)} \rangle + \|\tilde{x} - x^{(n+1)}\|^2 + \frac{1}{2}\beta^{(n)} \left[\|Kb^{(n)}\|^2 + \|Kb^{(n+1)}\|^2 \right] \\ &\quad - \frac{1}{2}\beta^{(n)} \|Kb^{(n)} - Kb^{(n+1)}\|^2 \le 0. \end{aligned} \tag{5.46}$$

Adding $\frac{1}{2}\beta^{(n)}\|K(b^{(n)} - b^{(n+1)})\|^2 \le \frac{r}{2}\|x^{(n)} - x^{(n+1)}\|^2$ to (5.46), we have

$$\begin{aligned} &\langle x^{(n)} - \tilde{x}, \tilde{x} - x^{(n+1)} \rangle + \|\tilde{x} - x^{(n+1)}\|^2 + \frac{1}{2}\beta^{(n)} \left[\|Kb^{(n)}\|^2 + \|Kb^{(n+1)}\|^2 \right] \\ &\quad \le \frac{r}{2}\|x^{(n)} - x^{(n+1)}\|^2 \\ &\quad = \frac{r}{2}\left[\|x^{(n)} - \tilde{x}\|^2 + \|x^{(n+1)} - \tilde{x}\|^2 - 2\langle x^{(n)} - \tilde{x}, x^{(n+1)} - \tilde{x} \rangle \right]. \end{aligned}$$

It follows that

$$\begin{aligned} &\left(1 - \frac{r}{2}\right) \|x^{(n+1)} - \tilde{x}\|^2 + (1 - r)\langle \tilde{x} - x^{(n)}, x^{(n+1)} - \tilde{x} \rangle - \frac{r}{2}\|\tilde{x} - x^{(n)}\|^2 \\ &\quad \le -\frac{1}{2}\beta^{(n)} \left[\|Kb^{(n)}\|^2 + \|Kb^{(n+1)}\|^2 \right] \le 0, \end{aligned} \tag{5.47}$$

which, in turn, implies that

$$\left(1 - \frac{r}{2}\right) \|x^{(n+1)} - \tilde{x}\|^2 - (1 - r)\|\tilde{x} - x^{(n)}\| \|x^{(n+1)} - \tilde{x}\| - \frac{r}{2}\|\tilde{x} - x^{(n)}\|^2 \le 0. \tag{5.48}$$

This can be rewritten as

$$\left[\|\tilde{x}-x^{(n+1)}\|-\|\tilde{x}-x^{(n)}\|\right]\left[\left(1-\frac{r}{2}\right)\|x^{(n+1)}-\tilde{x}\|+\frac{r}{2}\|\tilde{x}-x^{(n)}\|\right]\leq 0, \quad (5.49)$$

which implies $\|x^{(n+1)}-\tilde{x}\|\leq\|x^{(n)}-\tilde{x}\|$. □

We are now ready to state the main result of our work.

Theorem 1 *The sequence* $(x^{(n)})_{n\in\mathbb{N}}$ *as defined in* (3.2), *where the* step-length sequence $(\beta^{(n)})_{n\in\mathbb{N}}$ *satisfies Condition* (*B*) *with respect to the* $x^{(n)}$, *converges in norm to a minimizer of* $\mathcal{D}$ *on* B_R.

Proof The sequence $(x^{(n)})_{n\in\mathbb{N}}$ has a least one accumulation point $x^{\#}$. By Proposition 1 $x^{\#}$ minimizes $\mathcal{D}$ in B_R. By Proposition 2 $(x^{(n)})_{n\in\mathbb{N}}$ has a subsequence $(x^{(n_\ell)})_{\ell\in\mathbb{N}}$ that converges to $x^{\#}$. By Lemma 13 $\|x^{(n)}-x^{\#}\|$ decreases monotonically, hence it has a limit for $n\to\infty$, and

$$\lim_{n\to\infty}\|x^{(n)}-x^{\#}\|=\lim_{\ell\to\infty}\|x^{(n_\ell)}-x^{\#}\|=0. \quad (5.50)$$

□

6 Numerical Experiments and Additional Algorithms

6.1 Numerical Examples

We conduct a number of numerical experiments to gauge the effectiveness of the different algorithms we discussed. All computations were done in Mathematica 5.2 [46] on a 2 GHz workstation with 2 Gb memory.

