

FIXED-POINT CONTINUATION FOR ℓ_1 -MINIMIZATION: METHODOLOGY AND CONVERGENCE*

ELAINE T. HALE[†], WOTAO YIN[†], AND YIN ZHANG[†]

Abstract. We present a framework for solving the large-scale ℓ_1 -regularized convex minimization problem:

$$\min \|x\|_1 + \mu f(x).$$

Our approach is based on two powerful algorithmic ideas: operator-splitting and continuation. Operator-splitting results in a fixed-point algorithm for any given scalar μ ; continuation refers to approximately following the path traced by the optimal value of x as μ increases. In this paper, we study the structure of optimal solution sets, prove finite convergence for important quantities, and establish q -linear convergence rates for the fixed-point algorithm applied to problems with $f(x)$ convex, but not necessarily strictly convex. The continuation framework, motivated by our convergence results, is demonstrated to facilitate the construction of practical algorithms.

Key words. ℓ_1 regularization, fixed-point algorithm, q -linear convergence, continuation, compressed sensing

AMS subject classifications. 65K05, 90C06, 90C25, 90C90

DOI. 10.1137/070698920

1. Introduction. Under suitable conditions, minimizing the ℓ_1 -norm is equivalent to minimizing the so-called “ ℓ_0 -norm,” that is, the number of nonzeros in a vector. The former is always more computationally tractable than the latter. Thus, minimizing or limiting the magnitude of $\|x\|_1$ has long been recognized as a practical avenue for obtaining sparse solutions x . Some early work is in the area of geophysics, where sparse spike train signals are often of interest, and data may include large sparse errors [10, 39, 55, 57]. The signal processing and statistics communities use the ℓ_1 -norm to describe a signal with just a few waveforms or a response quantity with just a few explanatory variables [9, 24, 44, 58]. More references on ℓ_1 -regularization for signal processing and statistics can be found in [46].

In this work, we present an algorithmic framework and related convergence analysis for solving general problems of the form

$$(1.1) \quad \min_{x \in \mathbb{R}^n} \|x\|_1 + \mu f(x),$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable and convex, but not necessarily strictly convex, and $\mu > 0$. Interesting special cases of this problem include

$$(1.2) \quad \min_{x \in \mathbb{R}^n} \|x\|_1 + \frac{\mu}{2} \|Ax - b\|_2^2,$$

*Received by the editors July 31, 2007; accepted for publication (in revised form) June 17, 2008; published electronically October 31, 2008.

<http://www.siam.org/journals/siopt/19-3/69892.html>

[†]Department of Computational and Applied Mathematics, Rice University, 6100 Main Street, MS-134, Houston, TX 77005 (ehale@rice.edu, wotao.yin@rice.edu, yzhang@rice.edu). The work of E. Hale was supported by an NSF VIGRE grant (DMS-0240058). The work of W. Yin was supported in part by NSF CAREER award DMS-0748839 and ONR grant N00014-08-1-1101. The work of Y. Zhang was supported in part by NSF grants DMS-0405831 and DMS-0811188 and ONR grant N00014-08-1-1101.

and its generalization

$$(1.3) \quad \min_{x \in \mathbb{R}^n} \|x\|_1 + \frac{\mu}{2} \|Ax - b\|_M^2,$$

where $M \in \mathbb{R}^{m \times m}$ is a positive definite matrix, $\|x\|_M := \sqrt{x^\top M x}$ is the associated M -norm, $A \in \mathbb{R}^{m \times n}$ is dense, $m \leq n$ or even $m \ll n$, and n is large.

As a general principle, a sparse solution, $x \in \mathbb{R}^n$, of an underdetermined linear system of equations, $Ax = b$, may be obtained by minimizing the ℓ_1 -norm of x . If the “observation” b is contaminated with noise, then an appropriate norm of the residual $Ax - b$ should be minimized or constrained. Such considerations yield several related optimization problems. For instance, if there is Gaussian noise distributed as $N(0, \sigma^2 I)$ in b , then the ℓ_1 -regularized least squares problem (1.2) would be appropriate, as would the least absolute shrinkage and selection operator (LASSO) problem (6.3) [58].

Such problems are of fundamental importance to *compressed sensing*. Compressed sensing is the name assigned to the idea of encoding a large sparse signal using a relatively small number of linear measurements, and minimizing the ℓ_1 -norm (or its variants) in order to decode the signal. Recent results reported by Candes et al. [4, 5, 6], Donoho et al. [16, 20, 61], and others ([54, 60], for example) stimulated the current burst of research in this area. Applications of compressed sensing include compressive imaging [56, 65, 66], medical imaging [42], multisensor and distributed compressed sensing [1], analog-to-information conversion [36, 37, 38, 59], and missing data recovery [67]. Compressed sensing is attractive for these and other potential applications because it reduces the number of measurements required to obtain a given amount of information. The tradeoff is the addition of a nontrivial decoding process that consists of solving problems like (1.2), where the data matrix A is usually either random or has its rows taken from an orthogonal matrix such as a discrete cosine transform (DCT) matrix. Such data matrices are invariably dense and large in applications of interest. Thus we are motivated to study algorithms that do not require any linear system solves or matrix factorizations, and are able to take advantage of available fast transforms like FFT and DCT.

1.1. Our approach and main results. The objective function in (1.1) is the sum of two convex functions. While the ℓ_1 -norm term is not smooth, it is easily transformed into a linear function plus some linear constraints, such that standard interior-point methods utilizing a direct linear solver can be applied to, say, problem (1.2). However, such a standard approach is too costly for large-scale problems with dense data.

Our approach is based on operator splitting. It is well known in convex analysis that minimizing a convex function $\phi(x)$ is equivalent to finding a zero of the sub-differential $\partial\phi(x)$, i.e., finding x such that $\mathbf{0} \in \partial\phi(x) := T(x)$, where T is a maximal monotone operator [53]. In many cases, one can split ϕ into the sum of two convex functions, $\phi = \phi_1 + \phi_2$, which implies the decomposition of T into the sum of two maximal monotone operators T_1 and T_2 , i.e., $T = T_1 + T_2$. For $\tau > 0$, if T_2 is single-valued and $(I + \tau T_1)$ is invertible, then

$$(1.4) \quad \begin{aligned} \mathbf{0} \in T(x) &\iff \mathbf{0} \in (x + \tau T_1(x)) - (x - \tau T_2(x)) \\ &\iff (I - \tau T_2)x \in (I + \tau T_1)x \\ &\iff x = (I + \tau T_1)^{-1}(I - \tau T_2)x. \end{aligned}$$

Equation (1.4) leads to the *forward-backward splitting* algorithm for finding a zero of T :

$$(1.5) \quad x^{k+1} := (I + \tau T_1)^{-1}(I - \tau T_2)x^k,$$

which is a fixed-point algorithm. For the minimization problem (1.1), $T_2 = \mu \nabla f$ and $(I + \tau T_1)^{-1}$ is component-wise shrinkage (or soft-thresholding), which is related to the ℓ_1 -norm term in (1.1) and is described fully in sections 3 and 4 below.

The forward-backward splitting method was first proposed by Lions and Mercier [40] and Passty [51] at about the same time in 1979. Over the years, this scheme and its modifications have been extensively studied by various authors, including, to name a few, Mercier [45], Gabay [30], Glowinsky and Le Tallec [31], Eckstein [22], Chen and Rockafellar [8], Haubruge, Nguyen, and Strodiot [33], Noor [48], and Tseng [62]. The idea of splitting operators can be traced back to the mid-1950's in the works of Peaceman and Rachford [52] and Douglas and Rachford [21] for solving second-order elliptic and parabolic partial differential equations.

General convergence theory exists for forward-backward splitting methods [8, 30, 45]. Unfortunately, it requires rather strong conditions on T_2 , or on T as a whole. In short, when reduced to our setting with $\phi_1 = \|x\|_1$ and $\phi_2 = \mu f(x)$, the classical convergence theory requires either f or the whole objective in (1.1) to be strictly convex (though such strong assumptions may be weakened with modifications to the basic algorithm [62]). An anonymous referee brought our attention to the recent paper [12] by Combettes and Wajs, which applies forward-backward splitting methods to various concrete forms of minimizing a sum of two convex functions, including problem (1.1). Combettes and Wajs [12] show that the fixed-point iterations (1.5), and some extensions of it, converge to a global minimum without a strict convexity assumption.

In the present work, we aim to address the following two questions, “Can stronger convergence results be obtained for algorithm (1.5) applied to problem (1.1)?” and “Can algorithm (1.5) be computationally competitive when applied to problem (1.2) or (1.3)?” Our answers to both questions are affirmative.

On the theoretical side, we have obtained finite convergence for some interesting quantities (cf. Theorems 4.5 and 4.7) and q -linear¹ rates of convergence (cf. Proposition 4.9 and Theorems 4.10 and 4.11) without assuming strict convexity, nor uniqueness of solution. Furthermore, we show that these q -linear rates of convergence are not determined by the conditioning of the Hessian of f , as is normally the case for gradient-type methods, but by that of a “reduced” Hessian whose condition number can be much smaller than that of the full Hessian when the solution x is sparse.

On the computational side, we devised a continuation strategy that significantly reduces the number of iterations required for a given value of μ . Our extensive numerical results, which will be presented in a separate paper [32] due to space limitation, indicate that our algorithm is especially well suited for large-scale instances of problem (1.2) when x^* is sufficiently sparse and A is a partial transform matrix such as a partial DCT matrix. In comparison with several recently developed algorithms, our algorithm appears to be the most robust, and in many cases also the fastest.

1.2. Related work. Recently, solving problems (1.1) or (1.2), especially (1.2), has been actively studied by many authors, largely because of its newly found applications in signal and image processing. These problems can be solved by the forward-backward operator-splitting method given by (1.5) if one substitutes T_1 and T_2 by the subdifferential of $\|x\|_1$ and the gradient of $\mu f(x)$, respectively, resulting in the formula (3.1) of section 3. Because the operator-splitting approach was priorly not widely known in signal and image processing areas, in most recent works the

¹ $\{x^k\}$ converges to x^* q -linearly, where q stands for “quotient,” if $\limsup_k \|x^{k+1} - x^*\| / \|x^k - x^*\| < 1$.

fixed-point iteration (3.1), namely (1.5) specialized to (1.1), was derived by different authors, often independently, based on different motivations and approaches.

