Compressed Sensing Recovery via Nonconvex Shrinkage Penalties

Joseph Woodworth, Rick Chartrand *Senior Member, IEEE*

Abstract

The ℓ^0 minimization of compressed sensing is often relaxed to ℓ^1, which yields easy computation using the shrinkage mapping known as soft thresholding, and can be shown to recover the original solution under certain hypotheses. Recent work has derived a general class of shrinkages and associated nonconvex penalties that better approximate the original ℓ^0 penalty and empirically can recover the original solution from fewer measurements. We specifically examine *p-shrinkage* and *firm thresholding*. In this work, we prove that given data and a measurement matrix from a broad class of matrices, one can choose parameters for these classes of shrinkages to guarantee exact recovery of the sparsest solution. We further prove convergence of the algorithm *iterative p-shrinkage* (IPS) for solving one such relaxed problem.

Index Terms

compressed sensing, nonconvexity, relaxation, exact recovery, stability, convergence

I. Introduction

Compressed sensing has been successfully applied in a multitude of scientific fields, ranging from image processing tasks to radar to coding theory, making the potential impact of advancements in theory and practice rather large. Compressed sensing methods rely on the notion of sparsity, which is primarily approximated via the ℓ^1 norm [1], [2]. The nature and limitations

J. Woodworth is with the Department of Mathematics, University of California, Los Angeles, CA, 90095, USA; e-mail: jwoodworth@math.ucla.edu

R. Chartrand is with the Theoretical Division, MS B284, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA; e-mail: rickc@lanl.gov

Manuscript received ????, 2014; revised ????.

of this relaxation have been well-studied [3]–[11], as well as some alternative relaxations, such as the ℓ^p quasinorm [5], [10]–[20]. The nonconvex ℓ^p quasinorm approaches present a tradeoff: closer approximation of sparsity for harder analysis and computation. Recent work has introduced generalized nonconvex penalties [21]–[27] that have thus far demonstrated strong empirical performance [21], [23], [25], [28]. In this paper, we prove conditions that guarantee good performance of these generalized penalties.

A. Compressed Sensing

Compressed sensing seeks to represent a signal from a small number of linear measurements. We let the vector $x \in \mathbb{R}^n$ represent the original signal. The linear measurements are the result of an application of the short and fat measurement matrix $A \in \mathbb{R}^{m \times n}$, with $m \ll n$. One is given the measurements $b := Ax$ and wants to recover x. Of course $m \ll n$ implies that $Ax = b$ is an underdetermined linear system in x, so additional assumptions must be made about x. Thus one assumes that x is *sparse*, meaning that it has few nonzero entries. By considering the standard definition of p norms for vectors,

$$\|w\|_p^p := \sum_i |w_i|^p, \tag{1}$$

and taking the limit as p approaches 0 from above, we get the ℓ^0 penalty, $\|w\|_0$, which counts the number of nonzero entries of w. One would like to find the sparsest vector $w \in \mathbb{R}^n$ whose measurements are b, which suggests the following optimization problem:

$$\min_w \|w\|_0 \text{ subject to } Aw = b. \tag{2}$$

Unfortunately, this problem is known to be NP-hard (Non-deterministic Polynomial-time hard) in general [29, Sec. 9.2.2]. In other words, without making further assumptions on A and x, an algorithm solving this problem would be computationally intractable. For this reason, one relaxes the problem, replacing the ℓ^0 penalty with other penalties.

B. ℓ^1 relaxation

The ℓ^1 relaxed version of the compressed sensing problem is as follows:

$$\min_w \|w\|_1 \text{ subject to } Aw = b. \tag{3}$$

In contrast to the combinatorial ℓ^0 problem, this problem minimizes a convex energy subject to linear constraints, and can be recast as a linear program. Extensive theory has been developed to study the properties of solutions to convex problems [30]. Further, a subproblem related to the ℓ^1 relaxation of compressed sensing has a closed-form solution, given by an application of a shrinkage operator:

Definition I.1. Soft thresholding is given by the following formula:

$$S_{\lambda,1}(x)_i = s_{\lambda,1}(|x_i|)\,\mathrm{sign}(x_i) = \max\{|x_i| - \lambda, 0\}\,\mathrm{sign}(x_i). \tag{4}$$

The role soft thresholding plays is as the *proximal mapping* of the ℓ^1 norm:

$$S_{\lambda,1}(x) = \mathrm{prox}_\lambda \|\cdot\|_1(x) := \arg\min_w \lambda\|w\|_1 + \tfrac{1}{2}\|w - x\|_2^2. \tag{5}$$

Several algorithms for compressed sensing make use of this proximal mapping, such as iterative soft thresholding [31], alternating direction method of multipliers (ADMM) [32]–[35], and the Chambolle-Pock algorithm [36]. The explicit formula for (5) makes the use of ℓ^1 regularization particularly convenient.

All of this suggests why the ℓ^1 relaxation of compressed sensing is nice to solve, but does not motivate it as the right problem to solve. In particular, one is interested in conditions under which the solution to the ℓ^1 relaxation (3) of compressed sensing equals or approximately equals the solution of the original ℓ^0 compressed sensing problem (2). The papers [1], [2] developed theory for the recovery of the ℓ^0 solution by the ℓ^1 problem. In the years the followed, getting looser conditions for exact ℓ^1 recovery received continuing interest [3]–[11], [16]. One type of condition for recovery of the ℓ^0 solution from the ℓ^1 problem relies on the *restricted isometry constants* associated with the measurement matrix A. The restricted isometry constant of order k associated with the matrix $A \in \mathbb{R}^{m\times n}$ is the smallest $\delta_k \geq 0$ such that the following holds for all $x \in \mathbb{R}^n$ with $\|x\|_0 \leq k$ [37]:

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2. \tag{6}$$

Note that when $\delta_k > 1$ the lower bound becomes trivial and the upper bound can be improved by rescaling A. Thus any measurement matrix, with appropriate rescaling, can achieve $\delta_k = 1$, so one typically only regards $\delta_k \in [0, 1)$. One of the best current ℓ^1 recovery results states that for sufficiently large n, a sparse vector $x \in \mathbb{R}^n$ with $\|x\|_0 = k$ can be recovered by ℓ^1 minimization

as long as $k < m/2$ and the restricted isometry constant of order $2k$ associated with A satisfies $\delta_{2k} \leq 1/2$ [4].

C. ℓ^p relaxations ($0 < p < 1$)

A similar relaxation of the ℓ^0 problem that achieves recovery results in broader cases is ℓ^p minimization for $0 < p < 1$. In contrast to the ℓ^1 norm, the ℓ^p quasinorms for $0 < p < 1$ are not convex. Hence much of the theory of convex analysis no longer applies, making solution uniqueness and convergence results more complicated. However, the loss of convexity comes with the benefit that ℓ^p is better able to approximate the original ℓ^0 than ℓ^1 can. As a result, one can show that for any given measurement matrix with restricted isometry constant $\delta_{2k} < 1$, there exists some $p \in (0, 1)$ that will guarantee exact recovery of signals with support smaller than $k < m/2$ by the ℓ^p minimization problem [13]. It has also been demonstrated empirically that ℓ^p minimization gives better sparse recovery results than ℓ^1 minimization [38]–[40], with improved robustness [14], [18], [19].

Consider the proximal mapping of the ℓ^p quasinorm (to the p^{th} power, for simplicity), that is,

$$\operatorname{prox}_\lambda \|\cdot\|_p^p(x) := \arg\min_w \lambda\|w\|_p^p + \tfrac{1}{2}\|w - x\|_2^2. \tag{7}$$

Unfortunately, (7) is a discontinuous mapping [41], and there is no closed-form expression for (7) for general p. (The expression given in [42] is incorrect. For the special cases of $p = 1/2$ or $2/3$, the proximal mapping can be expressed in terms of the solution of a cubic or quartic equation, explicitly but cumbersomely.) This prevents several efficient algorithms from being generalized from ℓ^1 to ℓ^p minimization.

D. Generalized shrinkage

The need for an explicit proximal mapping motivates the approach of specifying a shrinkage mapping, and minimizing an implicitly-defined penalty function whose proximal mapping is the specified shrinkage [21]–[23], [27]. In this work, we extend theoretical results for recovery of sparse signals to the case of penalty functions induced by two families of shrinkages, p-shrinkage and firm thresholding (see Defs. II.1, II.2 below). In Section II we describe these shrinkage mappings, and how they are the proximal mappings of nonconvex penalty functions. In Section III we prove conditions for the exact recovery of sparse signals via minimizing such

nonconvex penalty functions. In Section IV we demonstrate the stability of signal recovery to noisy measurements and approximately sparse signals, and in Section V we show the algorithmic convergence of *iterative p-shrinkage* (IPS).

II. Generalized shrinkage penalties

As described above, nonconvex penalty functions have been shown both theoretically and empirically to give better results for compressed sensing than the ℓ^1 norm. In order to make use of any of several efficient algorithms, we wish to consider penalty functions with explicit proximal mappings. In this section, we consider two such families of functions.