We are primarily interested in the behavior, as a function of time (not number of iterations), of the relative error $\|x^{(n)}-\bar{x}\|/\|\bar{x}\|$. To this end, and for a given operator K and data y, we need to know in advance the actual minimizer $\bar{x}(\tau)$ of the functional (2.3).

One can calculate the minimizer exactly (in practice up to computer round-off) with a finite number of steps using the LARS algorithm described in [29] (the variant called 'Lasso', implemented independently by us). This algorithm scales badly, and is useful in practice only when the number of non-zero entries in the minimizer $\bar{x}(\tau)$ is sufficiently small. We made our own implementation of this algorithm to make it more directly applicable to our problem (i.e., we do not renormalize the columns of the matrix to have zero mean and unit variance, as it is done in the statistics context [29]). We also double-check the minimizer obtained in this manner by verifying that it is indeed a fixed point of the iterative thresholding algorithm (2.5) (up to machine epsilon). We then have an 'exact' minimizer $\bar{x}$ together with its radius $R=\|\bar{x}\|_1$ (used in the projected algorithms) and, according to Lemma 6, the corresponding threshold $\tau=\max_i|\bar{r}_i|$ with $\bar{r}=K^*(y-K\bar{x})$ (used in the iterative thresholding algorithm).

The numerical examples below are listed in order of increasing complexity; they illustrate that the algorithms can behave differently for different examples. In these

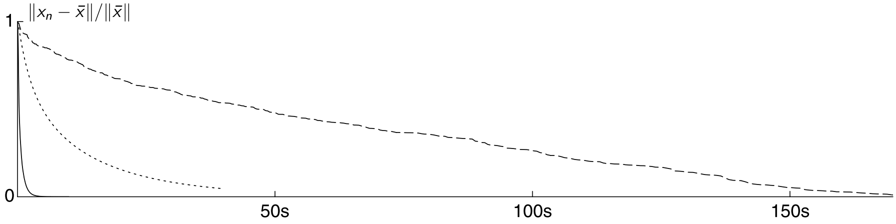

Fig. 3 The different convergence rates of the thresholded Landweber algorithm (*dotted line*), the projected steepest descent algorithm (*solid line*, near vertical axis) and the LARS algorithm (*dashed line*), for the second example. The projected steepest descent algorithm converges much faster than the thresholded Landweber iteration. They both do better than the LARS method

experiments we choose $\beta^{(n)} = \beta_{\text{st.}}^{(n)} := \|r^{(n)}\|^2/\|Kr^{(n)}\|^2$, (where, as before, $r^{(n)} = K^*(y - Kx^{(n)})$); $\beta_{\text{st.}}^{(n)}$ is the standard descent parameter from the classical linear steepest descent algorithm.

1. When K is a partial Fourier matrix (i.e., a Fourier matrix with a prescribed number of deleted rows), there is no advantage in using a dynamical step size $\beta_{\text{st.}}^{(n)} = \|r^{(n)}\|^2/\|Kr^{(n)}\|^2$ as this ratio is always equal to 1. This trivially fulfills Condition (B) in Sect. 5.1. The performance of the projected steepest descent iteration simply equals that of the projected Landweber iterations.
2. By combining a scaled partial Fourier transform with a rank 1 projection operator, we constructed our second example, in which K is a 1536×2049 matrix, of rank 1536, with largest singular value equal to 0.99 and all the other singular values between 0.01 and 0.11. Because of the construction of the matrix, the FFT algorithm provides a fast way of computing the action of this matrix on a vector. For the y and τ that were chosen, the limit vector $\bar{x}_\tau$ has 429 nonzero entries. For this example, the LARS procedure is slower than thresholded Landweber, which in turn is significantly slower than projected steepest descent. To get within a distance of the true minimizer corresponding to a 5% relative error, the projected steepest descent algorithm takes 2 s, the thresholded Landweber algorithm 39 s, and LARS 151 s. (The relatively poor performance of LARS in this case is due to the large number of nonzero entries in the limit vector $\bar{x}_\tau$; the complexity of LARS is cubic in this number of nonzero entries.) In this case, the $\beta_{\text{st.}}^{(n)} = \|r^{(n)}\|^2/\|Kr^{(n)}\|^2$ are much larger than 1; moreover, they satisfy Condition (B) of Sect. 5.1 at every step. We illustrate the results in Fig. 3.
3. The last example is inspired by a real-life application in geoscience [38], in particular an application in seismic tomography based on earthquake data. The object space consists of the wavelet coefficients of a 2D seismic velocity perturbation. There are 8192 degrees of freedom. In this particular case the number of data is 1848. Hence the matrix K has 1848 rows and 8192 columns. We apply the different methods to the same noisy data that are used in [38] and measure the time to convergence up to a specified relative error (see Table 1 and Fig. 4). This example illustrates the slow convergence of the thresholded Landweber algorithm (2.5), and the improvements made by a projected steepest descent iteration (3.2) with the special choice $\beta^{(n)} = \beta_{\text{st.}}^{(n)}$ above. In this case, this choice turns out *not* to satisfy