Here we mention a number of recent works that proposed, derived, or analyzed the fixed-point iteration scheme (1.5) or its variants when applied to either problem (1.2) or (1.1). These contributions include [29] by Figueiredo and Nowak, [15] by De Mol and Defrise, [13] by Daubechies, Defrise, and De Mol, [2] by Aubert, Bect, Blanc-Feraud, and Chambolle, [25, 26, 27] by Elad et al., and [?] by Darbon and Osher (though this list is unavoidably nonexhaustive). Some of these works were done independently around the same time, such as the two earlier papers [29] and [15]. Although the derivations and analyses in these papers were conducted through different means other than forward-backward operator splitting, the theoretical and numerical results in these papers contribute to a better understanding on the behavior of the fixed-point iterations (1.5) when applied to problem (1.2) or (1.1). For example, the authors of [2] and those of [13] analyzed the model (1.2) and independently proved global convergence without a strict convexity assumption, while classic convergence results for forward-backward operator splitting require stronger assumptions, and the authors of [26] proposed extensions and enhancements to the basic fixed-point iterations to improve its practical performance. As is already mentioned, Combettes and Wajs [12] recently established global convergence of the fixed-point iterations (1.5) when applied to (1.1) without a strict convexity assumption. A more recent paper by Combettes and Pesquet [11] studies closely related proximal soft-thresholding algorithms.

Alternative algorithms for the unconstrained ℓ_1 -problem (1.3) include an iterative linear solver in an interior-point framework [35] by Kim et al., a gradient projection and Barzilai–Borwein method applied to an equivalent box-constrained QP [28] by Figueiredo, Nowak, and Wright, a direct and accelerated projected gradient method [14] by Daubechies, Fornasier, and Loris, an accelerated multistep gradient method [47] with an error convergence rate $O(1/k^2)$ by Nesterov, and a “two-step” shrinkage-based algorithm [3] by Bioucas–Dias and Figueiredo. In [64], Van den Berg and Friedlander apply an iterative method for solving the LASSO problem (6.3).

1.3. Notation and organization. For simplicity, we let $\|\cdot\| := \|\cdot\|_2$, the Euclidean norm, unless otherwise specified. The *support* of $x \in \mathbb{R}^n$ is $\text{supp}(x) := \{i : x_i \neq 0\}$. Let

$$g(x) := \nabla f(x)$$

be the gradient of $f(x)$; in particular, $g(x) = A^\top M(Ax - b)$ for f defined by (1.3), that is

$$(1.6) \quad f = \frac{1}{2} \|Ax - b\|_M^2.$$

For a set E , we use $|E|$ to denote its cardinality. For any symmetric matrix $B \in \mathbb{R}^{n \times n}$, we denote its eigenvalues as $\lambda_i(B)$, $i = 1, \dots, n$, and its maximum and minimum eigenvalues as, respectively, $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$.

The signum function of $t \in \mathbb{R}$ is

$$\text{sgn}(t) := \begin{cases} +1 & t > 0, \\ 0 & t = 0, \\ -1 & t < 0, \end{cases}$$

while the signum multifunction (i.e., set-valued function) of $t \in \mathbb{R}$ is

$$\text{SGN}(t) := \partial|t| = \begin{cases} \{+1\} & t > 0, \\ [-1, 1] & t = 0, \\ \{-1\} & t < 0, \end{cases}$$

which is also the subdifferential of $|t|$. For $x \in \mathbb{R}^n$, we define $\text{sgn}(x) \in \mathbb{R}^n$ and $\text{SGN}(x) \subset \mathbb{R}^n$ component-wise as $(\text{sgn}(x))_i := \text{sgn}(x_i)$ and $(\text{SGN}(x))_i := \text{SGN}(x_i)$, $i = 1, 2, \dots, n$, respectively. Clearly,

$$\text{sgn}(x) = \text{sgn}(x') \iff \text{SGN}(x) = \text{SGN}(x'), \forall x, x'.$$

For $x, y \in \mathbb{R}^n$, let $x \odot y \in \mathbb{R}^n$ denote the component-wise product of x and y , i.e., $(x \odot y)_i = x_i y_i$. Furthermore, vector operators such as $|x|$ and $\max\{x, y\}$ are defined to operate component-wise as well, analogous to the definitions of sgn and SGN .

For any index set $I \subseteq \{1, 2, \dots, n\}$ (later, we will use index sets E and L), x_I is defined as the subvector of x of length $|I|$ consisting only of components x_i , $i \in I$. Similarly, if g is a vector-valued function, then $g_I(x)$ denotes the subvector of $g(x)$ consisting of $g_i(x)$, $i \in I$.

This paper is organized as follows. In section 2, we recall the classic optimality (or in general, stationarity) conditions for problem (1.1), and then characterize the optimal solution sets of problems (1.1) and (1.3). In section 3, we present a fixed-point optimality condition for (1.1). This optimality condition motivates a fixed-point algorithm and introduces the shrinkage operator, the properties of which conclude section 3. In section 4, we present our results on the convergence and rates of convergence of the fixed-point algorithm; the proofs of the main results are given in section 5. We motivate and propose a continuation method in section 6, support this proposal with a few numerical results, and briefly discuss some possible extensions. Finally, we conclude the paper in section 7.

2. Optimality and optimal solution sets. Recall that $f(x)$ in (1.1) is convex, and let X^* be the set of optimal solutions of (1.1). It is well known from convex analysis (see, for example, [53]) that an optimality condition for (1.1) is

$$(2.1) \quad x \in X^* \iff \mathbf{0} \in \text{SGN}(x) + \mu g(x),$$

where $\mathbf{0}$ is the zero vector in \mathbb{R}^n , or equivalently,

$$(2.2) \quad x \in X^* \iff \mu g_i(x) \begin{cases} = -1, & x_i > 0, \\ \in [-1, 1], & x_i = 0, \\ = 1, & x_i < 0. \end{cases}$$

It follows readily from (2.2) that $\mathbf{0}$ is an optimal solution of (1.1) if and only if $\mu \|g(\mathbf{0})\|_\infty \leq 1$; therefore, it is easy to check whether $\mathbf{0}$ is a solution of (1.1).

We note that the solution set X^* may have more than one element. The following theorem establishes some properties of X^* that are of interest in their own right, but will also be useful in later developments.

THEOREM 2.1. *Let $f \in C^2$ be convex and X^* be the set of optimal solutions of (1.1), which is nonempty.*

1. *If $x^1 \in X^*$ and $x^2 \in X^*$, then $g(x^1) = g(x^2)$.*

2. $x \in X^*$ if and only if $g(x) \equiv g^*$, where for $i = 1, 2, \dots, n$,

$$(2.3) \quad \mu g_i^* \begin{cases} = -1, & \max\{x_i : x \in X^*\} > 0, \\ = +1, & \min\{x_i : x \in X^*\} < 0, \\ \in [-1, 1], & \text{otherwise.} \end{cases}$$

3. X^* is contained in a single orthant of \mathbb{R}^n ; more precisely

$$(2.4) \quad X^* \subset O := \{x \in \mathbb{R}^n : -\text{sgn}^+(g_i^*)x_i \geq 0, \forall i\},$$

where $\text{sgn}^+(\cdot)$ is equal to $\text{sgn}(\cdot)$ except that $\text{sgn}^+(0) := 1$, i.e.,

$$\text{sgn}^+(t) := \begin{cases} +1 & t \geq 0, \\ -1 & t < 0. \end{cases}$$

(In addition, we let $\text{sgn}^+(x)$ be defined component-wise for any $x \in \mathbb{R}^n$.)

Furthermore, if $f(x)$ is the quadratic defined as in (1.6), then

4. If $x^1 \in X^*$ and $x^2 \in X^*$, then $Ax^1 = Ax^2$.
5. $\|x\|_1$ and $\|Ax - b\|_M$ are constant for all $x \in X^*$.
6. X^* is a bounded polyhedron, i.e., a polytope.

Proof. We prove the statements one by one.

1. This part will be proven later as Corollary 4.2 under Assumption 1, which is slightly weaker than what is assumed for this theorem. That proof is independent of this theorem and the results that follow from it.
2. (2.3) follows directly from part 1 and (2.2) applied to all $x \in X^*$.
3. From (2.1) and (2.3), if there exists an $x \in X^*$ with a strictly positive (negative) x_i , then $\mu g_i^* = -1$ ($\mu g_i^* = 1$), so all other $x \in X^*$ must satisfy $x_i \geq 0$ ($x_i \leq 0$). Consequently, X^* lies in the orthant O .
4. From part 1 and for the quadratic $f(x)$ so specified, $g(x^1) - g(x^2) = A^\top M A (x^1 - x^2) = \mathbf{0}$, which immediately implies that $A(x^1 - x^2) = \mathbf{0}$, given that M is symmetric positive definite.
5. From part 4, Ax is constant over $x \in X^*$, and so is $\|Ax - b\|_M$. Since (1.3) has a unique optimal objective value, $\|x\|_1$ must also be constant.
6. Defining $p = -\text{sgn}^+(g^*)$, from the definition of O we have

$$p^\top x = \|x\|_1, \quad \forall x \in O.$$

Consider the linear program

$$(2.5) \quad \min_x \{p^\top x : Ax = c, x \in O\},$$

where $c = Ax$ for any $x \in X^*$. It is easy to verify that an optimal solution \bar{x} of (2.5) satisfies both $\|\bar{x}\|_1 = \|x\|_1$ and $\|A\bar{x} - b\|_M = \|Ax - b\|_M$ for any $x \in X^*$ and vice versa. So (2.5) is equivalent to (1.3), as long as c and O (or equivalently, g^*) are known. Consequently, X^* , as the solution set of the linear program (2.5), must be a polyhedron and must be bounded since $\|x\|_1$ is constant for all $x \in X^*$.

This completes the proof. \square

3. A fixed-point algorithm.

3.1. Optimality as a fixed-point equation. We start with another optimality condition for problem (1.1): for any scalar $\tau > 0$, $x^* \in X^*$ if and only if

$$(3.1) \quad x^* = \text{sgn}(x^* - \tau g(x^*)) \odot \max \left\{ |x^* - \tau g(x^*)| - \frac{\tau}{\mu}, \mathbf{0} \right\}.$$

The derivation of (3.1) can be found, for example, in [12].

It is straightforward to verify that optimality condition (3.1) can be replaced by

$$(3.2) \quad x^* = \text{sgn}(x^* - d(x^*) \odot g(x^*)) \odot \max \left\{ |x^* - d(x^*) \odot g(x^*)| - \frac{d(x^*)}{\mu}, \mathbf{0} \right\},$$

where the positive scalar τ in (3.1) is replaced by any mapping d from \mathbb{R}^n to \mathbb{R}^n such that $(d(x))_i = d_i(x_i) > 0$. The algorithm based on (3.1), and its analysis as well, can be readily extended to those based on (3.2) (see [26] for a study of such an extension).