A. p-shrinkage and firm thresholding

First we consider a shrinkage mapping, a version of which first appeared in [21], that has some qualitative resemblance to the ℓ^p proximal mapping, while being continuous and explicit:

Definition II.1. For $\lambda > 0$, the p-shrinkage mapping $S_p = S_{\lambda,p}$ for $p \in \mathbb{R}$ is defined by $S_p(x)_i = s_p(|x_i|)\,\mathrm{sign}(x_i)$, where the shrinkage function $s_p = s_{\lambda,p}$ is defined by

$$s_p(t) = \max\{t - \lambda^{2-p} t^{p-1}, 0\}. \tag{8}$$

See Fig. 1 for example plots. When $p = 1$, p-shrinkage and soft thresholding coincide. The smaller the value of p, the less p-shrinkage shrinks large inputs. In the limit as $p \to -\infty$, p-shrinkage tends pointwise to *hard thresholding*:

Definition II.2. For $\lambda > 0$, the hard thresholding mapping H_λ is defined by

$$H_\lambda(x)_i = \begin{cases} 0 & \text{if } |x_i| \le \lambda, \\ x_i & \text{if } |x_i| > \lambda. \end{cases} \tag{9}$$

Hard thresholding is related to the proximal mapping of the ℓ^0 penalty function:

$$H_{\sqrt{2\lambda}} \in \mathrm{prox}_\lambda \|\cdot\|_0, \tag{10}$$

the right side of (10) being two-valued in components satisfying $x_i^2 = 2\lambda$. Hard thresholding imposes no bias on large inputs, but its discontinuity makes it very unstable when used with ADMM [43].

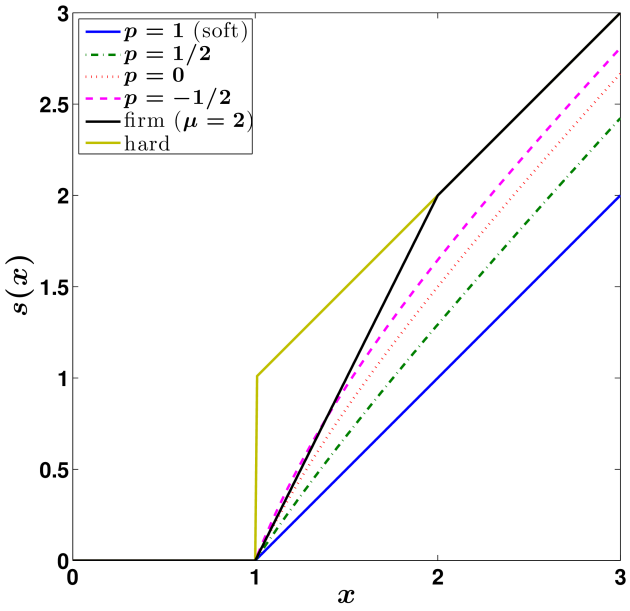


Fig. 1. Plot of several shrinkage functions, all with $\lambda = 1$. The smaller the value of p, the smaller the bias applied to large inputs. Firm thresholding removes the bias completely for large enough inputs, without the discontinuity of hard thresholding.

Another shrinkage mapping we consider is *firm thresholding*, a continuous, piecewise-linear approximation of hard thresholding. Firm thresholding was first introduced in [44] in connection with the WaveShrink procedure for denoising and non-parametric regression. It was not known at the time to be the proximal operator of a given penalty function.

Definition II.3. For $\lambda > 0$ and $\mu > \lambda$, the firm thresholding mapping $S_{\text{firm}} = S_{\lambda,\mu,\text{firm}}$ is defined by $S_{\text{firm}}(x)_i = s_{\text{firm}}(|x_i|)$, where $s_{\text{firm}} = s_{\lambda,\mu,\text{firm}}$ is defined by

$$s_{\text{firm}}(t) = \begin{cases} 0 & \text{if } t \leq \lambda, \\ \frac{\mu}{\mu-\lambda}(t-\lambda) & \text{if } \lambda \leq t \leq \mu, \\ t & \text{if } t \geq \mu. \end{cases} \tag{11}$$

Note that $S_{\lambda,\lambda,\text{firm}} = H_\lambda$, and $\lim_{\mu\to\infty} S_{\lambda,\mu,\text{firm}}(x) = S_{\lambda,1}(x)$ pointwise. Thus both p-shrinkage and firm thresholding can be seen as generalizing both soft and hard thresholding.

B. Shrinkage-induced penalty functions

Our motivation for considering alternative shrinkage mappings is to have them as closed-form proximal mappings. This requires that the shrinkages actually be the proximal mappings

of penalty functions. The following theorem guarantees this. It is proved in [23, Thm. 1], and strengthens the earlier result of Antoniadis [27, Prop. 3.2].

Theorem II.4. *Suppose $s : [0, \infty) \to \mathbb{R}$ is continuous, satisfies $x \leq \lambda \Rightarrow s(x) = 0$ for some $\lambda \geq 0$, is strictly increasing on $[\lambda, \infty)$, and $s(x) \leq x$. Define $S(x)_i = s(|x_i|)\operatorname{sign}(x_i)$, for each i. Then S is the proximal mapping of a penalty function $G(w) = \sum_i g(w_i)$ where g is even, strictly increasing and continuous on $[0, \infty)$, differentiable on $(0, \infty)$, and nondifferentiable at 0 iff $\lambda > 0$ (in which case $\partial g(0) = [-1, 1]$). If also $x - s(x)$ is nonincreasing on $[\lambda, \infty)$, then g is concave on $[0, \infty)$ and G satisfies the triangle inequality.*

Both p-shrinkage and firm thresholding satisfy all hypotheses of the theorem for all parameter values. The proof of the theorem constructs g using the Legendre-Fenchel transform [45] of an antiderivative of s. Because of the nature of the Legendre-Fenchel transform, this often does not produce a closed-form expression for g. We consider this as an acceptable price to pay for having an explicit proximal mapping, which is much more useful for most of today's state-of-the-art algorithms for compressed sensing than having an explicit penalty function. In the case of the penalty function G_p induced by p-shrinkage, we can compute $g_p(w)$ numerically, and example plots are in Fig. 2. In addition to the properties guaranteed by Thm. II.4, it can be shown that $\lim_{w\to\infty} g_p(w) - w^p/p - C_p = 0$ for $p \neq 0$ and constant C_p depending only on p. This includes $p < 0$, in which case it follows that $g_p(w)$ is bounded above. For $p = 0$, we have $\lim_{w\to\infty} g_0(w) - \log w - C = 0$ instead.

In the case of the penalty function G_{firm} induced by firm thresholding, g_{firm} does have a closed form:

$$g_{\text{firm}}(w) = \begin{cases} |w| - w^2/(2\mu) & \text{if } |w| \leq \mu, \\ \mu/2 & \text{if } |w| \geq \mu. \end{cases} \tag{12}$$

Note that $g_{\text{firm}}(w)$ is independent of λ, except that $\mu \geq \lambda$ is required by the definition of g_{firm}.

Although the statement of Thm. II.4 excludes hard thresholding (being discontinuous), the construction in the proof does produce a penalty function G_{hard}. It coincides with G_{firm} for $\mu = \lambda$. The part of the conclusion of the theorem that doesn't hold is that $\operatorname{prox}_\lambda G_{\text{hard}}(\lambda)$ is the entire interval $[0, \lambda]$, while $H_\lambda(\lambda)$ is generally defined to take on a single value from this interval (namely 0 in our definition (9)).

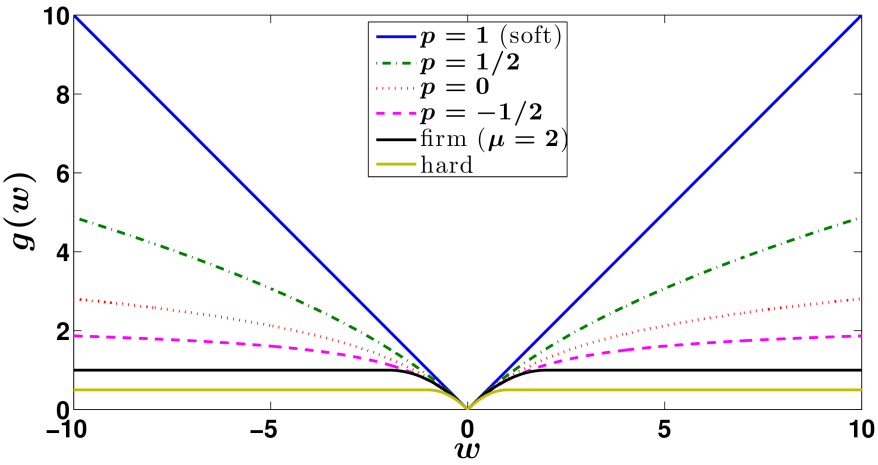


Fig. 2. Plot of penalty component function g induced by several shrinkage mappings, all with $\lambda = 1$. The smaller the value of p, the slower the growth of the p-shrinkage penalty function, being bounded above when $p < 0$. Both firm and hard thresholding have penalty functions that are quadratic near the origin, then constant.

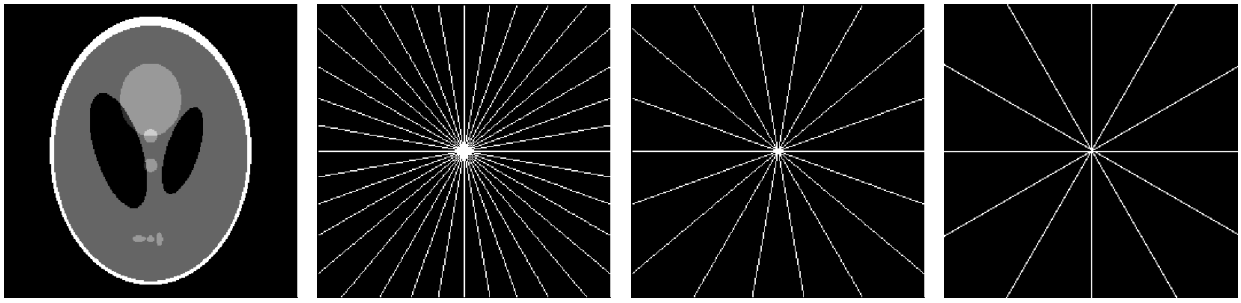
C. Example

To motivate the consideration of p-shrinkage and firm thresholding, we consider a generalization of an example appearing in the first compressed sensing paper [1]. We seek to reconstruct the 256×256 Shepp-Logan phantom image from samples of its 2-D discrete Fourier transform (DFT), taken along radial lines, thereby simulating both MRI and X-ray CT data (the latter by way of the Fourier slice theorem). See Fig. 3. Since the phantom has a sparse gradient, we seek to solve the following optimization problem:

$$\min_x G(\nabla x) \text{ subject to } \mathcal{F}x = b, \tag{13}$$

where G is one of the penalty functions being compared, ∇ is a discrete gradient using forward differences and periodic boundary conditions, $\mathcal{F}$ is the 2-D DFT, and b contains the sample data. We solve (13) with ADMM, where the shrinkage mapping is p-shrinkage with $p \leq 1$ or firm thresholding. See [25] for details, being also a straightforward generalization of the algorithm of [34].