Table 1 Table illustrating the relative performance of three algorithms: thresholded Landweber, projected Landweber and projected steepest descent, for the third example

Relative	Thresholded Landweber		Projected st. descent		Projected Landweber	
error	n	Time	n	Time	n	Time
0.90	3	1 s	2	1 s	3	2 s
0.80	20	8 s	8	7 s	15	11 s
0.70	163	1 m 8 s	20	17 s	59	44 s
0.50	3216	22 m 9 s	340	4 m 56 s	2124	27 m 17 s
0.20	55473	6 h 23 m	6738	1 h 37 m		
0.10	100620	11 h 38 m	11830	2 h 51 m		
0.03	198357	21 h 47 m	22037	5 h 20 m		

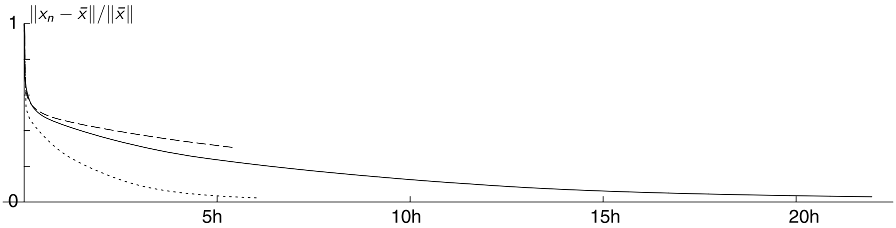

Fig. 4 The different convergence rates of the thresholded Landweber algorithm (*solid line*), the projected Landweber algorithm (*dashed line*) and the projected steepest descent algorithm (*dotted line*), for the third example. The projected steepest descent algorithm converges about four times faster than the thresholded Landweber iteration. The projected Landweber iteration does better at first (not visible in this plot), but looses with respect to iterative thresholding afterwards. The horizontal axis has time (in hours), the vertical axis displays the relative error

Condition (B) in general. One could conceivably use successive corrections, e.g. by a line-search, to determine, starting from $\beta_{\text{st.}}^{(n)}$, values of $\beta^{(n)}$ that would satisfy condition (B), and thus guarantee convergence as established by Theorem 5.18. This would slow down the method considerably. The $\beta_{\text{st.}}^{(n)}$ seem to be in the right ballpark, and provide us with a numerically converging sequence. We also implemented the projected Landweber algorithm (3.1); it is listed in Table 1 and illustrated in Fig. 4.

The matrix K in this example is extremely ill-conditioned: its largest singular value was normalized to 1, but the remaining singular values quickly tend to zero. The threshold was chosen, according to the (known or estimated) noise level in the data, so that $\mathcal{D}(\bar{x})/\sigma^2 = 1848$ (= the number of data points), where σ is the data noise level; this is a standard choice that avoids overfitting.

In Fig. 4, we see that the thresholded Landweber algorithm takes more than 21 hours (corresponding to 200,000 iterations) to converge to the true minimizer within a 3% relative error, as measured by the usual ℓ_2 distance. The projected steepest descent algorithm is about four times faster and reaches the same reconstruction error in about 5.5 hours (25,000 iterations). Due to one additional matrix-vector multiplication and, to a minor extent, the computation of the projection onto an ℓ_1-ball, one step in the projected steepest descent algorithm takes approximately twice as long as one step in the thresholded Landweber algorithm. For the projected Landweber algorithm there is an advantage in the first few iterations, but after a short while, the additional time needed to compute the projection $\mathbb{P}_R$ (i.e., to compute the corresponding variable thresholds) makes this algorithm slower than the iterative soft-thresholding. We illustrate the corresponding CPU time in Table 1.