The right-hand side of the fixed-point equation (3.1) is a composition of two mappings from \mathbb{R}^n to \mathbb{R}^n defined as

$$(3.3) \quad h(\cdot) := I(\cdot) - \tau g(\cdot),$$

$$(3.4) \quad s_\nu(\cdot) := \text{sgn}(\cdot) \odot \max\{|\cdot| - \nu, 0\}, \text{ where } \nu > 0.$$

Intuitively, $h(\cdot)$ resembles a gradient descent step for $f(x)$ with the stepsize $\tau > 0$, and $s_\nu(\cdot)$ reduces the magnitude of each nonzero component of the input vector by an amount less than or equal to ν , thus reducing the ℓ_1 -norm. Later we will also use s_ν as a mapping from \mathbb{R} to \mathbb{R} in composition with $h_i(\cdot) = (h(\cdot))_i$ from \mathbb{R}^n to \mathbb{R} .

Equation (3.1) leads to *the fixed-point iterations*

$$(3.5) \quad x^{k+1} = s_\nu \circ h(x^k) \text{ with } \tau > 0, \nu = \tau/\mu,$$

which can be derived by operator-splitting or other means as has been done by the authors mentioned in subsection 1.2. As many others before us, we originally derived the fixed-point scheme (3.5) from a totally different approach, and later found that it can be interpreted as the forward-backward splitting algorithm (1.5) with

$$T_1(x) = \partial\|x\|_1/\mu, \text{ and } T_2(x) = g(x),$$

since simple calculations show that

$$s_\nu = (I + \tau T_1)^{-1}, \text{ and } h = I - \tau T_2.$$

However, some special properties of the operator s_ν , given below, will allow us to obtain strong convergence results that do not directly follow from the existing theory for forward-backward splitting algorithms applied to more general operators.

The main algorithm of the paper, Algorithm 1, is based on (3.5) and will be presented in subsection 6.2 along with choices for τ and μ .

3.2. Properties of the shrinkage operator. It is easy to verify that $s_\nu(y)$ is the unique solution of

$$\min_{x \in \mathbb{R}^n} \nu\|x\|_1 + \frac{1}{2}\|x - y\|^2$$

for any $y \in \mathbb{R}^n$. Wavelet analysts refer to $s_\nu(\cdot)$ as the soft-thresholding [19] or wavelet shrinkage [7] operator. For convenience, we will refer to $s_\nu(\cdot)$ as the *shrinkage operator*.

The two lemmas below establish some useful properties of the shrinkage operator. Both make immediate use of the component-wise separability of s_ν , that is, for all indices i

$$(s_\nu(y))_i = s_\nu(y_i).$$

The alternative representation of s_ν in Lemma 3.1 will be used to prove Lemma 3.2. Lemma 3.2 proves a number of component-wise properties of s_ν , including nonexpansiveness. These results will be used to prove convergence results for (3.5) that are not implied by convergence results in [12].

With a slight abuse of notation, we let $\mathcal{P}(x)$ denote the projection of a vector $x \in \mathbb{R}^k$ onto the k -cube $[-\nu, \nu]^k$ for any positive integer k , since the projection onto any k -cube is done component-wise. Now Lemma 3.1 below can be trivially verified by enumerating all possible cases.

LEMMA 3.1. *The shrinkage operator can be written as*

$$(3.6) \quad s_\nu(y) = y - \mathcal{P}(y), \quad \forall y \in \mathbb{R}^n,$$

and the equation holds component-wise; i.e., $(s_\nu(y))_i \equiv s_\nu(y_i) = y_i - \mathcal{P}(y_i)$. Moreover, both $s_\nu(y)$ and $\mathcal{P}(y)$ are component-wise monotone.

LEMMA 3.2. *The operator $s_\nu(\cdot)$ is component-wise nonexpansive and for any $y^1, y^2 \in \mathbb{R}^n$,*

$$(3.7) \quad |s_\nu(y_i^1) - s_\nu(y_i^2)| = |y_i^1 - y_i^2| - |\mathcal{P}(y_i^1) - \mathcal{P}(y_i^2)|, \quad \forall i.$$

Consequently, s_ν is nonexpansive in any ℓ_p (quasi-)norm with $p \geq 0$, and if h is nonexpansive in a given norm, then $s_\nu \circ h$ is as well. Moreover, consider the case when

$$(3.8) \quad |s_\nu(y_i^1) - s_\nu(y_i^2)| = |y_i^1 - y_i^2|,$$

which we refer to as the no-shrinkage condition. We have, for each index i :

1. (3.8) $\implies \mathcal{P}(y_i^1) = \mathcal{P}(y_i^2)$, $\text{sgn}(y_i^1) = \text{sgn}(y_i^2)$, $s_\nu(y_i^1) - s_\nu(y_i^2) = y_i^1 - y_i^2$.
2. (3.8) and $y_i^1 \neq y_i^2 \implies |y_i^1| \geq \nu$, $|y_i^2| \geq \nu$ and $|y_i^1| \neq |y_i^2|$.
3. (3.8) and $|y_i^2| < \nu \implies |y_i^1| < \nu$, $y_i^1 = y_i^2$, $s_\nu(y_i^1) = s_\nu(y_i^2) = 0$.
4. (3.8) and $|y_i^2| \geq \nu \implies |y_i^1| \geq \nu$.
5. $|y_i^2| \geq \nu$ and $\text{sgn}(y_i^1) \neq \text{sgn}(y_i^2) \implies |s_\nu(y_i^1) - s_\nu(y_i^2)| \leq |y_i^1 - y_i^2| - \nu$.
6. $s_\nu(y_i^1) \neq 0 = s_\nu(y_i^2) \implies |y_i^1| > \nu$, $|y_i^2| \leq \nu$, $|s_\nu(y_i^1) - s_\nu(y_i^2)| \leq |y_i^1 - y_i^2| - (\nu - |y_i^2|)$.

Proof. For ease of notation, we drop the subscript i and let $p^1 = y_i^1$ and $p^2 = y_i^2$. Without loss of generality we assume that $p^1 \geq p^2$. Hence, from the monotonicity of s_ν and \mathcal{P} we have $s_\nu(p^1) \geq s_\nu(p^2)$ and $\mathcal{P}(p^1) \geq \mathcal{P}(p^2)$. Therefore,

$$\begin{aligned} |s_\nu(p^1) - s_\nu(p^2)| &= s_\nu(p^1) - s_\nu(p^2) = p^1 - p^2 - (\mathcal{P}(p^1) - \mathcal{P}(p^2)) \\ &= |p^1 - p^2| - |\mathcal{P}(p^1) - \mathcal{P}(p^2)|, \end{aligned}$$

which proves (3.7).

Next, we move to proving parts 5 and 6, omitting the proofs for parts 1 through 4 since, given (3.6) and (3.7), they all can be similarly and easily verified.

For part 5, it suffices to show that $|\mathcal{P}(p^1) - \mathcal{P}(p^2)| \geq \nu$. Without loss of generality, we assume that $p^1 > 0$ and $p^2 < 0$. Hence, $\mathcal{P}(p^1) \geq 0$, $\mathcal{P}(p^2) = -\nu$, and $|\mathcal{P}(p^1) - \mathcal{P}(p^2)| = \mathcal{P}(p^1) - \mathcal{P}(p^2) \geq \nu$.

Finally, we verify part 6. First, it is easy to see that $|p^1| > \nu$ and $|p^2| \leq \nu$. Hence, $|\mathcal{P}(p^1)| = \nu$ and $\mathcal{P}(p^2) = p^2$, and $|\mathcal{P}(p^1) - \mathcal{P}(p^2)| \geq |\mathcal{P}(p^1)| - |\mathcal{P}(p^2)| = \nu - |p^2|$. \square

4. Convergence analysis. In this section, we study the convergence of the fixed-point iterations (3.5) applied to the general ℓ_1 -regularized minimization problem (1.1) and the quadratic case (1.3). Assumption 1 below, which states that f is a convex function with bounded Hessian in a neighborhood of an optimal solution of (1.1), is sufficient for our global convergence result and will be applied throughout. Further assumptions (primarily on the rank of a particular minor of the Hessian of f) will be made to obtain linear convergence rate results in section 4.2.

Assumption 1. Problem (1.1) has an optimal solution set $X^* \neq \emptyset$, and there exists a bounded convex set $\Omega \supset X^*$ such that $f \in C^2(\Omega)$, $H(x) := \nabla^2 f(x) \succeq 0$ for $x \in \Omega$ and

$$(4.1) \quad \hat{\lambda}_{\max} := \max_{x \in \Omega} \lambda_{\max}(H(x)) < \infty.$$

For simplicity, we will use a constant parameter τ in the fixed-point iterations (3.5): $x^{k+1} = s_\nu(x^k - \tau g(x^k))$, where

$$(4.2) \quad \nu = \tau/\mu.$$

In particular, we will always choose

$$(4.3) \quad \tau \in \left(0, 2/\hat{\lambda}_{\max}\right),$$

which guarantees that $h(\cdot) = I(\cdot) - \tau g(\cdot)$ is nonexpansive in Ω , and contractive in the range space of H in the quadratic case. Our analysis can be extended to the case of variable τ , but this would require more complicated notation and a reduction of clarity.

4.1. Global and finite convergence. From the mean-value theorem, we recall that for any $x, x' \in \Omega$

$$(4.4) \quad g(x) - g(x') = \left(\int_0^1 H(x' + t(x - x')) dt\right) (x - x') := \bar{H}(x, x')(x - x').$$

This fact is used to verify the nonexpansiveness of h and the result that noncontraction between any two points under h implies that the gradient of f is equal at those points.

LEMMA 4.1. *Under Assumption 1 and the choice of τ specified in (4.3), $h(\cdot) = I(\cdot) - \tau g(\cdot)$ is nonexpansive in Ω , i.e., for any $x, x' \in \Omega$,*

$$(4.5) \quad \|h(x) - h(x')\| \leq \|x - x'\|.$$

Moreover, $g(x) = g(x')$ whenever equality holds in (4.5).