With $G = G_1 = \|\cdot\|_1$, 18 lines are required for exact reconstruction, while using $G = G_{-1/2}$, 9 lines suffice, as shown in [21], the latter being the fewest that had been demonstrated at that time. In [22] (see also [23]), 6 lines were shown to suffice using the G induced by a shrinkage mapping that is a C^∞ approximation of hard thresholding. This is the fewest possible, since



(a) Shepp-Logan phantom (b) $p = 1$, 18 lines (c) $p = -1/2$, 9 lines (d) firm, 6 lines

Fig. 3. The Shepp-Logan phantom, and the number of radial lines of Fourier samples needed to reconstruct the phantom perfectly using different penalty functions.

with 5 lines, there are fewer measurements than nonzero gradient pixels, so that the phantom will not even be a local minimizer of the problem with $G = \|\cdot\|_0$. However, here we report that using $G = G_{\mathrm{firm}}$ (with $\lambda = 0.1$ and $\mu = 2.5$), 6 lines also suffice, and many fewer ADMM iterations are needed (337 versus 2213).

While this example is an ideal case, using a very sparse image and noisefree measurements, this does demonstrate that p-shrinkage and firm thresholding induce penalty functions that can be useful for recovering sparse signals. Now we turn to a theoretical analysis of the sparse recovery performance of minimizing these penalty functions.

III. Exact recovery

In this section, we establish sufficient conditions for exact recovery of sparse signals from noisefree measurements by solving a minimization problem with penalty function G:

$$\min_w G(w) \text{ subject to } Aw = b. \tag{14}$$

Our objective is to determine sufficient conditions in the case where G is a penalty function induced by a shrinkage mapping; however, we will establish conditions for a somewhat more general class of penalty functions G. We shall assume that the measurement matrix $A \in \mathbb{R}^{m\times n}$ has the Unique Representation Property (URP), i.e., any m columns of A are linearly independent. This implies that any vector in $\ker(A)$ has at least $m+1$ nonzero entries. The URP can be regarded as a *generic* property of matrices; for example, a matrix whose entries are independently and

identically distributed samples drawn from any absolutely continuous probability distribution will have URP with probability 1.

Remark III.1. The URP implies that the m rows of A are linearly independent. Thus an orthonormal basis for the span of the rows can be formulated as linear combinations of the rows of A. So if we multiply A by a product of elementary matrices, E, corresponding to the necessary elementary row operations, the resulting product will have orthonormal rows. Since elementary matrices are invertible, $Aw = b$ is equivalent to $EAw = Eb$. Also, since each elementary matrix is invertible, A_T being full rank for $|T| = m$ implies EA_T is full rank as well, and so A satisfying the URP implies EA satisfies the URP. Thus we can always transform the problem so that the rows of A are orthonormal, i.e., $AA^T = I$, and so without loss of generality, we assume that the A given satisfies $AA^T = I$.

We shall also assume that $G(w) = \sum_i g(w_i)$ with

I) $g(0) = 0$, and g even on $\mathbb{R}$; and

II) g is continuous on $\mathbb{R}$, and either strictly increasing and strictly concave on $\mathbb{R}$, or strictly increasing and strictly concave on $(0, \gamma]$ and constant on $[\gamma, \infty)$ for some $\gamma > 0$.

These conditions imply that g is nondecreasing and concave on $[0, \infty)$, is everywhere nonnegative, and satisfies the triangle inequality.

Lemma III.2. *The penalty functions G_{firm} and G_p (for $-\infty < p < 1$) satisfy the above conditions.*

Proof: It is clear from the expression (12) for g_{firm} that G_{firm} satisfies the conditions with $\gamma = \mu$.

For G_p, by Thm. II.4 we get condition I, and that g_p is differentiable on $(0, \infty)$ with $g_p' > 0$. It suffices to prove that g_p is twice differentiable on $(0, \infty)$ with $g_p'' < 0$; it will be no more difficult to show that $g_p \in C^\infty(0, \infty)$. We need some details from the construction of g_p, from [23]. We have

$$g_p(w) = (f_p^*(w) - w^2/2)/\lambda, \tag{15}$$

where $f_p' = s_p$ and f_p^* is the Legendre-Fenchel transform of f_p. Since s_p is continuous and nondecreasing, f_p is C^1 and convex. Then by [45, Prop. 11.3], we have that

$$x \in \partial f_p^*(w) \Leftrightarrow w = f_p'(x) = s_p(x). \tag{16}$$

Fix $w > 0$, and let x be such that $w = s_p(x)$. From (8), we must have $x > \lambda$, so $w = x - \lambda^{2-p}x^{p-1}$. If we define $F(x, w) = x - \lambda^{2-p}x^{p-1} - w$, we have that $F(\cdot, w)$ is C^∞ on $(0, \infty)$, and $\frac{\partial^k F}{\partial x^k}(x, w) \neq 0$ for $x \in (\lambda, \infty)$. Thus by the implicit function theorem, f_p^* is C^∞ on $(0, \infty)$, hence g_p is as well by (15).

Returning to $w = x - \lambda^{2-p}x^{p-1}$, by (15), (16), and the differentiability of f_p^*, we have

$$g_p'(w) = ((f_p^*)'(w) - w)/\lambda = (\lambda/x)^{1-p}. \tag{17}$$

Thus $g_p'(w)$ is decreasing in x on (λ, ∞), and since x is a strictly increasing function of w on $(0, \infty)$, $g_p''(w) < 0$ on $(0, \infty)$. ■

Lemma III.3. *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP and G satisfies (I,II) above. Then the global minimizer of* (14) *has m or fewer nonzero entries.*

Proof: Consider w such that $Aw = b$ and $\|w\|_0 > m$. Define the matrix M to have the columns $-w_i e_i$. The set of vectors Mv with $\mathrm{supp}(v) \subset \mathrm{supp}(w)$ span a subspace of dimension greater than m. Since $\dim(\ker(A)) = n - m$, we can choose a v with $Mv \in \ker(A)$ and $\|v\|_\infty = 1$.

For all $t \in \mathbb{R}$, $w + tMv$ is feasible. Define $T = \{i : v_i \neq 0 \text{ and } |w_i| < \gamma\}$ (taking $\gamma = +\infty$ if the first case of assumption II holds). First suppose $T \neq \emptyset$. Then by assumption II, the function $t \mapsto G(w + tMv)$ is strictly concave on an interval $[-\delta, \delta]$, with $\delta > 0$ chosen small enough that every $(w + tMv)_i$ has the same sign as w_i for all $|t| \leq \delta$. Then $G(w) > \min\{G(w - \delta Mv), G(w + \delta Mv)\}$, and w is not a global minimizer.

Otherwise, we have $v_i \neq 0 \Rightarrow |w_i| \geq \gamma$. Let $t_0 = \sup\{t : \forall i \min\{|(w - tMv)_i|, |(w + tMv)_i|\} \geq \gamma\}$. Then taking $t_1 = t_0 + \delta$ with $\delta > 0$ again small enough that every $(w \pm t_1 Mv)_i$ has the same sign as w_i, then one of $|(w \pm t_1 Mv)_i|$ is less than γ for at least one i, giving a smaller value of g. Since all other components keep g constant, we have one of $G(w \pm t_1 Mv)$ being smaller than $G(w)$. ■

Lemma III.4. *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP. Then the magnitudes of nonzero entries of vectors y satisfying $Aw = b$ with m or fewer nonzero entries are uniformly bounded below by some positive constant α and uniformly above by some positive constant β.*

Proof: By the URP, every m columns of A can admit no more than one solution. Thus

there are no more than $\binom{n}{m}$ vectors w satisfying $Aw = b$ with m or fewer nonzero entries. Thus the set of nonzero entries of these vectors is finite and bounded below and above by α, β respectively. Neither constant depends on G in any way. ■

Note that Lemma III.3 and Lemma III.4 imply that the global minimizer of the equality-constrained G minimization problem has nonzero entries with magnitude bounded below by α and above by β.

Next we introduce the G Nullspace Property, a generalization of the ℓ^1 Nullspace Property introduced in [46] for norms and implicitly in [11] for penalty functions belonging to a particular class. We denote $\{1, 2, \ldots, n\} = [n]$, and T^c denotes the complement of T in $[n]$.

Definition III.5. The *G Nullspace Property* (or *G NSP*) of order k for the matrix A is satisfied when for all $h \in \ker(A)\backslash\{0\}$ and $T \subset [n]$ with $|T| \leq k$, one has $G(h_T) < G(h_{T^c})$.