It is worthwhile noticing that for the three algorithms the value of the functional (2.3) converges much faster to its limit value than the minimizer itself: When the reconstruction error is 10%, the corresponding value of the functional is already accurate up to three digits with respect to the value of the functional at $\bar{x}$. We can imagine that in this case the functional has a long narrow "valley" with a very gentle slope in the direction of the eigenvectors with small (or zero) singular values. The path in the $\|x\|_1$ vs. $\|Kx-y\|^2$ plane followed by the iterates is shown in Fig. 1. The projected steepest descent algorithm, by construction, stays within a fixed ℓ_1-ball, and, as already mentioned, converges faster than the thresholded Landweber algorithm. The path followed by the LARS algorithm is also pictured. It corresponds with the so-called *trade-off* curve which can be interpreted as the border of the area that is reachable by any element of the model space, i.e., it is generated by $\bar{x}(\tau)$ for decreasing values of $\tau > 0$.

In this particular example, the number of nonzero components of $\bar{x}$ equals 128. The LARS (exact) algorithm only takes 55 seconds, which is *much* faster than any of the iterative methods demonstrated here. However, as illustrated above, by the second example, LARS looses its advantage when dealing with larger problems where the minimizer is not sparse in absolute number of entries, as is the case in, e.g., realistic problems of global seismic tomography. Indeed, the example presented here is a "toy model" for proof-of-concept for geoscience applications. The 3D model will involve millions of unknowns and solutions that may be sparse compared with the total number of unknowns, but not sparse in absolute numbers. Because the complexity of LARS is cubic in the number of nonzero components of the solution, such 3D problems are expected to lie beyond its useful range.

6.2 Relationship to Other Methods

The projected iterations (5.2) and (5.3) are related to the POCS (Projection on Convex Sets) technique [3]. The projection of a vector a on the solution space $\{x : Kx = y\}$ (a convex set, assumed here to be non-empty; no such assumption was made before because the functional (2.3) always has a minimum) is given by:

$$x = a - K^*(KK^*)^{-1}(y - Ka). \tag{6.1}$$

Hence, alternating projections on the convex sets $\{x : Kx = y\}$ and B_R give rise to the algorithm [5]: Pick an arbitrary $x^{(0)} \in \ell_2(\Lambda)$, for example $x^{(0)} = 0$, and iterate

$$x^{(n+1)} = \mathbb{P}_R(x^{(n)} - K^*(KK^*)^{-1}(y - Kx^{(n)})). \tag{6.2}$$

This may be practical in case of a small number of data or when there is structure in K, i.e., when KK^* is efficiently inverted. Approximating KK^* by the unit matrix, yields the projected Landweber algorithm (5.2); approximating $(KK^*)^{-1}$ by a

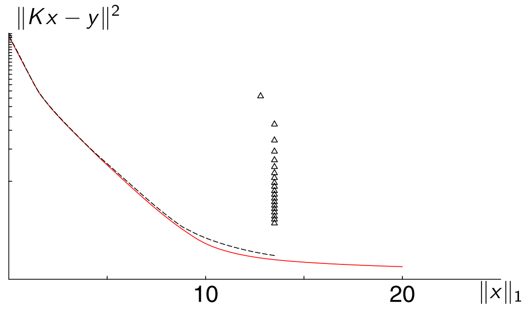

Fig. 5 Trade-off curve (*solid line*) and its approximation with algorithm (6.3) in 200 steps (*dashed line*). For comparison, the iterates of projected steepest descent are also indicated (*triangles*)

constant multiple of the unit matrix yields the projected gradient iteration (5.3) if one chooses the constant equal to $\beta^{(n)}$. The projected methods discussed in this paper produce iterates that (except for the first few) live on the 'skin' of the ℓ_1-ball of radius R, as shown in Fig. 1. We have found even more promising results for an 'interior' algorithm in which we still project on ℓ_1-balls, but now with a slowly increasing radius, i.e.,

$$x^{(n+1)} = \mathbb{P}_{R^{(n)}}\big(x^{(n)} + \beta^{(n)} r^{(n)}\big), \quad R^{(n)} = (n+1)R/N, \text{ and } n = 0, \dots, N, \quad (6.3)$$

where N is the prescribed maximum number of iterations (the origin is chosen as the starting point of this iteration). We do not have a proof of convergence of this 'interior point type' algorithm. We observed (also without proof) that the path traced by the iterates $x^{(n)}$ (in the $\|x\|_1$ vs. $\|Kx - y\|^2$ plane) is very close to the trade-off curve (see Fig. 5); this is a useful property in practice since at least part of the trade-off curve should be constructed anyway.

Note that the strategy followed by these algorithms is similar to that of LARS [29], in that they both start with $x^{(0)} = 0$ and slowly increase the ℓ_1 norm of the successive approximations.