Proof. Let $\bar{H} := \bar{H}(x, x')$. We first note that

$$h(x) - h(x') = x - x' - \tau(g(x) - g(x')) = (I - \tau\bar{H})(x - x').$$

Hence, in view of (4.3),

$$\begin{aligned} \|h(x) - h(x')\| &= \|(I - \tau\bar{H})(x - x')\| \\ &\leq \max\{|1 - \tau\lambda_{\max}(\bar{H})|, |1 - \tau\lambda_{\min}(\bar{H})|\} \|x - x'\| \\ &\leq \max\{|1 - \tau\hat{\lambda}_{\max}|, 1\} \|x - x'\| \\ &\leq \|x - x'\|. \end{aligned}$$

To prove the second statement, let $s := x - x'$ and $p := \bar{H}^{1/2}s$. Then

$$\begin{aligned} \|h(x) - h(x')\| = \|x - x'\| &\iff \|s - \tau\bar{H}s\| = \|s\| \\ &\iff -2\tau s^T \bar{H}s + \tau^2 s^T \bar{H}^2 s = 0 \\ &\iff \tau p^T \bar{H}p = 2p^T p \\ &\implies \tau \frac{p^T \bar{H}p}{p^T p} = 2 \text{ if } p \neq 0, \end{aligned}$$

which contradicts (4.3) since $\frac{p^T \bar{H}p}{p^T p} \leq \hat{\lambda}_{\max} < \frac{2}{\tau}$. Hence, $p = 0$ so that

$$g(x) - g(x') = \bar{H}^{1/2}p = 0$$

whenever h is noncontractive. \square

Since any two fixed points, say x and x' , of the nonexpansive mapping $s_\nu \circ h$ must satisfy the equality

$$(4.6) \quad \|x - x'\| = \|s_\nu \circ h(x) - s_\nu \circ h(x')\| = \|h(x) - h(x')\|,$$

Lemma 4.1 shows that the gradient of f evaluated at any two fixed points must be equal. Hence, we have the following corollary and the first statement of Theorem 2.1.

COROLLARY 4.2 (Constant optimal gradient). *Under Assumption 1, there is a vector $g^* \in \mathbb{R}^n$ such that*

$$(4.7) \quad g(x^*) \equiv g^*, \quad \forall x^* \in X^*.$$

We will use the following partition of all indices $\{1, \dots, n\}$ into L and E to obtain finite convergence for components in L and linear convergence for components in E .

DEFINITION 4.3. *Let $X^* \neq \emptyset$ be the solution set of (1.1) and g^* be the vector specified in Corollary 4.2. Define*

$$(4.8) \quad L := \{i : \mu|g_i^*| < 1\} \quad \text{and} \quad E := \{i : \mu|g_i^*| = 1\}.$$

It is clear from the optimality condition (2.2) that $L \cup E = \{1, 2, \dots, n\}$,

$$(4.9) \quad \text{supp}(x^*) \subseteq E, \quad \text{and} \quad x_i^* = 0, \quad \forall i \in L, \quad \forall x^* \in X^*.$$

There are examples in which $\text{supp}(x^*) \neq E$, so the two vectors $|x^*|$ and $\mathbf{1} - \mu|g^*|$ are always complementary, but may not be strictly complementary.

The positive scalar ω defined below will also play a key role in the finite convergence property of the fixed-point iterations:

DEFINITION 4.4. *Let g^* be the vector specified in Corollary 4.2. Define*

$$(4.10) \quad \omega := \min\{\nu(1 - \mu|g_i^*|) : i \in L\}.$$

From the definition, clearly $\omega > 0$. In addition, since (4.9) implies that for all $x^* \in X^*$ and all $i \in L$

$$\nu(1 - \mu|g_i^*|) = \nu - \tau|g_i^*| = \nu - |x_i^* - \tau g_i(x^*)| = \nu - |h_i(x^*)|,$$

and consequently, for any $x^* \in X^*$,

$$(4.11) \quad \min\{\nu - |h_i(x^*)| : i \in L\} = \omega > 0.$$

We now claim that Assumption 1 is sufficient for obtaining convergence of the fixed-point iterations (3.5) and finite convergence for components in L and signs in E . We reiterate that under similar conditions, convergence has been established in [12].

THEOREM 4.5 (the general case). *Under Assumption 1, the sequence $\{x^k\}$ generated by the fixed-point iterations (3.5) applied to problem (1.1) from any starting point $x^0 \in \Omega$ converges to some $x^* \in X^* \cap \Omega$. In addition, for all but finitely many iterations, we have*

$$(4.12) \quad x_i^k = x_i^* = 0, \quad \forall i \in L,$$

$$(4.13) \quad \text{sgn}(h_i(x^k)) = \text{sgn}(h_i(x^*)) = -\mu g_i^*, \quad \forall i \in E,$$

where the numbers of iterations not satisfying (4.12) and (4.13) do not exceed $\|x^0 - x^*\|^2/\omega^2$ and $\|x^0 - x^*\|^2/\nu^2$, respectively, for ω defined in (4.10) and $\nu = \tau/\mu$.

The proof of Theorem 4.5 is rather lengthy and is therefore relegated to the next section. A majority of the proof concerns the finite convergence properties which are new. For the sake of completeness, we also include a proof for global convergence which is known.

In light of this theorem, every starting point $x^0 \in \Omega$ determines a converging sequence $\{x^k\}$ whose limit is a solution of (1.1). Generally, the solutions of (1.1) may be nonunique, as it is not difficult to construct simple examples for which different starting points lead to different solutions.

We recall that x_E and g_E^* are defined as the subvectors of x and g^* with components x_i and g_i^* , $i \in E$, respectively. Without loss of generality, we assume $E = \{1, 2, \dots, |E|\}$, and let $(x_E; \mathbf{0})$ denote the vector in \mathbb{R}^n obtained from x by setting the components $x_i \forall i \in L$ to zero. The following corollary enables one to apply any convergence results for the gradient projection method to the fixed-point iterations (3.5).

COROLLARY 4.6. *Under Assumption 1 and starting from some $x^0 \in \Omega$, after a finite number of iterations the fixed-point iterations (3.5) reduce to gradient projection iterations for minimizing $\phi(x_E)$ over a constraint set O_E , where*

$$(4.14) \quad \phi(x_E) := -(g_E^*)^\top x_E + f((x_E; \mathbf{0})), \quad \text{and}$$

$$(4.15) \quad O_E = \{x_E \in \mathbb{R}^{|E|} : -\text{sgn}(g_E^*) \odot x_E \geq 0\}.$$

Specifically, we have $x^{k+1} = (x_E^{k+1}; \mathbf{0})$ in which

$$(4.16) \quad x_E^{k+1} := P_{O_E}(x_E^k - \tau \nabla \phi(x_E^k)),$$

where P_{O_E} is the orthogonal projection onto O_E , and $\nabla \phi(x_E) = -g_E^* + g_E((x_E; \mathbf{0}))$.

Proof. From Theorem 4.5, there exists $K > 0$ such that for $k \geq K$ (4.12)–(4.13) hold. Let $k > K$. Since $x_i^k = 0$ for $i \in L$, it suffices to consider $i \in E$. For

$i \in E$, we have $x_i^k \geq 0$ if $\text{sgn}(h_i(x^{k-1})) = 1$ (equivalently, $g_i^* < 0$) and $x_i^k \leq 0$ if $\text{sgn}(h_i(x^{k-1})) = -1$ ($g_i^* > 0$). Therefore, for any i , $-g_i^* x_i^k \geq 0$ for all $k > K$. Hence, $x^k \in O$ according to the definition (2.4) of O and $x_E^k \in O_E$.

For $i \in E$, we calculate the quantity

$$\begin{aligned} y_i^{k+1} &:= x_i^k - \tau (\nabla \phi(x^k))_i \\ &= x_i^k - \tau (-g_i^* + g_i(x^k)) \\ &= h_i(x^k) + \nu \mu g_i^* \\ &= \text{sgn}(h_i(x^k)) (|h_i(x^k)| - \nu), \end{aligned}$$

where (4.13) was used to obtain the last expression. Clearly, the fixed-point iterations (3.5) restricted to the components $i \in E$ are

$$(x_E^{k+1})_i = s_\nu \circ h_i(x^k) = \begin{cases} y_i^{k+1}, & -g_i^* y_i^{k+1} \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Equivalently,

$$(x_E^{k+1})_i = (P_{O_E}(x_E^k - \tau \nabla \phi(x_E^k)))_i,$$

which completes the proof. \square

Finally, a stronger global convergence result for convex quadratic functions, namely, f as in (1.6), follows directly from the general convergence result. We note that Assumption 1 is no longer necessary if the convex quadratic is bounded below. Due to the importance of the quadratic case, we state a separate theorem.

THEOREM 4.7 (the quadratic case). *Let f be a convex quadratic function that is bounded below, H be its Hessian, and τ satisfy*

$$(4.17) \quad 0 < \tau < 2/\lambda_{\max}(H).$$

Then the sequence $\{x^k\}$, generated by the fixed-point iterations (3.5) from any starting point x^0 , converges to some $x^ \in X^*$. In addition, (4.12)–(4.13) hold for all but finitely many iterations.*

4.2. Linear rate of convergence. Let $\{x^k\}$ be generated by the fixed-point iterations (3.5) starting from any $x^0 \in \Omega$. We know that the sequence converges to some point in X^* . Throughout this subsection, we let $x^0 \in \Omega$,

$$x^* := \lim_{k \rightarrow \infty} x^k,$$

and study the rate of convergence of $\{x^k\}$ to x^* under different assumptions. Note that $\Omega = \mathbb{R}^n$ if f is convex quadratic, and recall that a sequence $\{\|x^k - x^*\|\}$ converges to zero q -linearly if its q_1 -factor is less than one, i.e., if

$$q_1 := \limsup_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} < 1,$$

while it is r -linearly convergent if it is bounded by a q -linearly convergent sequence.

As we will show, under appropriate assumptions q -linear convergence holds for any $\tau \in (0, 2/\hat{\lambda}_{\max})$. However, the q_1 -factor may vary with different choices of τ . In particular, we consider choices of the form

$$(4.18) \quad \tau(\lambda) := \frac{\gamma(\lambda)}{\gamma(\lambda) + 1} \frac{2}{\hat{\lambda}_{\max}}, \quad \gamma(\lambda) := \frac{\hat{\lambda}_{\max}}{\lambda},$$

where $\hat{\lambda}_{\max}$ is defined in (4.1) and $\lambda > 0$ will take different values under different assumptions. It is easy to see that $\tau(\lambda) \in (0, 2/\hat{\lambda}_{\max})$ since $\gamma(\lambda) > 0$.

Some of our assumptions will involve the matrix H_{EE} :

DEFINITION 4.8. Let $H(x)$ denote the Hessian of f evaluated at $x \in \Omega$, and

$$H_{EE}(x) := [H_{i,j}(x)]_{i,j \in E}$$

denote the square submatrix of H corresponding to the index set E defined in (4.8).

Based on Corollary 4.6, we first apply existing convergence results for the gradient projection method applied to (4.16).

PROPOSITION 4.9. Let Assumption 1 hold. If (i) $H_{EE}(x^*)$ has full rank, or (ii) f is defined as in (1.6), then $\{\|x^k\|_1 + \mu f(x^k)\}$ converges to $\|x^*\| + \mu f(x^*)$ q -linearly and $\{x^k\}$ converges to x^* r -linearly.