Proposition III.6. *For a penalty function G satisfying the triangle inequality, the G NSP implies exact recovery.*

Proof: We simply observe that the proof of [11] works assuming only that the penalty function satisfies the triangle inequality. ■

Definition III.7. Let the matrix $A \in \mathbb{R}^{m\times n}$ and the vector $b \in \mathbb{R}^m$ be given. Let x be the sparsest solution to $Aw = b$, $k = \|x\|_0$ with $2k \leq m$, and $T = \operatorname{supp}(x)$. We say the *G Restricted Nullspace Property* (or *G RNSP*) of order k is satisfied if whenever w satisfies $Aw = b$ and $\|w\|_0 \leq m$, then for $h = x - w$, we have either $h = 0$ or $G(h_T) < G(h_{T^c})$.

Note that the G NSP of order k for A implies the G RNSP of order k for A. However, examining the proof of Proposition III.6 from [11] and applying Lemma III.3 shows that in fact G RNSP suffices for exact recovery. We assume $2k \leq m$ to guarantee that the sparsest solution of $Aw = b$ is unique, as URP ensures that a second solution must have more than $m-k$ nonzero components.

Proposition III.8. *For penalty function G satisfying the triangle inequality, G RNSP implies exact recovery.*

Theorem III.9 (G exact recovery)**.** *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP and G satisfies (I,II)*

above. For given b, let x^* be the global minimizer of (14) *and x the sparsest feasible vector. Let $k = \|x\|_0$, and define α, β to be the lower and upper bound of magnitudes of nonzero entries of feasible vectors with m or fewer nonzero components as in Lemma III.4. If $2k \leq m$ and $kg(2\beta) < (m+1-k)g(\alpha)$ then $x^* = x$.*

Proof: Let $h = x^* - x$. Since x is supported on T, $h_{T^c} = x^*_{T^c}$, and so for all $t \in T^c$, $|h(t)|$ is either zero or at least α. Also, since $h \in \ker(A)$, if $h \neq 0$, then $\|h_{T^c}\|_0 \geq m+1-k$ (otherwise we would have $\|h\|_0 \leq m$, violating URP), so that $G(h_{T^c}) \geq (m+1-k)g(\alpha)$. Also,

$$G(h_T) \leq \sum_{i \in T} g(|x^*_i| + |x_i|) \leq kg(2\beta) < (m+1-k)g(\alpha) \tag{18}$$

by assumption. Thus either $h = 0$ or $G(h_T) < G(h_{T^c})$, so G RNSP is satisfied. ■

Corollary III.10 (G_{firm} exact recovery)**.** *Assume $A \in \mathbb{R}^{m \times n}$ satisfies URP and $G = G_{\text{firm}}$, the penalty corresponding to firm thresholding. For given b, let x^* be the global minimizer of* (14) *and x the sparsest feasible vector. Let $k = \|x\|_0$. If $2k \leq m$ and*

$$\mu < \min\left\{\alpha\frac{m+1-k}{k}\left(1+\sqrt{1-\frac{k}{m+1-k}}\right), 2\beta\right\}, \tag{19}$$

then $x^* = x$.

Proof: Since A satisfies URP and G satisfies (I,II), we may apply Theorem III.9. The inequality conditions from Theorem III.9 are $2k \leq m$ and $kg(2\beta) < (m+1-k)g(\alpha)$. We know $\alpha < 2\beta$. If we have $\mu \leq \alpha$, then the inequality becomes $k\mu/2 < (m+1-k)\mu/2$ which follows automatically from $2k \leq m$. And so we satisfy the hypotheses of Theorem III.9, and thus have exact recovery. If instead we have $\alpha < \mu < 2\beta$, we can evaluate the desired inequality as follows:

$$\frac{k\mu}{2} \leq (m+1-k)(\alpha - \alpha^2/2\mu), \tag{20}$$

$$\mu^2 - 2\frac{m+1-k}{k}\alpha\mu + \frac{m+1-k}{k}\alpha^2 < 0, \tag{21}$$

$$\left|\mu - \alpha\frac{m+1-k}{k}\right| < \alpha\frac{m+1-k}{k}\sqrt{1-\frac{k}{m+1-k}}, \tag{22}$$

$$\alpha\frac{m+1-k}{k}\left(1-\sqrt{1-\frac{k}{m+1-k}}\right) < \mu < \alpha\frac{m+1-k}{k}\left(1+\sqrt{1-\frac{k}{m+1-k}}\right). \tag{23}$$

The left bound is always looser than the assumed $\alpha < \mu$ (for $2k < m+1$), so the condition $\mu < \alpha\frac{m+1-k}{k}\left(1+\sqrt{1-\frac{k}{m+1-k}}\right)$ gives the desired inequality and guarantees exact recovery. ■

Corollary III.11 (G_p exact recovery)**.** *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP and $G = G_p$, the $p-$shrinkage penalty. For given b, let x^* be the global minimizer of* (14) *and x the sparsest feasible vector. Let $k = \|x\|_0$. If $2k \leq m$ then there exist $\lambda > 0$ and $0 < p < 1$ sufficiently small that $x^* = x$. For any $p < 0$ there also exists $\lambda > 0$ sufficiently small that $x^* = x$.*

Proof: Since A satisfies the URP and G_p satisfies (I,II), we may apply Theorem III.9. The inequality conditions from Theorem III.9 are $2k \leq m$ and $kg(2\beta) < (m+1-k)g(\alpha)$.

Fix $w > 0$. As in the proof of Lemma III.2, we have

$$g_p(w) = (f_p^*(w) - w^2/2)/\lambda, \tag{24}$$

where $f_p' = s_p$ and f_p^* is the Legendre-Fenchel transform of f_p and is smooth at w. Let $x = (f_p^*)'(w)$, noting that while w is fixed, x depends on λ and p. By (16), we have $s_p(x) = w$, so that

$$x - w = \lambda^{2-p}x^{p-1}. \tag{25}$$

Furthermore, by [45, Prop. 11.3], we have

$$x = \arg\min_x \big(xw - f_p(x)\big), \tag{26}$$

so that by definition of the Legendre-Fenchel transform,

$$f_p^*(w) = xw - f_p(x). \tag{27}$$

Combining (24), (25), and (27), we obtain

$$\begin{aligned} g_p(w) &= (xw - f_p(x) - w^2/2)/\lambda \\ &= \big(xw - x^2/2 + \lambda^{2-p}x^p/p - \lambda^2(1/p - 1/2) - w^2/2\big)/\lambda \qquad (28)\\ &= \lambda^{1-p}x^p/p - (x-w)^2/(2\lambda) - \lambda(1/p - 1/2) \\ &= \frac{\lambda}{p}(x/\lambda)^p - \frac{\lambda}{2}(x/\lambda)^{2p-2} - \lambda(1/p - 1/2). \qquad (29) \end{aligned}$$

(In (28), the expression for $f_p(x)$ is obtained by antidifferentiating s_p with $f_p(0) = 0$.)

a) Case $0 < p < 1$*:* We want to show that for sufficiently small $0 < \lambda$ and $0 < p < 1$, $g(2\beta)/g(\alpha) < (m+1-k)/k$. By hypothesis, $(m+1-k)/k > 1$. So it suffices to show for any fixed α, β with $0 < \alpha < 2\beta$, that $g(2\beta)/g(\alpha) \to 1$ as $(p, \lambda) \to (0^+, 0^+)$.

By (25), $x > w$ for any λ and p, so $\lim_{\lambda \to 0^+}(x/\lambda) = \infty$. Then for $p < 1$,

$$\lim_{\lambda \to 0^+} g_p(w) - \frac{\lambda}{p}\big[(x/\lambda)^p - 1\big] = 0. \tag{30}$$

Now

$$\frac{\lambda}{p}\big[(x/\lambda)^p - 1\big] = \frac{\lambda}{p}\big[\exp\big(p\log(x/\lambda)\big) - 1\big] = \frac{\lambda}{p}\big[p\log(x/\lambda) + o\big(p\log(x/\lambda)\big)\big], \tag{31}$$

where the little-o is as $p\log(x/\lambda) \to 0^+$, which we wish to establish as $p, \lambda \to 0^+$. Since $x > w$, we have that

$$p\log(x/\lambda) = p\log(w/\lambda + (x/\lambda)^{p-1}) < p\log(w/\lambda + (w/\lambda)^{p-1}) \to 0^+ \tag{32}$$

provided $p \to 0^+$ fast enough, such as if $p \sim \lambda^q$ for any $q > 0$. This yields

$$\begin{aligned}\liminf_{(\lambda,p)\to(0^+,0^+)} \frac{g_p(2\beta)}{g_p(\alpha)} &= \liminf_{(\lambda,p)\to(0^+,0^+)} \frac{\lambda\log(x(2\beta)/\lambda)}{\lambda\log(x(\alpha)/\lambda)} \\ &\le \liminf_{(\lambda,p)\to(0^+,0^+)} \frac{\log(2\beta/\lambda + (2\beta/\lambda)^{p-1})}{\log(\alpha/\lambda)} = \liminf_{\lambda\to 0^+} \frac{\log(2\beta) - \log(\lambda)}{\log(\alpha) - \log(\lambda)} = 1. \end{aligned} \tag{33}$$

Therefore, there exist $\lambda > 0, p > 0$ sufficiently small that $kg(2\beta) < (m+1-k)g(\alpha)$.