While we were finishing this paper, Michael Friedlander informed us of their numerical results in [45] which are closely related to our approach, although their analysis is limited to finite dimensions.

Different, but closely related is also the recent approach by Figueiredo, Nowak, and Wright [32]. The authors first reformulate the minimization of (2.3) as a bound-constrained quadratic program in standard form, and then they apply iterative projected gradient iterations, where the projection act componentwise by clipping to zero negative components.

7 Conclusions

We have presented convergence results for accelerated projected gradient methods to find a minimizer of an ℓ_1 penalized functional. The innovation due to the introduction of 'Condition (B)' is to guarantee strong convergence for the full sequence. Numerical examples confirm that this algorithm can outperform (in terms of CPU time) existing methods such as the thresholded Landweber iteration or even LARS.

It is important to remark that the speed of convergence may depend strongly on how the operator is available. Because most of the time in the iterations is consumed by matrix-vector multiplications (as is often the case for iterative algorithms), it makes a big difference whether K is given by a full matrix or a sparse matrix (perhaps sparse in the sense that its action on a vector can be computed via a fast algorithm, such as the FFT or a wavelet transform). The applicability of the projected algorithms hinges on the observation that the ℓ_2 projection on an ℓ_1 ball can be computed with a $\mathcal{O}(m \log m)$-algorithm, where m is the dimension of the underlying space.

There is no universal method that performs best for any choice of the operator, data, and penalization parameter. As a general rule of thumb we expect that, among the algorithms discussed in this paper for which we have convergence proofs,

- the thresholded Landweber algorithm (2.5) works best for an operator K close to the identity (independently of the sparsity of the limit),
- the projected steepest descent algorithm (3.2) works best for an operator with a relatively nice spectrum, i.e., with not too many zeroes (also independently of the sparsity of the minimizer), and
- the exact (LARS) method works best when the minimizer is sparse in absolute terms.

Obviously, the three cases overlap partially, and they do not cover the whole range of possible operators and data. In future work we intend to investigate algorithms that would further improve the performance for the case of a large ill-conditioned matrix and a minimizer that is relatively sparse with respect to the dimension of the underlying space. We intend, in particular, to focus on proving convergence and other mathematical properties of (6.3).

Acknowledgements M.F. acknowledges the financial support provided by the European Union's Human Potential Programme under the contract MOIF-CT-2006-039438. I.L. is a post-doctoral fellow with the F.W.O.-Vlaanderen (Belgium). M.F. and I.L. thank the Program in Applied and Computational Mathematics, Princeton University, for the hospitality during the preparation of this work. I.D. gratefully acknowledges partial support from NSF grants DMS-0245566 and 0530865.

References

1. Alber, Y.I., Iusem, A.N., Solodov, M.V.: On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. Math. Program. Ser. A **81**(1), 23–35 (1998)
2. Anthoine, S.: Different Wavelet-based Approaches for the Separation of Noisy and Blurred Mixtures of Components. Application to Astrophysical Data. Ph.D. Thesis, Princeton University, 2005
3. Brègman, L.M.: Finding the common point of convex sets by the method of successive projection. Dokl. Akad. Nauk SSSR **162**, 487–490 (1965)
4. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. Commun. Pure Appl. Math. **57**(2), 219–266 (2004)
5. Candès, E.J., Romberg, J.: Practical signal recovery from random projections. In: Wavelet Applications in Signal and Image Processing XI. Proc. SPIE Conf., vol. 5914 (2004)
6. Candès, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
7. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
8. Candès, E.J., Romberg, J., Tao, T.: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)