Under the first condition, the above result follows from [50] and [41], while under the second condition it follows from [41]. However, by directly analyzing the original fixed-point iterations, we can strengthen the convergence rate of $\{x^k\}$ from r -linear to q -linear. Theorem 4.10 does this under the assumption that $H_{EE}(x^*)$ is full rank; Theorem 4.11 instead assumes that $\text{supp}(x^*) = E$. We first define

$$(4.19) \quad \bar{H}^k \equiv \bar{H}(x^k, x^*) := \int_0^1 H(x^* + t(x^k - x^*)) dt.$$

THEOREM 4.10. Let Assumption 1 hold, and assume that

$$(4.20) \quad \lambda_{\min}^E := \lambda_{\min}(H_{EE}(x^*)) > 0.$$

Then for any $\tau \in (0, 2/\hat{\lambda}_{\max})$, $\{x^k\}$ converges to x^* q -linearly. Moreover, if τ is chosen as in (4.18) with $\lambda = \lambda_{\min}^E$, then the q_1 -factor satisfies

$$(4.21) \quad q_1 \leq \frac{\gamma(\lambda_{\min}^E) - 1}{\gamma(\lambda_{\min}^E) + 1}.$$

Proof. Without loss of generality, we assume that all iteration counts, k , are large enough so that $x_i^k = x_i^* = 0$ for all $i \in L$, and that the spectrum of \bar{H}_{EE}^k falls in the interval $[\lambda_{\min}^E - \epsilon, \hat{\lambda}_{\max}]$ for an arbitrary $\epsilon > 0$. (The first assumption on k is valid because of the finite convergence properties of Theorem 4.5; the second follows from the continuity of the Hessian.) Since $x_i^k = x_i^* = 0, \forall i \in L$, the mean-value theorem yields

$$h_E(x^k) - h_E(x^*) = x_E^k - x_E^* - \tau(g_E(x^k) - g_E(x^*)) = (I - \tau \bar{H}_{EE}^k)(x_E^k - x_E^*).$$

Recall that $x^{k+1} = s_\nu \circ h(x^k)$ and $s_\nu(\cdot)$ is nonexpansive. Hence,

$$\begin{aligned} \|x^{k+1} - x^*\| &\equiv \|x_E^{k+1} - x_E^*\| \\ &\leq \|h_E(x_E^k) - h_E(x_E^*)\| \\ &\leq \|I - \tau \bar{H}_{EE}^k\| \|x_E^k - x_E^*\| \\ &\leq \max \left\{ |1 - \tau \hat{\lambda}_{\max}|, |1 - \tau \lambda_{\min}^E| + \tau \epsilon \right\} \|x_E^k - x_E^*\| \\ &\equiv \max \left\{ |1 - \tau \hat{\lambda}_{\max}|, |1 - \tau \lambda_{\min}^E| + \tau \epsilon \right\} \|x^k - x^*\|. \end{aligned}$$

Clearly, $\max\{|1 - \tau\hat{\lambda}_{\max}|, |1 - \tau\lambda_{\min}^E| + \tau\epsilon\}$ is less than one for any $\tau \in (0, 2/\hat{\lambda}_{\max})$ and ϵ sufficiently small; in particular, it equals the right-hand side of (4.21) plus $\tau\epsilon$ when $\tau = \tau(\lambda_{\min}^E)$. Since ϵ is arbitrary, (4.21) must hold. \square

THEOREM 4.11. *Let Assumption 1 hold, and also assume that x^* satisfies (i) $\text{supp}(x^*) = E$ or, equivalently, the strict complementarity condition*

$$(4.22) \quad |x^*| + (1 - \mu|g^*|) > 0,$$

and (ii) the range space $\mathcal{R}(H_{EE}(x))$ of $H_{EE}(x)$ is invariant in a neighborhood N^ of x^* . Whenever $H_{EE}(x^*) \neq \mathbf{0}$, let*

$$(4.23) \quad \lambda_{\min}^{\mathcal{R}} := \lambda_{\min}(V^{\top}H_{EE}(x^*)V) > 0,$$

where V is any orthonormal basis of $\mathcal{R}(H_{EE}(x^))$.*

If $H_{EE}(x^) = \mathbf{0}$, then $x^k = x^*$ for all k sufficiently large; otherwise $\{x^k\}$ converges to x^* q -linearly for any $\tau \in (0, 2/\hat{\lambda}_{\max})$. In the latter case, if τ is chosen as in (4.18) with $\lambda = \lambda_{\min}^{\mathcal{R}}$, then the q_1 -factor satisfies*

$$(4.24) \quad q_1 \leq \frac{\gamma(\lambda_{\min}^{\mathcal{R}}) - 1}{\gamma(\lambda_{\min}^{\mathcal{R}}) + 1}.$$

The proof of this theorem is given in section 5.2. We note that $\mathcal{R}(H_{EE}(x))$ is invariant near x^* if either f is a quadratic function or $H_{EE}(x^*)$ has full rank.

Since Assumption 1 is not required in the proof of global convergence for convex quadratic f , we can directly derive the following results for this case, which is the situation one encounters with compressed sensing. The proof, which is similar to those of Theorems 4.10 and 4.11, is left to the reader.

COROLLARY 4.12. *Let f be a convex quadratic function that is bounded below, and $\{x^k\}$ be the sequence generated by the fixed-point iterations (3.5) with $\tau \in (0, 2/\lambda_{\max}(H))$.*

1. *If H_{EE} has full rank, then $\{x^k\}$ converges to x^* q -linearly. Moreover, if τ is chosen as in (4.18) with $\lambda = \lambda_{\min}(H_{EE})$, then the q_1 -factor satisfies*

$$q_1 \leq \frac{\gamma(\lambda_{\min}(H_{EE})) - 1}{\gamma(\lambda_{\min}(H_{EE})) + 1}.$$

2. *Let x^* satisfy the strict complementarity condition (4.22). Then if $H_{EE} = \mathbf{0}$, $\{x^k\}$ converges to x^* in a finite number of steps. Otherwise $\{x^k\}$ converges to x^* q -linearly, and if τ is chosen as in (4.18) with $\lambda := \lambda_{\min}(V^{\top}H_{EE}V)$, where V is an orthonormal basis for the range space of H_{EE} , then the q_1 -factor satisfies*

$$q_1 \leq \frac{\gamma(\lambda_{\min}(V^{\top}H_{EE}V)) - 1}{\gamma(\lambda_{\min}(V^{\top}H_{EE}V)) + 1}.$$

4.3. Discussion. The assumptions of Theorems 4.10 and 4.11 usually hold for compressed sensing reconstruction problems posed as in (1.3) or (1.2), in which case A is often a Gaussian random matrix or has rows randomly chosen from an orthogonal matrix such as an FFT, DCT, or wavelets transform matrix. It is well known that a randomly generated matrix is full rank with probability one (unless elements of the matrix are generated from a restricted space) [23]. Therefore, when $A \in \mathbb{R}^{m \times n}$

is a random matrix, the reduced Hessian for problem (1.3), i.e., $A_E^\top M A_E$, where A_E consists of columns of A with indices in E , will have full rank with probability one as long as $|E| \leq m$, which is generally the case. A similar argument can be made for partial orthogonal matrices. We believe that the strict complementarity assumption in Theorem 4.11 should also hold for random matrices with a prevailing probability, though we do not currently have a proof for this. We have observed the general convergence behavior predicted by our theorems empirically in computational studies; see section 6.2.

In our convergence theorems, the choice of τ is restricted by the upper bound $2/\hat{\lambda}_{\max}$, where $\hat{\lambda}_{\max}$ is an upper bound for the largest eigenvalue of the Hessian. In compressed sensing applications, the quantity $\hat{\lambda}_{\max}$ is often easily obtained when $M = I$. When A is a partial orthogonal matrix, $\hat{\lambda}_{\max} = \lambda_{\max}(A^\top A) = 1$ and $\tau \in (0, 2)$ will suffice. When A is a Gaussian random matrix (with elements independently drawn from the standard normal distribution), well-known random matrix theory (see [34] or [23], for example) yields

$$n \left(1 - \sqrt{\frac{m}{n}}\right)^2 \leq \lambda_i(A^\top A) \leq n \left(1 + \sqrt{\frac{m}{n}}\right)^2$$

with prevailing probability for large n . In either case, upper bounding τ is not an issue as long as $M = I$.

For simplicity, we have used a fixed $\tau \in (0, 2/\hat{\lambda}_{\max})$ in our analysis. However, this requirement could be relaxed in the later stages of the iterations when the actions of the mapping $h = I - \tau g$ concentrate on a “reduced space.” In this stage, h can remain contractive even if the maximum eigenvalue bound on the Hessian is replaced by that on the reduced Hessian, which will generally increase the upper bound on τ . For example, consider the quadratic problem (1.2) where A is a partial orthogonal matrix. Then $\lambda_{\max}(A^\top A) = 1$, but $\lambda_{\max}(A_E^\top A_E) < 1$, such that h remains contractive even if τ is chosen close to $2/\lambda_{\max}(A_E^\top A_E) > 2$. Such a dynamic strategy, though theoretically feasible, is not straightforward to implement. It should be an interesting topic for further research.

5. Proofs of convergence results. In this section, Theorems 4.5 and 4.11 are proved through several technical results that lead to the final arguments.

5.1. Proof of Theorem 4.5. The lemma below establishes a sufficient condition for $x \in \Omega$ to be a fixed point of $s_\nu \circ h(\cdot)$.

LEMMA 5.1. *Under Assumption 1, if*

$$(5.1) \quad \|s_\nu \circ h(x) - s_\nu \circ h(x^*)\| \equiv \|s_\nu \circ h(x) - x^*\| = \|x - x^*\|,$$

then x is a fixed point, and therefore a solution of (1.1); that is,

$$(5.2) \quad x = s_\nu \circ h(x).$$

Proof. Recall that s_ν is component-wise nonexpansive and h is nonexpansive in $\|\cdot\|$. From (5.1),

$$(5.3) \quad \|x - x^*\| = \|s_\nu \circ h(x) - s_\nu \circ h(x^*)\| \leq \|h(x) - h(x^*)\| \leq \|x - x^*\|.$$

Hence, both inequalities hold as equalities. In particular, the no-shrinkage condition (3.8) holds for $y^1 = h(x)$ and $y^2 = h(x^*)$, so Part 1 of Lemma 3.2 yields

$$s_\nu \circ h(x) - s_\nu \circ h(x^*) = h(x) - h(x^*).$$

Rewriting this equation, we get

$$s_\nu \circ h(x) = x - \tau(g(x) - g(x^*)),$$

and since the last inequality in (5.3) also holds as equality, we have $g(x) - g(x^*) = 0$ according to Lemma 4.1, and hence the conclusion. \square

The next lemma establishes the finite convergence properties stated in Theorem 4.5.