b) Case $p < 0$*:* Since g_p is strictly increasing on $[0, \infty)$, we take $w \to \infty$ to determine an upper bound. Note that $x(w) > w$ implies that $x(w) \to \infty$ as $w \to \infty$. Then from (29), since now $p < 0$, we obtain

$$\lim_{w\to\infty} g_p(w) = \lambda(1/2 - 1/p). \tag{34}$$

Thus for $p < 0$ and all w, λ, we have $g_p(w) \le \lambda(1/2 - 1/p)$. Applying this with $w = 2\beta$ and using (29),

$$\liminf_{\lambda\to 0^+} \frac{g_p(2\beta)}{g_p(\alpha)} \le \liminf_{\lambda\to 0^+} \frac{\lambda(1/2 - 1/p)}{\lambda\big[\frac{1}{p}(x(\alpha)/\lambda)^p - \frac{1}{2}(x(\alpha)/\lambda)^{2p-2} - (1/p - 1/2)\big]}. \tag{35}$$

As before, $(x/\lambda) \to \infty$ as $\lambda \to 0^+$. Then

$$\liminf_{\lambda\to 0^+} \frac{g_p(2\beta)}{g_p(\alpha)} \le \lim_{\lambda\to 0^+} \frac{\lambda(1/2 - 1/p)}{\lambda(1/2 - 1/p)} = 1. \tag{36}$$

Thus for every $p < 0$ there exists $\lambda > 0$ sufficiently small that $kg(2\beta) < (m+1-k)g(\alpha)$. ■

IV. STABILITY

Next we consider the case of noisy measurements of an approximately sparse signal. Let x be the original signal with $\|Ax-b\|_2 \leq \epsilon$ whose k-sparse approximation is supported on T, i.e. $x_T = \arg\min_w G(x-w)$ subject to $\|w\|_0 = k$. We wish to bound $G(x^*-x)$ where

$$x^* = \arg\min_w G(w) \text{ subject to } \|Aw-b\|_2 \leq \epsilon. \tag{37}$$

We shall bound the recovery error by the sum of a term dependent on the noise level and a term dependent on the sparse approximation error.

We shall first need two results: bounds on the magnitudes of nonzero entries of local minima of (37) and an extension of those bounds to the error vector projected onto the null space of A. Recall that $\|w\|_{-\infty} := \min_i |w_i|$.

Lemma IV.1. *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP and G satisfies (I,II) above. Let $b \in \mathbb{R}^m$ be given. For $S \subset [n]$ with $|S| = m$ define $\alpha_S = \|A_S^{-1}b\|_{-\infty}$ and $\beta_S = \|A_S^{-1}b\|_\infty$. If $\epsilon < \min_S(\alpha_S/\|A_S^{-1}\|)$, then the magnitudes of components of feasible vectors of* (37) *are bounded below by $\alpha := \min_S(\alpha_S - \|A_S^{-1}\|\epsilon) > 0$ and bounded above by $\beta := \max_S(\beta_S + \|A_S^{-1}\|\epsilon)$.*

The assumption that $\alpha_S > 0$ for all S has a similar character to the URP, in that it is true with probability 1 for random data drawn from an absolutely continuous distribution.

Proof: First, note that the error-bounded problem (37) is equivalent to taking the G minimizer from a set of equality-constrained G minimizers (with different equality constraints): For all feasible w, we must have $Aw = b+\eta$ for some $\|\eta\|_2 \leq \epsilon$. Thus by Lemma III.3 the minimizer of (37) has m or fewer nonzero entries. By the URP, any m columns S of A give exactly one solution to $A_S w = b+\eta$. So we have

$$\begin{aligned}
\|w\|_{-\infty} &= \|A_S^{-1}(b+\eta)\|_{-\infty} \geq \min_i(|A_S^{-1}b| - |A_S^{-1}\eta|)_i \\
&\geq \|A_S^{-1}b\|_{-\infty} - \|A_S^{-1}\eta\|_\infty \geq \alpha_S - \|A_S^{-1}\eta\|_2 \geq \alpha_S - \|A_S^{-1}\|\epsilon \\
&\geq \alpha,
\end{aligned} \tag{38}$$

and

$$\begin{aligned}\|w\|_\infty = \|A_S^{-1}(b+\eta)\|_\infty &\le \|A_S^{-1}b\|_\infty + \|A_S^{-1}\eta\|_\infty \\ &\le \beta_S + \|A_S^{-1}\eta\|_2 \le \beta_S + \|A_S^{-1}\|\epsilon \\ &\le \beta. \end{aligned} \tag{39}$$

■

Lemma IV.2. *Assume G satisfies (I,II). Let x^* be the global minimizer of* (37)*, x the original signal with $\|Ax-b\| \le \epsilon$, and let T be the support of the k-sparse approximation of x. Let α_S, α, and β be as in Lemma IV.1. Define $\alpha' := \alpha - \|x_{T^c}\|_\infty - 2\epsilon$ and $\beta' := \beta + \epsilon$. If A satisfies the URP, $AA^T = I$, $\min_S \alpha_S > \|x_{T^c}\|_\infty$ (requiring that x be nearly k sparse), and $\epsilon < \min_S\{(\alpha_S - \|x_{T^c}\|_\infty)/(2+\|A_S^{-1}\|)\}$, then the orthogonal projection w of $h = x^* - x$ onto the nullspace of A satisfies*

$$\alpha' \le \|w_{T^c}\|_{-\infty} \text{ and } \|w_{T^c}\|_\infty \le 2\beta'. \tag{40}$$

Proof: First, consider the bound $\epsilon < \min_S\{(\alpha_S - \|x_{T^c}\|_\infty)/(2+\|A_S^{-1}\|)\}$. Note that this is stronger than the bound on ϵ from Lemma IV.1, and it implies $2\epsilon + \|x_{T^c}\|_\infty < \alpha$. We see this from the following inequalities:

$$\begin{aligned}\alpha &= \min_S\big\{\alpha_S - \epsilon\|A_S^{-1}\|\big\} \\ &> \min_S\big\{\alpha_S - (\alpha_S - \|x_{T^c}\|_\infty)\|A_S^{-1}\|/(2+\|A_S^{-1}\|)\big\} \\ &= \min_S\left\{\frac{2\alpha_S}{2+\|A_S^{-1}\|} + \frac{\|A_S^{-1}\|\|x_{T^c}\|_\infty}{2+\|A_S^{-1}\|}\right\} \\ &= \min_S\left\{\frac{2\alpha_S - 2\|x_{T^c}\|_\infty}{2+\|A_S^{-1}\|} + \frac{(2+\|A_S^{-1}\|)\|x_{T^c}\|_\infty}{2+\|A_S^{-1}\|}\right\} \\ &> 2\epsilon + \|x_{T^c}\|_\infty. \end{aligned} \tag{41}$$

We shall use this below to guarantee $\alpha' > 0$.

Note that the hypotheses of Lemma IV.1 are satisfied, giving $\|x^*\|_{-\infty} \ge \alpha$ and $\|x^*\|_\infty \le \beta$, $\|x\|_\infty \le \beta$. Since $AA^T = I$, the orthogonal projection of h onto the nullspace of A is $(I - A^TA)h$. The desired lower bound comes from the following sequence of inequalities, using the given lower bound on nonzero elements of x^*, the feasibility of x^* and x, the fact $\|A^TA\| = 1$, and

the assumed bound on ϵ:

$$\begin{aligned}\|[(I - A^T A)h]_{T^c}\|_{-\infty} &\geq \|h_{T^c}\|_{-\infty} - \|A^T Ah\|_\infty \\ &\geq \|x^*_{T^c} - x_{T^c}\|_{-\infty} - \|A^T Ah\|_2 \\ &\geq \|x^*_{T^c}\|_{-\infty} - \|x_{T^c}\|_\infty - \|h\|_2 \\ &\geq \alpha - \|x_{T^c}\|_\infty - 2\epsilon = \alpha' > 0.\end{aligned}$$

The upper bound comes from a completely analogous argument:

$$\begin{aligned}\|(I - A^T A)h\|_\infty &\leq \|h\|_\infty + \|A^T Ah\|_\infty \\ &\leq \|x^* - x\|_\infty + \|A^T Ah\|_2 \\ &\leq 2\beta + 2\epsilon = 2\beta'.\end{aligned}$$

■

Definition IV.3. The *G Noisy Nullspace Property* (or *G NNSP*) of order k for the matrix A is satisfied when for all $h \in \mathbb{R}^n$ and $S \subset [n]$ with $|S| \leq k$, there are constants $0 \leq \tau < 1$ and $D \geq 0$ such that

$$G(h_S) \leq \tau G(h_{S^c}) + D\|Ah\|_2. \tag{42}$$

Proposition IV.4. *Assume G satisfies the triangle inequality. For given A, b, let x^* be the global minimizer of* (37) *and let x be the original signal with $\|Ax - b\|_2 \leq \epsilon$ whose k-sparse approximation is supported on T. Then the G NNSP of order k for A implies the following stability bound:*

$$G(x^* - x) \leq C_1\epsilon + C_2 G(x_{T^c}) \tag{43}$$

with $C_1 = 4D/(1-\tau)$ and $C_2 = 2(1+\tau)/(1-\tau)$, where τ and D satisfy (42)*.*

Proof: Define the error vector $h = x^* - x$. Since x^* and x are both feasible and $\|A\| = 1$, $\|Ah\|_2 \leq 2\epsilon$. Then by the triangle inequality of G,

$$G(x_T) - G(-h_T) \leq G(x_T + h_T). \tag{44}$$

Since G decouples across components,

$$G(x_T + h_T) + G(h_{T^c}) = G(x_T + h_T + h_{T^c}) = G(x^* - x_{T^c}). \tag{45}$$

Then

$$\begin{aligned} G(h_{T^c}) &\le G(x^* - x_{T^c}) + G(h_T) - G(x_T) \\ &\le G(x^*) + G(x_{T^c}) + G(h_T) - G(x_T) \\ &\le G(x) + G(x_{T^c}) + G(h_T) - G(x_T) \\ &= 2G(x_{T^c}) + G(h_T). \end{aligned} \tag{46}$$