9. Canuto, C., Urban, K.: Adaptive optimization of convex functionals in Banach spaces. SIAM J. Numer. Anal. **42**(5), 2043–2075 (2004)
10. Chambolle, A., DeVore, R.A., Lee, N.-Y., Lucier, B.J.: Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. IEEE Trans. Image Process. **7**(3), 319–335 (1998)
11. Christensen, O.: An Introduction to Frames and Riesz Bases. Birkhäuser, Boston (2003)
12. Cohen, A.: Numerical Analysis of Wavelet Methods. Studies in Mathematics and Its Applications, vol. 32. North-Holland, Amsterdam (2003)
13. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods for elliptic operator equations—Convergence rates. Math. Comput. **70**, 27–75 (2001)
14. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods II: Beyond the elliptic case. Found. Comput. Math. **2**(3), 203–245 (2002)
15. Cohen, A., Hoffmann, M., Reiss, M.: Adaptive wavelet Galerkin methods for linear inverse problems. SIAM J. Numer. Anal. **42**(4), 1479–1501 (2004)
16. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
17. Dahlke, S., Maass, P.: An outline of adaptive wavelet Galerkin methods for Tikhonov regularization of inverse parabolic problems. In: Hon, J.-C., et al. (eds.) Recent Development in Theories and Numerics. Proceedings of the International Conference on Inverse Problems, Hong Kong, China, January 9–12 2002, pp. 56–66. World Scientific, Singapore (2003)
18. Dahlke, S., Fornasier, M., Raasch, T.: Adaptive frame methods for elliptic operator equations. Adv. Comput. Math. **27**(1), 27–63 (2007)
19. Dahlke, S., Fornasier, M., Raasch, T., Stevenson, R., Werner, M.: Adaptive frame methods for elliptic operator equations: The steepest descent approach. IMA J. Numer. Anal. **27**, 717–740 (2007)
20. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
21. Daubechies, I., Teschke, G.: Variational image restoration by means of wavelets: Simultaneous decomposition, deblurring, and denoising. Appl. Comput. Harmon. Anal. **19**(1), 1–16 (2005)
22. Daubechies, I., Defrise, M., DeMol, C.: An iterative thresholding algorithm for linear inverse problems. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
23. Donoho, D.L.: Superresolution via sparsity constraints. SIAM J. Math. Anal. **23**(5), 1309–1331 (1992)
24. Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory **41**(3), 613–627 (1995)
25. Donoho, D.L.: Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. Appl. Comput. Harmon. Anal. **2**(2), 101–126 (1995)
26. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
27. Donoho, D.L., Logan, B.F.: Signal recovery and the large sieve. SIAM J. Appl. Math. **52**, 577–591 (1992)
28. Donoho, D.L., Starck, P.: Uncertainty principles and signal recovery. SIAM J. Appl. Math. **49**, 906–931 (1989)
29. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
30. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Mathematics and Its Applications, vol. 375. Kluwer Academic, Dordrecht (1996)
31. Figueiredo, M.A.T., Nowak, R.D.: An EM algorithm for wavelet-based image restoration. IEEE Trans. Image Proc. **12**(8), 906–916 (2003)
32. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE J. Sel. Top. Signal Process. **1**(4), 586–597 (2007)
33. Fornasier, M.: Domain decomposition methods for linear inverse problems with sparsity constraints. Inverse Probl. **23**, 2505–2526 (2007)
34. Fornasier, M., Pitolli, F.: Adaptive iterative thresholding algorithms for magnetoenceophalography (MEG). J. Comput. Appl. Math. (2008, to appear). doi:10.1016/j.cam.2007.10.048
35. Fornasier, M., Rauhut, H.: Iterative thrsholding algorithms. Appl. Comput. Harmon. Anal. **25**(2), 187–208 (2008)
36. Fornasier, M., Rauhut, H.: Recovery algorithms for vector valued data with joint sparsity constraints. SIAM J. Numer. Anal. **46**(2), 577–613 (2008)
37. Logan, B.F.: Properties of High-Pass Signals. Ph.D. Thesis, Columbia University, 1965
38. Loris, I., Nolet, G., Daubechies, I., Dahlen, F.A.: Tomographic inversion using ℓ_1-norm regularization of wavelet coefficients. Geophys. J. Int. **170**(1), 359–370 (2007)

39. Mallat, S.: A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, San Diego (1999)
40. Ramlau, R., Teschke, G.: Tikhonov replacement functionals for iteratively solving nonlinear operator equations. Inverse Probl. **21**(5), 1571–1592 (2005)
41. Rauhut, H.: Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. **22**(1), 16–42 (2007)
42. Starck, J.-L., Candès, E.J., Donoho, D.L.: Astronomical image representation by curvelet transform. Astron. Astrophys. **298**, 785–800 (2003)
43. Starck, J.-L., Nguyen, M.K., Murtagh, F.: Wavelets and curvelets for image deconvolution: a combined approach. Signal Process. **83**, 2279–2283 (2003)
44. Teschke, G.: Multi-frame representations in linear inverse problems with mixed multi-constraints. Appl. Comput. Harmon. Anal. **22**(1), 43–60 (2007)
45. van den Berg, E., Friedlander, M.P.: In pursuit of a root. Preprint (2007)
46. Wolfram, S.: The Mathematica Book, 5th edn. Wolfram Media/Cambridge University Press, Cambridge (2003)