LEMMA 5.2. *Let Assumption 1 hold and $\{x^k\}$ be generated by the fixed-point iterations (3.5) starting from any $x^0 \in \Omega$. Then*

1. $x_i^k = 0 \ \forall i \in L$ for all but at most $\|x^0 - x^*\|^2/\omega^2$ iterations;
2. $\text{sgn}(h_i(x^k)) = \text{sgn}(h_i(x^*)) = -\mu g_i^*$, $\forall i \in E$, for all but at most $\|x^0 - x^*\|^2/\nu^2$ iterations.

Proof. We fix any $x^* \in X^*$ and consider $x_i^k \neq 0$ for some $i \in L$. In view of the nonexpansiveness of $s_\nu(\cdot)$ and the related property in Lemma 3.2 part 6, we have

$$\begin{aligned} |x_i^{k+1} - x_i^*|^2 &= |s_\nu \circ h_i(x^k) - s_\nu \circ h_i(x^*)|^2 \\ &\leq (|h_i(x^k) - h_i(x^*)| - (\nu - h_i(x^*)))^2 \\ &\leq |h_i(x^k) - h_i(x^*)|^2 - \omega^2, \end{aligned}$$

where the last inequality follows from (4.11). The component-wise nonexpansiveness of $s_\nu(\cdot)$ and the nonexpansiveness of $h(\cdot)$ imply that

$$\|x^{k+1} - x^*\|^2 \leq \|h(x^k) - h(x^*)\|^2 - \omega^2 \leq \|x^k - x^*\|^2 - \omega^2.$$

Therefore, the number of iterations where $x_i^k \neq 0$ for some $i \in L$ cannot be more than $\|x^0 - x^*\|^2/\omega^2$. This proves the first statement.

For the second statement, we recall (3.1) and note that if $i \in \text{supp}(x^*)$

$$0 \neq x_i^* = \text{sgn}(h_i(x^*)) \max\{|h_i(x^*)| - \nu, 0\},$$

so that $|h_i(x^*)| > \nu$ for $i \in \text{supp}(x^*)$. On the other hand,

$$|h_i(x^*)| = \tau|g^*| = \tau/\mu = \nu, \ \forall i \in E \setminus \text{supp}(x^*).$$

Therefore,

$$|h_i(x^*)| \geq \nu, \ \forall i \in E.$$

Now if $\text{sgn}(h_i(x^k)) \neq \text{sgn}(h_i(x^*))$ for some $i \in E$, then Lemma 3.2, Part 5 implies

$$\begin{aligned} |x_i^{k+1} - x_i^*|^2 &= |s_\nu \circ h_i(x^k) - s_\nu \circ h_i(x^*)|^2 \\ &\leq (|h_i(x^k) - h_i(x^*)| - \nu)^2 \\ &\leq |h_i(x^k) - h_i(x^*)|^2 - \nu^2. \end{aligned}$$

Hence, the number of iterations for which $\text{sgn}(h_i(x^k)) \neq \text{sgn}(h_i(x^*))$ for some $i \in E$ cannot be more than $\|x^0 - x^*\|^2/\nu^2$. Moreover, it follows directly from the definitions of E in (4.8), h in (3.3), and g^* in (2.3), and the equation $\tau = \nu\mu$, that $\text{sgn}(h_i(x^*)) = -\mu g_i^*$ for all $i \in E$. \square

Based on these lemmas, we provide a short proof of Theorem 4.5 for the sake of completeness.

Proof of Theorem 4.5. To show that $\{x^k\}$ converges, we (i) show that $\{x^k\}$ has a limit point, (ii) argue that it must be a fixed point because it satisfies the condition (5.1) of Lemma 5.1, and (iii) prove its uniqueness.

Since $s_\nu \circ h(\cdot)$ is nonexpansive, $\{x^k\}$ lies in a compact subset of Ω and must have a limit point, say,

$$\bar{x} = \lim_{j \rightarrow \infty} x^{k_j}.$$

Since for any given fixed point x^* the sequence $\{\|x^k - x^*\|\}$ is monotonically nonincreasing, it has a limit which can be written as

$$(5.4) \quad \lim_{k \rightarrow \infty} \|x^k - x^*\| = \|\bar{x} - x^*\|,$$

where \bar{x} can be any limit point of $\{x^k\}$. That is, all limit points, if more than one exists, must have an equal distance to any given fixed point $x^* \in X^*$.

By the continuity of $s_\nu \circ h(\cdot)$, the image of \bar{x} ,

$$s_\nu \circ h(\bar{x}) = \lim_{j \rightarrow \infty} s_\nu \circ h(x^{k_j}) = \lim_{j \rightarrow \infty} x^{k_j+1},$$

is also a limit point of $\{x^k\}$. Therefore, from (5.4) we have

$$\|s_\nu \circ h(\bar{x}) - s_\nu \circ h(x^*)\| = \|\bar{x} - x^*\|,$$

which allows us to apply Lemma 5.1 to \bar{x} and establish the optimality of \bar{x} .

By setting $x^* = \bar{x} \in X^*$ in (5.4), we establish the convergence of $\{x^k\}$ to its unique limit point \bar{x} :

$$\lim_{k \rightarrow \infty} \|x^k - \bar{x}\| = 0.$$

Finally, the finite convergence results (4.12)–(4.13) were proved in Lemma 5.2. \square

5.2. Proof of Theorem 4.11. The next lemma gives a useful update formula for k sufficiently large and $i \in \text{supp}(x^*)$.

LEMMA 5.3. *Under Assumption 1, after a finite number of iterations*

$$(5.5) \quad x_i^{k+1} = x_i^k - \tau (g_i(x^k) - g_i^*), \quad \forall i \in \text{supp}(x^*).$$

Proof. Since $x^k \rightarrow x^* \in X^*$ and $h(\cdot)$ is component-wise continuous, $h_i(x^k) \rightarrow h_i(x^*)$. The fact that $|h_i(x^*)| > \nu$ for $i \in \text{supp}(x^*)$ implies that after a finite number of iterations we have $|h_i(x^k)| > \nu$ for $i \in \text{supp}(x^*)$. This gives

$$\begin{aligned} x_i^{k+1} &= \text{sgn}(h_i(x^k)) (|h_i(x^k)| - \nu) \\ &= h_i(x^k) - \nu \text{sgn}(h_i(x^k)) \\ &= x_i^k - \tau g_i(x^k) - (\tau/\mu) (-\mu g_i^*) \\ &= x_i^k - \tau (g_i(x^k) - g_i^*), \end{aligned}$$

for any $i \in \text{supp}(x^*)$. \square

Proof of Theorem 4.11. Without loss of generality, we can assume that k is large enough so that (5.5) holds and $x^k \in N^*$, where N^* is defined in Theorem 4.11.

Since $x_i^k = 0$ for any $i \in L$, it suffices to consider the rate of convergence of x_i^k for $i \in E = \text{supp}(x^*)$, where equality follows from the strict complementarity assumption on x^* .

Let \bar{H}^k be defined as in (4.19). By assumption, the range and null spaces of \bar{H}_{EE}^k are now invariant for all k . Let $P = VV^\top \in \mathbb{R}^{|E| \times |E|}$ be the orthogonal projection onto the range space of $H_{EE}(x^*)$ such that $I - P$ is the orthogonal projection onto the null space of $H_{EE}(x^*)$. Also recall that x_E denotes the subvector of x corresponding to the index set E .

Since $E = \text{supp}(x^*)$, Lemma 5.3 implies that

$$(5.6) \quad x_E^{k+1} = x_E^k - \tau (g(x^k) - g(x^*))_E = x_E^k - \tau \bar{H}_{EE}^k (x_E^k - x_E^*).$$

At each iteration, the update, $-\tau \bar{H}_{EE}^k (x_E^k - x_E^*)$, stays in the range space of $H_{EE}(x^*)$. This implies that the null space components of the iterates have converged to the null space components of x^* , namely, for all k sufficiently large,

$$(5.7) \quad (I - P) (x_E^k - x_E^*) \equiv \mathbf{0}.$$

If $H_{EE}(x^*) = 0$, then the range space is empty and the update vanishes such that $x^k = x^*$ after a finite number of steps.

Now assume that $H_{EE}(x^*) \neq 0$ so that $\lambda_{\min}^{\mathcal{R}} > 0$ exists. It suffices to consider the rate of convergence of $\{Px_E^k\}$ to Px_E^* . It follows from (5.6) and (5.7) that

$$(5.8) \quad x_E^{k+1} - x_E^* = P (x_E^{k+1} - x_E^*) = P (I - \tau \bar{H}_{EE}^k) P (x_E^k - x_E^*).$$

By a routine continuity argument, we know that there exists an arbitrarily small constant $\epsilon > 0$ such that for all k sufficiently large the eigenvalues of $V^\top H_{EE}^k V$ satisfy

$$\hat{\lambda}_{\max} \geq \lambda_i (V^\top H_{EE}^k V) \geq \lambda_{\min}^{\mathcal{R}} - \epsilon > 0, \quad \forall i.$$

Consequently, given the definition of τ in (4.18) and noting that $P^2 = P = VV^\top$, we calculate from (5.8):

$$(5.9) \quad \begin{aligned} \|x_E^{k+1} - x_E^*\| &\leq \|P (I - \tau \bar{H}_{EE}^k) P\| \|x_E^k - x_E^*\| \\ &= \|I - \tau V^\top \bar{H}_{EE}^k V\| \|x_E^k - x_E^*\| \\ &= \max \left\{ |1 - \tau \hat{\lambda}_{\max}|, |1 - \tau \lambda_{\min}^{\mathcal{R}}| + \tau \epsilon \right\} \|x_E^k - x_E^*\| \\ &= \left(\frac{\gamma (\lambda_{\min}^{\mathcal{R}}) - 1}{\gamma (\lambda_{\min}^{\mathcal{R}}) + 1} + \tau \epsilon \right) \|x_E^k - x_E^*\|, \end{aligned}$$

which implies (4.24) since ϵ can be arbitrarily small. \square

6. A continuation method. Our algorithm for solving (1.1), that is,

$$(6.1) \quad \min_{x \in \mathbb{R}^n} \|x\|_1 + \mu f(x),$$

consists of applying the fixed-point iterations

$$x^{k+1} = s_\nu \circ h(x^k) := \text{sgn}(x^k - \tau g(x^k)) \odot \max\{|x^k - \tau g(x^k)| - \nu, 0\}, \quad \mu\nu = \tau$$

(see (3.5) and (4.3)) within the continuation (or path-following) framework described below. Further extensions that may improve our algorithm are certainly possible, but are beyond the scope of this paper.

