SHRINKAGE MAPPINGS AND THEIR INDUCED PENALTY FUNCTIONS

*Rick Chartrand**

Los Alamos National Laboratory
Los Alamos, NM 87545, USA

ABSTRACT

Many optimization problems that are designed to have sparse solutions employ the ℓ^1 or ℓ^0 penalty functions. Consequently, several algorithms for compressive sensing or sparse representations make use of soft or hard thresholding, both of which are examples of shrinkage mappings. Their usefulness comes from the fact that they are the proximal mappings of the ℓ^1 and ℓ^0 penalty functions, meaning that they provide the solution to the corresponding penalized least-squares problem. In this paper, we both generalize and reverse this process: we show that one can begin with any of a wide class of shrinkage mappings, and be guaranteed that it will be the proximal mapping of a penalty function with several desirable properties. Such a shrinkage-mapping/penalty-function pair comes ready-made for use in efficient algorithms. We give an example of such a shrinkage mapping, and use it to advance the state of the art in compressive sensing.

Index Terms— Compressive sensing, sparse representations, shrinkage, nonconvex optimization, alternating direction method of multipliers

1. INTRODUCTION

The ℓ^1 norm is extensively used as a penalty function to enforce sparsity of solutions to optimization problems. Among the very many algorithms that have been developed for ℓ^1 minimization, a common tool [1–4] is the use of soft thresholding, defined componentwise as follows:

$$S_1^\lambda(\mathbf{x})_i = \max\{|x_i| - \lambda, 0\}\,\mathrm{sign}(x_i). \tag{1}$$

The common appearance of S_1 in algorithms can be explained by the fact that it is the *proximal mapping* of the ℓ^1 norm:

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\|_1 + \tfrac{1}{2\lambda}\|\mathbf{w}-\mathbf{x}\|_2^2 = S_1^\lambda(\mathbf{x}). \tag{2}$$

While the convexity of the ℓ^1 norm is a major reason for its common usage, some algorithms attempt to minimize the ℓ^0 penalty function instead. This often leads to the use of hard thresholding [5, 6]:

$$S_{\text{hard}}^\lambda(\mathbf{x})_i = \begin{cases} 0 & \text{if } |x_i| \le \lambda \\ x_i & \text{if } |x_i| > \lambda \end{cases}, \tag{3}$$

in terms of which the proximal mapping of $\|\cdot\|_0$ is expressed:

$$\arg\min_{\mathbf{w}} \|\mathbf{w}\|_0 + \tfrac{1}{2\lambda}\|\mathbf{w}-\mathbf{x}\|_2^2 = S_{\text{hard}}^{\sqrt{2\lambda}}(\mathbf{x}). \tag{4}$$

Knowing the proximal mapping for a penalty function immediately provides possibilities for efficient algorithms for solving the penalized sparse-recovery problem. For example, consider the following problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \tfrac{\mu}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2. \tag{5}$$

This can be used to find a sparse representation of $\mathbf{y}$ with respect to a dictionary A, or to recover the sparse representation of a signal from undersampled data $\mathbf{y}$, where A is the product of the measurement matrix and a dictionary. A simple algorithm for solving (5) is Iterative Soft Thresholding (ISTA) [1]. ISTA combines a gradient descent step for minimizing the fidelity term with a shrinkage step, where soft thresholding is applied. Combining this with Nesterov acceleration [7] produces FISTA ('F' for "fast") [2]. Or, the ℓ^1 norm can be replaced with the ℓ^0 norm, producing Iterative Hard Thresholding [5] instead. These algorithms are simple to implement, and do not require solving a linear system at each iteration, making them scalable to large problems.