Now apply G NNSP to h on T:

$$G(h_T) \le \tau G(h_{T^c}) + D\|Ah\|_2 \le 2\tau G(x_{T^c}) + \tau G(h_T) + 2D\epsilon, \tag{47}$$

so that

$$G(h_T) \le \frac{2}{1-\tau}\big(D\epsilon + \tau G(x_{T^c})\big). \tag{48}$$

Using (46), we obtain

$$G(h_{T^c}) \le 2G(x_{T^c}) + G(h_T) \le \frac{2D\epsilon}{1-\tau} + \frac{2}{1-\tau}G(x_{T^c}). \tag{49}$$

Now we add (48) and (49) to get the desired inequality:

$$\begin{aligned} G(h) &= G(h_T) + G(h_{T^c}) \\ &\le \frac{2}{1-\tau}\big(D\epsilon + \tau G(x_{T^c})\big) + \frac{2D}{1-\tau}\epsilon + \frac{2}{1-\tau}G(x_{T^c}) \\ &= \frac{4D}{1-\tau}\epsilon + \frac{2(1+\tau)}{1-\tau}G(x_{T^c}). \end{aligned} \tag{50}$$

■

Theorem IV.5 (G stability)**.** *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP, $AA^T = I$, G satisfies (I,II) above, and $G(v) \le C\sqrt{n}\|v\|_2$ for some constant $C > 0$. For given b, let x be the original signal with $\|Ax-b\|_2 \le \epsilon$, let T be the support of its k-sparse approximation, and suppose $\min_S\{\alpha_S\} > \|x_{T^c}\|_\infty$. Let x^* be the global minimizer of* (37)*, where $\epsilon < \min_S\{(\alpha_S - \|x_{T^c}\|_\infty)/(2+\|A_S^{-1}\|)\}$ (with α_S defined as in Lemma IV.1). Define α', β' as in Lemma IV.2. Assume that $2k < n$ and $kg(2\beta') < (n-k)g(\alpha')$. Then*

$$G(x^* - x) \le 2\left(1 - \frac{kg(2\beta')}{(n-k)g(\alpha')}\right)^{-1}\left[2C\sqrt{n}\epsilon + \left(1 + \frac{kg(2\beta')}{(n-k)g(\alpha')}\right)G(x_{T^c})\right]. \tag{51}$$

Proof: We shall show that the given hypotheses allow for the same application of the G NNSP as in Proposition IV.4, and in a similar way, arrive at stability. Define $h = x^* - x$. Since G satisfies the triangle inequality, we have $G(h_{T^C}) \leq G(h_T) + 2G(x_{T^C})$, as in the proof of Proposition IV.4.

Next we write h as the sum of its orthogonal projections onto $\ker(A)$ and $\ker(A)^\perp$, which we denote by w and v respectively. First, suppose that there exists some $0 \leq \tau < 1$ such that $G(w_T) \leq \tau G(w_{T^c})$ (which we will prove below). Then we have:

$$\begin{aligned} G(h_T) &\leq G(w_T) + G(v_T) \\ &\leq \tau G(w_{T^c}) + G(v_T) = \tau G(w_{T^c} + v_{T^c} - v_{T^c}) + G(v_T) \\ &\leq \tau G(h_{T^c}) + G(v_{T^c}) + G(v_T) \\ &= \tau G(h_{T^c}) + G(v) \\ &\leq \tau G(h_{T^C}) + C\sqrt{n}\|v\|_2. \end{aligned} \tag{52}$$

Since $AA^T = I$ and $v \in \ker(A)^\perp$, it follows that $v = A^T Av$. Hence $\|v\|_2^2 = \|Av\|_2^2$. Then from (52) we obtain

$$G(h_T) \leq \tau G(h_{T^c}) + C\sqrt{n}\|Av\|_2. \tag{53}$$

And so we have the application of the G NNSP to h on T with constants τ and $D = C\sqrt{n}$. From here the stability inequality (51) follows as in Proposition IV.4.

Now we go back to prove $G(w_T) \leq \tau G(w_{T^c})$. We shall use the lower bound $\|w_{T^c}\|_{-\infty} \geq \alpha'$ and the upper bound $\|w_T\|_\infty \leq \beta'$ from Lemma IV.2. We overestimate $G(w_T)$ and underestimate $G(w_{T^c})$ as follows:

$$G(w_T) \leq kg(2\beta'), \quad G(w_{T^c}) \geq (n-k)g(\alpha'). \tag{54}$$

So to get $G(w_T) \leq \tau G(w_{T^c})$, it suffices to have $kg(2\beta') \leq \tau(n-k)g(\alpha')$, and thus $kg(2\beta') < (n-k)g(\alpha')$ guarantees some $0 \leq \tau < 1$. The condition $k < n-k$ gives $(n-k)/k > 1$ and thus makes the inequality possible for $\alpha' < 2\beta'$.

Plugging in $\tau = \frac{kg(2\beta')}{(n-k)g(\alpha')}$ to the stability inequality we get from the previous argument gives

$$G(h) \leq 2\left(1 - \frac{kg(2\beta')}{(n-k)g(\alpha')}\right)^{-1}\left[2C\sqrt{n}\epsilon + \left(1 + \frac{kg(2\beta')}{(n-k)g(\alpha')}\right)G(x_{T^c})\right]. \tag{55}$$

■

Corollary IV.6 (G_{firm} stability)**.** *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP, $AA^T = I$, and $G = G_{\text{firm}}$, the penalty corresponding to firm thresholding. For given b, let x be the original signal with $\|Ax - b\|_2 \leq \epsilon$ whose k-sparse approximation is supported on T, with $\min_S\{\alpha_S\} > \|x_{T^c}\|_\infty$, and x^* be the global minimizer of* (37)*, where $\epsilon < \min_S\{(\alpha_S - \|x_{T^c}\|_\infty)/(2+\|A_S^{-1}\|)\}$ (with α_S defined as in Lemma IV.1). Define α', β' as in Lemma IV.2. If $2k < n$ and $\mu < \min\{\alpha' \frac{n-k}{k}\left(1 + \sqrt{1 - \frac{k}{n-k}}\right), 2\beta'\}$ then x^* is stable, satisfying the following inequality:*

$$G_{\text{firm}}(x^* - x) \leq 2\left(1 - \frac{kg_{\text{firm}}(2\beta')}{(n-k)g_{\text{firm}}(\alpha')}\right)^{-1}\left[2C\sqrt{n}\epsilon + \left(1 + \frac{kg_{\text{firm}}(2\beta')}{(n-k)g_{\text{firm}}(\alpha')}\right)G_{\text{firm}}(x_{T^c})\right]. \tag{56}$$

The proof of Corollary IV.6 is an application of Theorem IV.5 combined with the corresponding computations from the proof of Corollary III.10.

Corollary IV.7 (G_p stability)**.** *Assume $A \in \mathbb{R}^{m\times n}$ satisfies the URP, $AA^T = I$, and $G = G_p$, the penalty corresponding p-shrinkage. For given b, let x be the original signal with $\|Ax - b\|_2 \leq \epsilon$ whose k-sparse approximation is supported on T, with $\min_S\{\alpha_S\} > \|x_{T^c}\|_\infty$, and x^* be the global minimizer of* (37)*, where $\epsilon < \min_S\{(\alpha_S - \|x_{T^c}\|_\infty)/(2+\|A_S^{-1}\|)\}$ (with α_S defined as in Lemma IV.1). If $2k < n$ then there exist $0 < p < 1, 0 < \lambda$ sufficiently small so that x^* is stable, satisfying the following inequality.*

$$G_p(x^* - x) \leq 2\left(1 - \frac{kg_p(2\beta')}{(n-k)g_p(\alpha')}\right)^{-1}\left[2C\sqrt{n}\epsilon + \left(1 + \frac{kg_p(2\beta')}{(n-k)g_p(\alpha')}\right)G(x_{T^c})\right]. \tag{57}$$

Also, for any $p < 0$ there exists $\lambda > 0$ sufficiently small such that x^* is stable, and the above inequality holds.

The proof of Corollary IV.7 is an application of Theorem IV.5 combined with the corresponding computations from the proof of Corollary III.11.

V. Convergence of iterative p-shrinkage

Now we consider an algorithm that employs generalized shrinkage. Consider the following optimization problem:

$$\min_x F_p(x) := \lambda G_p(x) + \tfrac{1}{2}\|Ax - b\|_2^2, \tag{58}$$

where $\|A\| < 1$. Applying forward-backward splitting to this problem gives *iterative p-shrinkage* (IPS):

$$x^{n+1} = S_p(x^n - A^T(Ax^n - b)). \tag{59}$$

This generalizes the *iterative soft thresholding* algorithm (ISTA) [31], which is the case $p = 1$. ISTA was shown in [31] to be globally convergent to a global minimizer (necessarily, since F_1 is convex). In this section, we prove global convergence of IPS for general $p < 1$, though only to a stationary point of F_p. Portions of the proof appeared in [28], though statements there concerning convergence to a local minimizer are incorrect.

Recall from Lemma III.2 that g_p is C^∞ on $(0, \infty)$. A closer examination of the proof shows that g_p on $[0, \infty)$ is the restriction of a function that is C^∞ on $\mathbb{R}$, so g_p is one-sided differentiable to all orders at $w = 0$.