6.1. Homotopy algorithms in statistics. Statisticians often solve (1.2) (which is (1.1) with $f(x) = \frac{1}{2}\|Ax - b\|^2$) in the context of regression. In Bayesian terminology, this corresponds to maximizing the *a posteriori* probability for recovering the signal x from the measurement $b = Ax + \epsilon$, where the prior on x is Laplacian and ϵ is Gaussian white noise. Practically, such a procedure may be preferred over standard least squares because a sparse solution of (1.2) explicitly identifies the most significant regressor variables.

As intimated in the Introduction, variations on (1.2) may be used in different applications and contexts. For example, problem (1.2) is closely related to this quadratically constrained ℓ_1 -minimization problem

$$(6.2) \quad \min_x \{ \|x\|_1 \mid \|Ax - b\|^2 \leq \sigma^2 \chi_{1-\alpha, m}^2 \},$$

which is often used when an estimated noise level σ is available. Alternatively, one can constrain the size of $\|x\|_1$ and minimize the sum of squares of the residual $Ax - b$:

$$(6.3) \quad \min_x \left\{ \frac{1}{2} \|Ax - b\|^2 \mid \|x\|_1 \leq t \right\}.$$

Statisticians often refer to the above problem as the Least Absolute Shrinkage and Selection Operator (LASSO) [58].

Problems (1.2), (6.2), and (6.3) are equivalent in the sense that once the value of one of μ , σ , or t is fixed, there are values for the other two quantities such that all three problems have the same solution. For a detailed explanation, please see [53].

Least Angle Regression (LARS) (see [24], for example) is a method for solving (6.3). LARS starts with the zero vector and gradually increases the number of nonzeros in the approximation x . In fact, it generates the full path of solutions that results from setting the right-hand side of the constraint to every value in the interval $[0, t]$. Thus, LARS is a homotopy algorithm. The construction of the path of solutions is facilitated by the fact that it is piecewise linear, such that any segment can be generated given the solutions at turning points, which are the points at which at least one component changes from zero to nonzero or vice versa. Thus LARS and other homotopy algorithms [43, 49, 63] solve (6.3) by computing the solutions at the turning points encountered as t increases from 0 to a given value. These algorithms require the solution of a least squares problem at every iteration, where the derivative matrix of the residuals consists of the columns of A associated with the nonzero components of the current iterate. For large-scale problems, solving these intermediate least squares problems may prove costly, especially when the solution is only moderately sparse, and/or A is a partial fast transform matrix that is not stored explicitly.

We found it helpful for our algorithm to adopt a continuation strategy similar to homotopy in the sense that we solve (1.1) for an increasing sequence of μ values. However, our algorithm does not track turning points or solve any least squares subproblems, and so only approximately follows the solution path.

6.2. A continuation strategy. The convergence analysis indicates that the speed of the fixed-point algorithm is determined by the values of $\nu = \tau/\mu$ and ω (see Theorem 4.5), and the spectral properties of the Hessian of $f(x)$ (see Theorems 4.10 and 4.11). The signs of $h_i(x^k)$ evolve to agree with those of $h_i(x^*)$ for $i \in E$ faster for larger ν (equivalently, for smaller μ). Similarly, large ω implies fast convergence of the $|x_i^k|$, $i \in L$, to zero. Once the finite convergence properties of Theorem 4.5 are satisfied, all action is directed towards reducing the errors in the E components, and the (worst-

case) convergence rate is dictated by $\|I - \tau \bar{H}_{EE}\|$, which can be considerably smaller than $\|I - \tau \bar{H}\|$, especially when $|E| \ll n$.

In general, we have little or no control over the value of ω , nor the spectral properties of the Hessian. On the other hand, we do have the freedom to choose τ and $\nu = \tau/\mu$. For fixed τ we found $\tau \in [1/\hat{\lambda}_{\max}, 2/\hat{\lambda}_{\max})$ to be superior to $\tau \in (0, 1/\hat{\lambda}_{\max})$. Beyond this, τ does not have much effect on ν and can be chosen empirically or based on considerations concerning $\|I - \tau \bar{H}_{EE}\|$. The value of μ , on the other hand, while it must eventually be equal to some specified value $\bar{\mu}$, can in the meantime be chosen freely to produce a wide range of ν values. Thus, since larger ν means faster convergence, we propose a continuation strategy for μ . In particular, if problem (6.1) is to be solved with $\bar{\mu}$, we propose solving a sequence of problems (6.1) defined by an increasing sequence $\{\mu_j\}$, as opposed to fixing $\nu = \tau/\bar{\mu}$. When a new problem, associated with μ_{j+1} , is to be solved, the approximate solution for the current (μ_j) problem is used as the starting point. In essence, this framework approximately follows the path $x^*(\mu)$ in the interval $[\mu_1, \bar{\mu}]$, where for any given μ value $x^*(\mu)$ is an optimal solution for (6.1). This path is well defined if the solution to (6.1) is unique for $\mu \in [\mu_1, \bar{\mu}]$. Even if this is not the case, it is reassuring to observe that the algorithm itself is well-defined. A formal statement of our fixed-point continuation method is given in Algorithm 1.

ALGORITHM 1 Fixed-point Continuation (FPC) Algorithm

Require: A , b , x^0 , and $\bar{\mu}$

- 1: Select $0 < \mu_1 < \mu_2 < \dots < \mu_L = \bar{\mu}$. Set $x = x^0$.
 - 2: **for** $\mu = \mu_1, \mu_2, \dots, \mu_L$ **do**
 - 3: **while** “not converged” **do**
 - 4: Select τ and set $\nu = \tau/\mu$
 - 5: $x \leftarrow s_\nu \circ h(x)$
 - 6: **end while**
 - 7: **end for**
-

Our computational experience indicates that the performance of this continuation strategy can be far superior to that of directly applying the fixed-point iterations (3.5) with a specified value $\bar{\mu}$. This is evident in Figure 6.1, where the convergence behavior

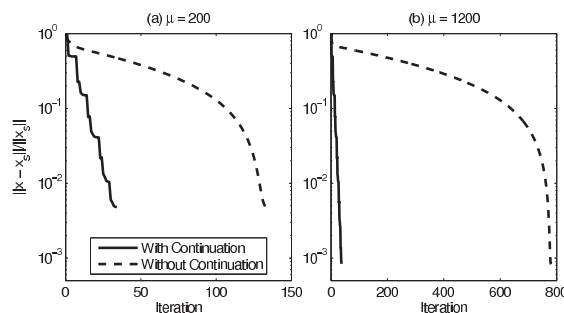


FIG. 6.1. Convergence acceleration via continuation. The relative error, $\|x^k - x^*\|/\|x^*\|$, and iteration data were obtained by applying the FPC algorithm with or without continuation to two instances of problem (1.2), where in each instance A is a 512×1024 partial DCT matrix, $b = Ax^*$ plus noise for a given sparse x^* , and $\mu = 200$ in the first case and 1200 in the second. The plots show that as μ increases, the advantages of continuation become more pronounced.

of the two approaches (with and without continuation) is plotted for two values of $\bar{\mu}$, and is in line with the observations of [18, 43, 49, 58]. Moreover, since $x^*(\mu)$ tends to be sparser for smaller μ , the reduced Hessian \bar{H}_{EE} tends to be smaller and better conditioned in this case, such that the continuation strategy should improve the convergence rate for the components with indices in E in addition to the rate of finite convergence of $\text{sgn}(h_i(x^k))$, $i \in E$. Overall, this fixed-point continuation algorithm produces competitive solution times (as compared to other state-of-the-art algorithms) for compressed sensing problems. For instance, we were able to reconstruct signals of length 2,097,152 in 2.5 to 7.5 minutes, depending on the level of sparsity and noise, when A was a $1,048,576 \times 2,097,152$ partial DCT matrix (using Matlab 7.3 on a Dell Optiplex GX620 with a 3.2 GHz processor and 4 GB RAM).

In principle, our fixed-point continuation algorithm can be used to solve problems (6.2) and (6.3) in addition to (6.1). Take the LASSO problem (6.3) as an example. When we start our algorithm with a small μ value, the corresponding optimal $\|x\|_1$ is also small; subsequent increases in μ correspond to increases in the optimal $\|x\|_1$. We can stop the process once $\|x\|_1$ approximately equals t , backtracking if necessary. As interesting as such extensions may be, they are not in the scope of the current paper. Indeed, as we observed in our computational study [32], a strength of this algorithmic framework is that a simple implementation is sufficient to obtain good results.

7. Conclusions. We investigated the use of the forward-backward operator splitting technique, combined with a continuation (path-following) strategy, for solving ℓ_1 -norm regularized convex optimization problems. Our theoretical analysis yields convergence results stronger than what could be obtained from applying existing general theory to our setting. In particular, we established finite convergence for some quantities and q -linear convergence rates without assuming strict convexity. Interestingly, our rate of convergence results imply, in a general sense, that sparser solutions correspond to faster rates of convergence, which agrees with what has been observed in practice. Our convergence analysis, however, is only for the fixed-point algorithm (3.5) with a fixed μ value. It remains an important, yet more difficult, research issue to study convergence behavior associated with specific continuation strategies.

We have conducted a comprehensive computational study to compare our fixed-point continuation (FPC) algorithm with three recent state-of-the-art compressed sensing recovery algorithms [17, 28, 35]. The numerical results, too lengthy to be included in the present paper, will be reported in a subsequent paper [32]. In brief, these numerical results indicate that FPC's overall performance is competitive with, and is often superior to, these state-of-the-art algorithms. The strong performance of FPC in computing sparse solutions to compressed sensing problems is certainly encouraging. However, it remains a research issue to carefully evaluate and possibly enhance the performance of FPC on other ℓ_1 -regularized optimization problems where solutions are not necessarily very sparse.

REFERENCES

- [1] D. BARON, M. WAKIN, M. DUARTE, S. SARVOTHAM, AND R. BARANIUK, *Distributed compressed sensing*, ECE Department, Rice University, preprint, 2005; also available online from <http://www.dsp.ece.rice.edu/cs/DCS112005.pdf>.
- [2] J. BECT, L. BLANC-FERAUD, G. AUBERT, AND A. CHAMBOLLE, *A ℓ_1 -unified variational framework for image restoration*, European Conference on Computer Vision, Prague, Lecture Notes in Computer Sciences 3024, 2004, pp. 1–13.