Now we generalize our regularization in terms of a penalty function G, while allowing an analysis operator T, chosen so that $T\mathbf{x}$ will be sparse:

$$\min_{\mathbf{x}} G(T\mathbf{x}) + \tfrac{\mu}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2. \tag{6}$$

We suppose we know the proximal mapping S of G, so that

$$\arg\min_{\mathbf{w}} G(\mathbf{w}) + \tfrac{1}{2\lambda}\|\mathbf{w}-\mathbf{x}\|_2^2 = S^\lambda(\mathbf{x}). \tag{7}$$

Adding a splitting variable $\mathbf{w}$, essentially a proxy for $T\mathbf{x}$, makes the utility of the proximal mapping clear:

$$\min_{\mathbf{w},\mathbf{x}} G(\mathbf{w}) + \tfrac{1}{2\lambda}\|\mathbf{w}-T\mathbf{x}\|_2^2 + \tfrac{\mu}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2. \tag{8}$$

*This research was supported by the U.S. Department of Energy through the LANL/LDRD Program, and by the University of California Laboratory Fees Research Program.

If we fix $\mathbf{x}$ and solve (8) for $\mathbf{w}$, our solution is simply $\mathbf{w} = S^\lambda(T\mathbf{x})$. This procedure is a key ingredient of the alternating directions, method of multipliers (ADMM) algorithm [8, 9]. Although the algorithm requires a linear solve at each iteration, the system matrix remains fixed throughout, so that a factorization or preconditioner need only be computed once. In some cases, the system can be solved very efficiently in the Fourier domain.

Thus we see that it is very useful to have a simple, explicit expression for the proximal mapping of a regularizing penalty function. For this reason, we propose to reverse the usual approach of designing a penalty function, then computing its proximal mapping. In the next Section, we generalize the notion of a *shrinkage* mapping, of which soft and hard thresholding are examples. We provide conditions on such a mapping that will guarantee the existence of a corresponding penalty function, and show rigorously that the penalty function will possess properties that are desirable for a sparsity-promoting regularizer. In Sec. 2.1, we provide examples of this procedure, and demonstrate numerically in Sec. 3 that one example gives better compressive sensing results than ever before.

1.1. Relation to Prior Work

A brief, one-page version of portions of this work appeared in [10]. This work provides many more details (including a proof of Theorem 1), as well as additional numerical experiments. The p-shrinkages of Sec. 2.1 appeared first in [11], then with theoretical justification in [12]. These can be seen as ad hoc instances of the general procedure described here. The proximal mapping for the $\ell^{1/2}$ quasinorm was computed in [13] (contemporaneously with [11]), but the approach cannot be extended to general p.

A referee pointed out substantial overlap between [14, Prop. 3.2] and Thm. 1. However, several results of Thm. 1 are not present in [14], or elsewhere to our knowledge.

2. GENERALIZED SHRINKAGE AND PENALTY FUNCTIONS

We proceed to generalize the notion of a shrinkage mapping. It will suffice to work in one dimension: we restrict our attention to shrinkages that are defined componentwise in a homogeneous manner, having the form $S(\mathbf{x})_i = s(|x_i|)\,\mathrm{sign}(x_i)$ for some function s. We call $S : \mathbb{R}^N \to \mathbb{R}^N$ a shrinkage mapping, and $s : \mathbb{R}_+ \to \mathbb{R}_+$ a shrinkage function (where $\mathbb{R}_+ = [0, \infty)$).

The reason we seek proximal mappings that "shrink" is that we want our mapping to sparsify its input. Hence we assume that $s(x) \leq x$ for all $x \in \mathbb{R}_+$. We furthermore assume that s sends some interval of inputs to zero, say $s(x) = 0$ for $x \leq \lambda$, for some $\lambda > 0$. It is also natural to suppose that s is a continuous, increasing function. This turns out to be enough to induce a penalty function with desirable properties.