The following follows exactly as in the known case of $p = 1$ [31]:

Lemma V.1 ([28])**.** *Let $\lambda > 0$ and $p \in \mathbb{R}$, and define $\{x^n\}$ by* (59)*, with x^0 arbitrary.*

1) *$F(x^{n+1}) \leq F(x^n)$ for all n, and $F(x^{n+1}) < F(x^n)$ unless x^n is a fixed point of the algorithm.*
2) *$\|x^{n+1} - x^n\|_2 \to 0$.*

Lemma V.2. *Let $\lambda > 0$ and $p \in \mathbb{R}$. The fixed points of* (59) *are precisely the stationary points of F_p.*

Proof: The iteration (59) can be seen as minimizing the surrogate functional

$$\lambda G_p(x) + \tfrac{1}{2}\|Ax - b\|_2^2 + \tfrac{1}{2}\|x - w\|_2^2 - \tfrac{1}{2}\|Ax - Aw\|_2^2 \tag{60}$$

with fixed $w = x^n$, by expanding the quadratic terms and rearranging to express the minimizer in terms of the proximal mapping of G_p. Therefore the first-order optimality condition of this functional is satisfied at $x = x^{n+1}$. Also, the first-order optimality condition of this functional at $x = x^n$ is the same as the first-order optimality condition of F_p at $x = x^n$. Hence $x^{n+1} = x^n$ if and only if the first-order optimality condition of F_p at $x = x^n$ is satisfied. ■

The lemma shows why it is not possible to show that IPS converges to a local minimizer: if the algorithm happens to be initialized with a stationary point that is not a local minimizer (*i.e.*,

a saddle point or local maximizer), then the initializer is a fixed point of the algorithm, so the algorithm cannot converge to a local minimizer in such a case.

Lemma V.3. *Fix $\lambda > 0$, $p \in (-\infty, 1)$. We have $g_p''' > 0$ on $(0, \infty)$, $g_p''' < 0$ on $(-\infty, 0)$, $g_p'''(0+) > 0$, and $g_p'''(0-) < 0$.*

Proof: Since g_p is even, it suffices to consider $w > 0$. Above we had that $x = x(w) = (f_p^*)'(w)$ satisfies $x - \lambda^{2-p} x^{p-1} = w$. Differentiating with respect to w, we have that

$$x' - \lambda^{2-p}(p-1)x^{p-2}x' = 1, \tag{61}$$

so

$$x' = \left(1 - \lambda^{2-p}(p-1)x^{p-2}\right)^{-1}. \tag{62}$$

Since $p < 1$, $(f_p^*)''(w) = x'(w) > 0$ for all $w > 0$.

Differentiating (61), we get

$$x'' - \lambda^{2-p}(p-1)\left[(p-2)x^{p-3}(x')^2 + x^{p-2}x''\right] = 0, \tag{63}$$

or

$$x''\left(1 - \lambda^{2-p}(p-1)x^{p-2}\right) = \lambda^{2-p}(p-1)(p-2)x^{p-3}(x')^2, \tag{64}$$

implying that x'' has the same sign as x. Since $x(w)$ has the same sign as w, we have that $(f_p^*)'''(w)$ has the same sign as w for $w \neq 0$.

Differentiating the relation (15) defining g_p, we obtain $w + \lambda g_p'(w) = (f_p^*)'(w)$, $1 + \lambda g_p''(w) = (f_p^*)''(w)$, and $\lambda g_p'''(w) = (f_p^*)'''(w)$. Thus $g_p'''(w)$ has the same sign as w for $w \neq 0$ as well. Also, $\lambda g_p'''(0+) = (f_p^*)'''(0+) = \lim_{w \to 0^+} x''(w)$. Since $\lim_{w \to 0^+} x(w) = \lambda$, we obtain from (62) and (64) that $(f_p^*)'''(0+) = \frac{1-p}{(2-p)^2}\lambda^{-1} > 0$. Thus $g'''(0+) > 0$. ■

Lemma V.4. *Let $p \geq 0$. Then $\{x^n\}$ is bounded.*

Proof: Since $\{F_p(x^n)\}$ decreases monotonically, it suffices to show that F_p is coercive, which we establish be showing coercivity of g_p. By (25), if $w \to \infty$, then $x \to \infty$. For $p > 0$,

that $g_p(w) \to \infty$ follows from (29). The $p = 0$ case is similar, but f_0 has a different form:

$$\begin{aligned} g_0(w) &= (xw - f_0(x) - w^2/2)/\lambda \\ &= \big(xw - x^2/2 + \lambda^2 \log x - \lambda^2(\log \lambda - 1/2) - w^2/2\big)/\lambda \\ &= \lambda \log x - (x-w)^2/(2\lambda) - \lambda(\log \lambda - 1/2) \\ &= \lambda \log x - \tfrac{\lambda}{2}(x/\lambda)^{-2} - \lambda(\log \lambda - 1/2). \end{aligned} \tag{65}$$

From this the coercivity of g_0 follows. ■

Lemma V.5. *Let $p < 0$, and assume $\lambda^2 > p\|b\|_2^2/(p-2)$. Let $x^0 = 0$. Then $\{x^n\}$ is bounded.*

Proof: From Lemma V.1, we know that $F_p(x^n)$ decreases (strictly except at a fixed point, in which case we are done). Then for $n \geq 1$,

$$F_p(x^n) < F_p(x^0) = \|b\|_2^2/2, \tag{66}$$

so

$$G_p(x^n) \leq F_p(x^n)/\lambda < \|b\|_2^2/(2\lambda). \tag{67}$$

By (34), $g_p(w) < (1/2 - 1/p)\lambda$. Combining this bound with (67), we obtain for each j,

$$g_p(x_j^n) \leq G_p(x^n) < \|b\|_2^2/(2\lambda) < (1/2 - 1/p)\lambda. \tag{68}$$

Letting t be the unique positive number satisfying $g(t) = \|b\|_2^2/(2\lambda)$, we obtain $\|x^n\|_\infty < t$ independently of n. ■

Now we can establish convergence of our algorithm.

Theorem V.6. *Let $\lambda > 0$, $p \in (-\infty, 1)$. Let the sequence $\{x^n\}$ be defined by* (59)*, with x^0 arbitrary for $p \geq 0$, and $x^0 = 0$ for $p < 0$ in which case we further assume $\lambda^2 > p\|b\|_2^2/(p-2)$. Then $\{x^n\}$ converges to a stationary point of F.*

Proof: We have that $F_p(x^{n+1}) < F_p(x^n)$ unless x^n is a fixed point, F is continuous, and the sequence $\{x^n\}$ is bounded. Then by [47, Thm. 3.1], we have that either $\{x^n\}$ converges or its limit points form a continuum. (A continuum is a compact, connected set; here we also exclude the degenerate case of a singleton.) Since we already know that any limit point of $\{x^n\}$ will be a stationary point of F_p, we complete the proof by showing that the stationary points of F_p cannot form a continuum.

Let E be the set of stationary points of F_p, and suppose E is a continuum. Fix $\bar{x} \in E$. For any $\epsilon > 0$, it cannot be that $\mathcal{N}(\bar{x}; \epsilon) \cap E = \{\bar{x}\}$, otherwise $\{\bar{x}\}$ would be both open and closed in E, contrary to E being connected. Thus there is a sequence of stationary points $\bar{x} + v^n$ with $v^n \neq 0$, $v^n \to 0$.

Since $\{v^n/\|v^n\|\}$ is a sequence of unit vectors, it cannot converge to zero. Then we can fix j such that $\{v_j^n/\|v^n\|\}$ does not tend to zero, though of course $v_j^n \to 0$. First suppose that $\bar{x}_j \neq 0$. By considering a tail of v_j^n, we can assume that $\bar{x}_j + v_j^n \neq 0$ for all n. Then g_p is differentiable at $\bar{x}_j$ and $\bar{x}_j + v_j^n$, and since $\bar{x}$ and $\bar{x} + v^n$ are fixed points,

$$\lambda^{2-p} g_p'(\bar{x}_j + v_j^n) + \left[A^T(A(\bar{x} + v^n) - b)\right]_j = 0 \tag{69}$$

and

$$\lambda^{2-p} g_p'(\bar{x}_j) + \left[A^T(A\bar{x} - b)\right]_j = 0. \tag{70}$$

Define $\varphi(x) = \lambda g_p'(x_j) + \left[A^T(Ax - b)\right]_j$. All derivatives of φ exist at every $x \neq 0$. Letting (a_i) denote the columns of A, if $i \neq j$, we have $\partial\varphi/\partial x_i(\bar{x}) = \langle a_i, a_j \rangle$, while $\partial\varphi/\partial x_j(\bar{x}) = \lambda g''(\bar{x}_j) + \|a_j\|^2$. Also, $\varphi(\bar{x}) = 0$ and each $\varphi(\bar{x} + v^n) = 0$. By differentiability of φ, we have

$$\frac{\varphi(\bar{x} + v^n) - \varphi(\bar{x}) - \nabla\varphi(\bar{x}) \cdot v^n}{\|v^n\|} \to 0. \tag{71}$$

Since the first two terms of (71) are zero, $\nabla\varphi(\bar{x}) \cdot v^n = o(\|v^n\|)$ as well. By continuity of $\nabla\varphi$ at $\bar{x}$, it is straightforward to show that $\nabla\varphi(\bar{x} + v^n) \cdot v^n = o(\|v^n\|)$ also.