- [3] J. BIOUCAS-DIAS AND M. FIGUEIREDO, *Two-step algorithms for linear inverse problems with non-quadratic regularization*, IEEE International Conference on Image Processing C ICIP' 2007, San Antonio, TX, 2007.
- [4] E. CANDÈS, J. ROMBERG, AND T. TAO, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.
- [5] E. CANDÈS AND J. ROMBERG, *Quantitative robust uncertainty principles and optimally sparse decompositions*, Found. Comput. Math., 6 (2006), pp. 227–254.
- [6] E. CANDÈS AND T. TAO, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.
- [7] A. CHAMBOLLE, R. A. DEVORE, N.-Y. LEE, AND B. J. LUCIER, *Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage*, IEEE Trans. Image Process., 7 (1998), pp. 319–335.
- [8] H.-G. CHEN AND R. T. ROCKAFELLAR, *Convergence rates in forward-backward splitting*, SIAM J. Optim., 7 (1997), pp. 421–444.
- [9] S. CHEN, D. DONOHO, AND M. A. SAUNDERS, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 20 (1998), pp. 33–61.
- [10] J. CLAERBOUT AND F. MUIR, *Robust modelling of erratic data*, Geophys., 38 (1973), pp. 826–844.
- [11] P. L. COMBETTES AND J.-C. PESQUET, *Proximal thresholding algorithm for minimization over orthonormal bases*, SIAM J. Optim., 18 (2007), pp. 1351–1376.
- [12] P. L. COMBETTES AND V. R. WAJS, *Signal recovery by proximal forward-backward splitting*, SIAM J. Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
- [13] I. DAUBECHIES, M. DEFRISE, AND C. DE MOL, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure Appl. Math., 57 (2004), pp. 1413–1457.
- [14] I. DAUBECHIES, M. FORNASIER, AND I. LORIS, *Accelerated projected gradient method for linear inverse problems with sparsity constraints*, arXiv:0706:4297, 2007.
- [15] C. DE MOL AND M. DEFRISE, *A note on wavelet-based inversion algorithms*, Contemp. Math., 313 (2002), pp. 85–96.
- [16] D. DONOHO AND J. TANNER, *Neighborliness of randomly-projected simplices in high dimensions*, Proc. Nat. Acad. Sci., 102 (2005), pp. 9452–9457.
- [17] D. DONOHO, Y. TSAIG, I. DRORI, AND J.-C. STARCK, *Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit*, Technical report 2006-02, Department of Statistics, Stanford University, Stanford, CA, 2006.
- [18] D. DONOHO AND Y. TSAIG, *Fast solutions of ℓ_1 -norm minimization problems when the solution may be sparse*, Technical report online, 2006.
- [19] D. DONOHO, *De-noising by soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.
- [20] D. DONOHO, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
- [21] J. DOUGLAS AND H. H. RACHFORD, *On the numerical solution of the heat conduction problem in 2 and 3 space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.
- [22] J. ECKSTEIN, *Splitting methods for monotone operators with applications to parallel optimization*, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.
- [23] A. EDELMAN, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560.
- [24] B. EFRON, T. HASTIE, I. JOHNSTONE, AND R. TIBSHIRANI, *Least angle regression*, Ann. Stat., 32 (2004), pp. 407–499.
- [25] M. ELAD, B. MATALON, J. SHTOK, AND M. ZIBULEVSKY, *A wide-angle view at iterated shrinkage algorithms*, SPIE (Wavelet XII), San Diego, CA, August 26–29, 2007.
- [26] M. ELAD, B. MATALON, AND M. ZIBULEVSKY, *Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization*, J. Appl. Comput. Harmonic Analysis, 23 (2006), pp. 346–367.
- [27] M. ELAD, *Why simple shrinkage is still relevant for redundant representations?*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5559–5569.
- [28] M. A. T. FIGUEIREDO, R. D. NOWAK, AND S. J. WRIGHT, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, reprint, 2007.
- [29] M. FIGUEIREDO AND R. NOWAK, *An EM algorithm for wavelet-based image restoration*, IEEE Trans. Image Process., 12 (2003), pp. 906–916.
- [30] D. GABAY, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983.

- [31] R. GLOWINSKI AND P. LE TALLEC, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, SIAM, Philadelphia, 1989.
- [32] E. HALE, W. YIN, AND Y. ZHANG, *A numerical study on fixed point continuation method applied to compressed sensing*, Rice University CAAM Technical report TR08-24, Rice University, Houston, TX, 2008, submitted.
- [33] S. HAUBRUGE, V. H. NGUYEN, AND J. J. STRODIOT, *Convergence analysis and applications of the Glowinski-Le Tallec splitting method for finding a zero of the sum of two maximal monotone operators*, *J. Optim. Theory Appl.*, 97 (1998), pp. 645–673.
- [34] D. JONSSON, *Some limit theorems for the eigenvalues of a sample covariance matrix*, *J. Multivariate Analysis*, 12 (1982), pp. 1–38.
- [35] S.-J. KIM, K. KOH, M. LUSTIG, S. BOYD, AND D. GORINEVSKY, *A method for large-scale l_1 -regularized least squares*, *IEEE J. Selected Topics Signal Process.*, 1 (2007), pp. 606–617.
- [36] S. KIROLOS, J. LASKA, M. WAKIN, M. DUARTE, D. BARON, T. RAGHEB, Y. MASSOUD, AND R. BARANIUK, *Analog-to-information conversion via random demodulation*, in *Proceedings of the IEEE Dallas Circuits and Systems Workshop (DCAS)*, Dallas, TX, 2006.
- [37] J. LASKA, S. KIROLOS, M. DUARTE, T. RAGHEB, R. BARANIUK, AND Y. MASSOUD, *Theory and implementaion of an analog-to information converter using random demodulation*, in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, New Orleans, LA, 2007.
- [38] J. LASKA, S. KIROLOS, Y. MASSOUD, R. BARANIUK, A. GILBERT, M. IWEN, AND M. STRAUSS, *Random sampling for analog-to-information conversion of wideband signals*, in *Proceedings of the IEEE Dallas Circuits and Systems Workshop*, Dallas, TX, 2006.
- [39] S. LEVY AND P. FULLAGAR, *Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution*, *Geophys.*, 46 (1981), pp. 1235–1243.
- [40] P. L. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two nonlinear operators*, *SIAM J. Numer. Anal.*, 16 (1979), pp. 964–979.
- [41] Z.-Q. LUO AND P. TSENG, *On the linear convergence of descent methods for convex essentially smooth minimization*, *SIAM J. Control Optim.*, 30 (1990), pp. 408–425.
- [42] M. LUSTIG, D. DONOHO, AND J. PAULY, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, *Magnetic Resonance in Medicine*, 58 (2007), pp. 1182–1195.
- [43] D. MALIOUTOV, M. ÇETIN, AND A. WILLSKY, *Homotopy continuation for sparse signal representation*, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5, Philadelphia, PA, 2005, pp. 733–736.
- [44] S. G. MALLAT AND Z. ZHANG, *Matching pursuits with time-frequency dictionaries*, *IEEE Trans. Signal Process.*, 41 (1993), pp. 3397–3415.
- [45] B. MERCIER, *Inéquations Variationnelles de la Mécanique*, Publications Mathématiques d’Orsay, Université de Paris-Sud, Orsay, France, 80.01 (1980).
- [46] A. MILLER, *Subset Selection in Regression*, Chapman and Hall, London, 2002.
- [47] Y. NESTEROV, *Gradient methods for minimizing composite objective function*, www.optimization-online.org, CORE Discussion Paper 2007/76, 2007.
- [48] M. A. NOOR, *Splitting methods for pseudomonotone mixed variational inequalities*, *J. Math. Anal. Appl.*, 246 (2000), pp. 174–188.
- [49] M. OSBORNE, B. PRESNELL, AND B. TURLACH, *A new approach to variable selection in least squares problems*, *IMA J. Numer. Anal.*, 20 (2000), pp. 389–403.
- [50] J.-S. PANG, *A posteriori error bounds for the linearly-constrained variational inequality problem*, *Math. Methods Oper. Res.*, 12 (1987), pp. 474–484.
- [51] G. B. PASSTY, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, *J. Math. Anal. Appl.*, 72 (1979), pp. 383–390.
- [52] D. H. PEACEMAN AND H. H. RACHFORD, *The numerical solution of parabolic elliptic differential equations*, *SIAM J. Appl. Math.*, 3 (1955), pp. 28–41.
- [53] R. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [54] M. RUDELSON AND R. VERSHYNIN, *Geometric approach to error-correcting codes and reconstruction of signals*, *Int. Math. Res. Not.*, (2005), pp. 4019–4041.
- [55] F. SANTOSA AND W. SYMES, *Linear inversion of band-limited reflection histograms*, *SIAM J. Sci. Stat. Comput.*, 7 (1986), pp. 1307–1330.
- [56] D. TAKHAR, J. LASKA, M. WAKIN, M. DUARTE, D. BARON, S. SARVOTHAM, K. KELLY, AND R. BARANIUK, *A new compressive imaging camera architecture using optical-domain compression*, in *Proceedings of Computational Imaging IV at SPIE Electronic Image*, San Jose, CA, 2006.
- [57] H. TAYLOR, S. BANK, AND J. MCCOY, *Deconvolution with the l_1 norm*, *Geophys.*, 44 (1979), pp. 39–52.
- [58] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, *J. Royal Statist. Soc. B*, 58 (1996), pp. 267–288.

- [59] J. TROPP, M. WAKIN, M. DUARTE, D. BARON, AND R. BARANIUK, *Random filters for compressive sampling and reconstruction*, in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, 2006.
- [60] J. TROPP, *Just relax: Convex programming methods for identifying sparse signals*, IEEE Trans. Inform. Theory, 51 (2006), pp. 1030–1051.
- [61] Y. TSAIG AND D. DONOHO, *Extensions of compressed sensing*, Signal Processing, 86 (2005), pp. 533–548.
- [62] P. TSENG, *A modified forward-backward splitting method for maximal monotone mappings*, SIAM J. Control Optim., 38 (2000), pp. 431–446.
- [63] B. TURLACH, *On algorithms for solving least squares problems under an L_1 penalty or an L_1 constraint*, in Proceedings of the American Statistical Association; Statistical Computing Section, Alexandria, VA, 2005, pp. 2572–2577.
- [64] E. VAN DEN BERG AND M. P. FRIEDLANDER, *In pursuit of a root*, UBC Computer Science Technical Report TR-2007-16, 2007.
- [65] M. WAKIN, J. LASKA, M. DUARTE, D. BARON, S. SARVOTHAM, D. TAKHAR, K. KELLY, AND R. BARANIUK, *An architecture for compressing image*, in Proceedings of the International Conference on Image Processing (ICIP), Atlanta, Georgia, 2006.
- [66] M. WAKIN, J. LASKA, M. DUARTE, D. BARON, S. SARVOTHAM, D. TAKHAR, K. KELLY, AND R. BARANIUK, *Compressive imaging for video representation and coding*, in Proceedings of Picture Coding Symposium (PCS), Beijing, China, 2006.
- [67] Y. ZHANG, *When is missing data recoverable?*, Rice University CAAM Technical Report TR06-15, 2006.