Now we consider second derivatives. $\partial^2\varphi/\partial x_i \partial x_k(\bar{x}) = 0$, unless $i = k = j$, while $\partial^2\varphi/\partial x_j^2(\bar{x}) = \lambda g_p'''(\bar{x}_j)$. Now by the differentiability of $\nabla\varphi$,

$$\|\nabla\varphi(\bar{x} + v^n) - \nabla\varphi(\bar{x}) - \nabla^2\varphi(\bar{x})\ v^n\| = o(\|v^n\|), \tag{72}$$

so

$$\nabla\varphi(\bar{x} + v^n) \cdot v^n - \nabla\varphi(\bar{x}) \cdot v^n - v^n \cdot \nabla^2\varphi(\bar{x})\ v^n = o(\|v^n\|^2). \tag{73}$$

But from the above we have that the first two terms are $o(\|v^n\|^2)$, so $v^n \cdot \nabla^2\varphi(\bar{x})\ v^n = o(\|v^n\|^2)$ as well. But this is $\lambda g_p'''(\bar{x}_j)(v_j^n)^2$; since $(v_j^n)^2/\|v^n\|^2$ does not tend to zero by choice of j, it must be that $g_p'''(\bar{x}_j) = 0$, a contradiction.

Thus we must have $\bar{x}_j = 0$. By choice of j, infinitely many $v_j^n \neq 0$, so by passing to a subsequence we may assume that either all $v_j^n > 0$ or $v_j^n < 0$. By the one-sided differentiability of g_p, we can then repeat the above argument using a smooth extension of g_p to $\mathbb{R}$. Since neither

$g_p'''(0+)$ nor $g_p'''(0-)$ are zero, we will obtain the same contradiction. Therefore E cannot be a continuum, and the sequence $\{x^n\}$ defined by (59) is convergent to a stationary point of F_p. ■

VI. Conclusion

We have shown that for given signals with reasonable sparsity assumptions and a broad class of measurement matrices, the families of penalties corresponding to p-shrinkage and firm thresholding, like the ℓ^p quasinorms, provide a candidate penalty that is able to exactly recover the given data with the given measurement matrix. Further we have shown that these penalties behave well with respect to the addition of noise in the measurements, or only approximately sparse signals (as is often the case in practical settings). Finally, we have shown that iterative p-shrinkage converges to stationary points of the unconstrained energy. These results, together with empirical results (see [23], and Fig. 3), further support the idea that generalized shrinkage penalties can be an advantageous alternative to standard ℓ^1 compressed sensing, or ℓ^p compressed sensing.

Further work could benefit from exploring in what generality these type of results hold. The theory of generalized shrinkage allows for an endless possibility of other shrinkages and penalties to study. Additionally, the methods of proof may apply to compressed sensing relaxations that arise in other ways. Generally speaking, determining conditions under which convex optimization results can be extended to handle nonconvex functionals may continue to be a fruitful area of research. Lastly, we make no claims that the approximations made in these proofs give the tightest results possible, so further refinement of these results may be possible and interesting.

Acknowledgment

The authors would like to acknowledge the support of the UC Lab Fees Research grant 12-LR-236660 in conducting this research. The first author also acknowledges the support of NSF grant no. DGE-1144087, and would like to thank his graduate advisor, Professor Andrea L. Bertozzi, and his other LANL mentor, Brendt Wohlberg, for their guidance. The second author also acknowledges the support of the U.S. Department of Energy through the LANL/LDRD Program.

REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[4] T. T. Cai and A. Zhang, "Sharp RIP bound for sparse signal and low-rank matrix recovery," *Appl. Comput. Harmon. Anal.*, vol. 35, no. 1, pp. 74–93, Jul. 2013.

[5] M. E. Davies and R. Gribonval, "Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2203–2214, May 2009.

[6] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, 2003.

[7] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep. 2002.

[8] S. Foucart, "A note on guaranteed sparse recovery via ℓ_1-minimization," *Appl. Comput. Harmon. Anal.*, vol. 29, no. 1, pp. 97–103, Jul. 2010.

[9] ——, "Sparse recovery algorithms: Sufficient conditions in terms of restricted isometry constants," in *Approximation Theory XIII: San Antonio 2010*, ser. Springer Proceedings in Mathematics, M. Neamtu and L. Schumaker, Eds. New York, NY: Springer, 2012, vol. 13, pp. 65–77.

[10] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[11] ——, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," *Appl. Comput. Harmon. Anal.*, vol. 22, no. 3, pp. 335–355, May 2007.

[12] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 035020, pp. 1–14, 2008.

[13] R. Wu and D.-R. Chen, "The improved bounds of restricted isometry constant for recovery via ℓ_p-minimization," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 6142–6147, Sep. 2013.

[14] A. Aldroubi, X. Chen, and A. Powell, "Stability and robustness of ℓ_q minimization using null space property," in *Proc. 10th Int. Conf. Sampl. Theory Appl.*, Singapore, May 2011.

[15] X. Chen, F. Xu, and Y. Ye, "Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2832–2852, 2010.

[16] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 395–407, May 2009.

[17] M.-J. Lai and L. Y. Liu, "The null space property for sparse recovery from multiple measurement vectors," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 3, pp. 402–406, May 2011.

[18] R. Saab, R. Chartrand, and Özgür Yilmaz, "Stable sparse approximations via nonconvex optimization," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Las Vegas, NV, Apr. 2008, pp. 3885–3888.

[19] Q. Sun, "Sparse approximation property and stable recovery of sparse signals from noisy measurements," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 5086–5090, Oct. 2011.

[20] ——, "Recovery of sparsest signals via ℓ^q-minimization," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 3, pp. 329–341, May 2012.

[21] R. Chartrand, "Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Boston, MA, Jun. 2009.

[22] ——, "Generalized shrinkage and penalty functions," in *Proc. IEEE Glob. Conf. Signal Inform. Process.*, Austin, TX, Dec. 2013.

[23] ——, "Shrinkage mappings and their induced penalty functions," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Florence, Italy, May 2014.

[24] R. Chartrand and B. Wohlberg, "A nonconvex ADMM algorithm for group sparsity with sparse groups," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, Canada, May 2013, pp. 6009–6013.

[25] R. Chartrand, "Nonconvex compressive sensing for X-ray CT: an algorithm comparison," in *Proc. Asilomar Conf. Signal Syst. Comput.*, Pacific Grove, CA, Nov. 2013.

[26] ——, "Nonconvex splitting for regularized low-rank + sparse decomposition," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5810–5819, Nov. 2012.

[27] A. Antoniadis, "Wavelet methods in statistics: Some recent developments and their applications," *Stat. Surv.*, vol. 1, pp. 16–55, 2007.

[28] S. Voronin and R. Chartrand, "A new generalized thresholding algorithm for inverse problems with sparsity constraints," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Vancouver, Canada, May 2013, pp. 1636–1640.

[29] S. Muthukrishnan, *Data streams: Algorithms and applications*. Hanover, MA: now Publishers Inc., 2005.

[30] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, 2009.

[31] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[32] R. Glowinski and A. Marrocco, "Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualité, d'une classe de problems de Dirichlet non lineares," *Revue Française d'Automatique, Informatique, et Recherche Opérationelle*, vol. 9, pp. 41–76, 1975.

[33] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comp. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[34] T. Goldstein and S. Osher, "The split Bregman method for L1 regularized problems," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[36] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vision*, vol. 40, no. 1, pp. 120–145, May 2011.

[37] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[38] R. Chartrand, "Exact reconstructions of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, October 2007.

[39] ——, "Nonconvex compressive sensing and reconstruction of gradient-sparse images: random vs. tomographic Fourier sampling," in *Proc. IEEE Int. Conf. Image Process.*, San Diego, CA, Oct. 2008.

[40] ——, "Nonconvex compressed sensing and error correction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Honolulu, HI, Apr. 2007.

[41] M. Yukawa and S.-i. Amari, "ℓ_p-regularized least squares ($0 < p < 1$) and critical path," 2013, arXiv preprint 1304.6591.

[42] A. Majumdar and R. K. Ward, "An algorithm for sparse MRI reconstruction by Schatten p-norm minimization," *Magn. Reson. Imaging*, vol. 29, no. 3, pp. 408–417, April 2011.

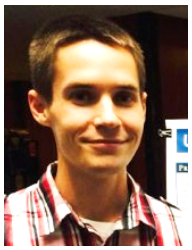
[43] B. Dong and Y. Zhang, "An efficient algorithm for ℓ_0 minimization in wavelet frame based image restoration," *J. Sci. Comput.*, vol. 54, no. 2-3, pp. 350–368, Feb. 2013.

[44] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, vol. 7, no. 4, pp. 855–874, 1997.

[45] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Berlin, Germany: Springer, 1998.

[46] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 211–231, Jan. 2009.

[47] R. Meyer, "Sufficient conditions for the convergence of monotonic mathematical programming algorithms," *J. Comput. Syst. Sci.*, vol. 12, no. 1, pp. 108–121, Feb. 1976.



Joseph Woodworth received a B.Sc.(Hons.) degree in mathematics from the University of Maryland, College Park in 2011. He is a fourth year graduate student at the Department of Mathematics at the University of California, Los Angeles in pursuit of a Ph.D. under the supervision of Professor Andrea L. Bertozzi. His research interests include variational methods, nonlocal operators, image processing, density estimation, networks, dictionary learning, compressed sensing, and optimization.

Rick Chartrand (M'06–SM'12) received a B.Sc.(Hons.) degree in mathematics from the University of Manitoba in 1993, and a Ph.D. in mathematics from the University of California, Berkeley in 1999. He held academic positions at Middlebury College and the University of Illinois at Chicago before coming to Los Alamos National Laboratory in 2003, where he is currently a technical staff member in the Applied Mathematics and Plasma Physics group. His research interests include compressive sensing, nonconvex continuous optimization, image processing, dictionary learning, computing on accelerated platforms, and geometric modeling of high-dimensional data.