Theorem 1. *Suppose $s = s^\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, satisfies $x \leq \lambda \Rightarrow s(x) = 0$ for some $\lambda > 0$, is strictly increasing on $[\lambda, \infty)$, and $s(x) \leq x$. Define $S = S^\lambda$ on $\mathbb{R}^n$ by $S(\mathbf{x})_i = s(|x_i|)\,\mathrm{sign}(x_i)$ for each i. Then S is the proximal mapping of a penalty function $G(\mathbf{x}) = \sum_i g(x_i)$ where g is even, nondecreasing and continuous on $[0, \infty)$, differentiable on $(0, \infty)$, and nondifferentiable at 0 with $\partial g(0) = [-1, 1]$. If also $x - s(x)$ is nonincreasing on $[\lambda, \infty)$, then g is concave on $[0, \infty)$ and G satisfies the triangle inequality.*

The proof is carefully presented in the Appendix.

2.1. Examples

The oldest example is soft thresholding, for which $s(x) = \max\{x - \lambda, 0\}$. The G constructed in the proof of Thm. 1 is precisely the ℓ^1 norm, independently of λ.

Let $s(x) = 0$ for $x \leq \lambda$, $s(x) = x$ for $x > \lambda$. Then S is hard thresholding. This s does not satisfy the conditions of the Theorem, but following the construction of G results in $G(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_0$.

A more recent class of examples is given by

$$s_p(x) = \max\{x - \lambda^{2-p}x^{p-1}, 0\}. \tag{9}$$

A variant appeared first in [11], also finding use in [12,15,16]. These reduce to soft thresholding when $p = 1$, and the limiting case as $p \to -\infty$ is hard thresholding. The resulting G_p has no explicit formula for general p. However, we are guaranteed by Thm. 1 that G_p is a suitable penalty function, and having a formula for S_p is more useful for several algorithms than having a formula for G_p. We can plot g_p numerically; see Fig. 1. For $p \neq 0$, g_p approaches $|w|^p/p - C_p$ for large $|w|$, for some constant C_p. In particular, when $p < 0$, it is bounded above by $-C_p$ (which is positive). The smaller the value of p, the slower the growth of g_p, which appears to be a desirable property for recovering sparse signals.

Now we give a new example:

$$s_{\mathrm{SH}}(x) = x\exp\{-\alpha/[\exp(x - \lambda) - 1]^2\} \tag{10}$$

for $x \geq \lambda$, otherwise 0. Here α is an extra tuning parameter. This s satisfies the conditions of the first part of Thm. 1 but not the second part, so the resulting G_{SH} is not necessarily concave and need not satisfy the triangle inequality. The construction of s_{SH} is designed to provide a close approximation of hard thresholding that is smooth (with "SH" standing for "smooth hard"). This is based upon the observation that hard thresholding itself does not give state-of-the-art results, and the conjecture that this is due to its discontinuity. A plot of the numerically computed g_{SH} is in Fig. 1; we see that it grows very slowly.

3. NUMERICAL RESULTS

We give numerical results to illustrate the greatly improved reconstruction performance that is possible using the new

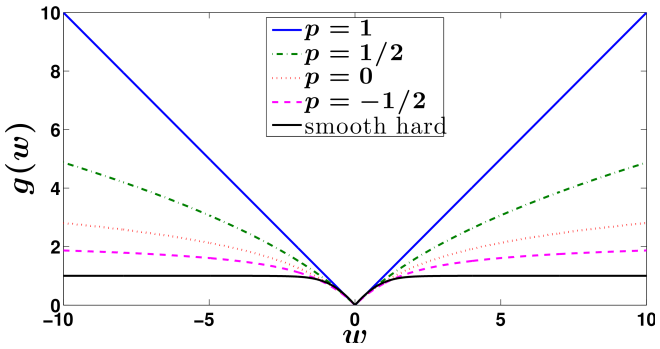


Fig. 1. Plots of g_p for four values of p, and the g_{SH} induced by the smooth approximation of hard thresholding with $\alpha = 10^{-2}$, all using $\lambda = 1$.

penalty function G_{SH} constructed in this paper. The context for the experiment is synthetic magnetic resonance imaging (MRI). We let $\mathbf{x}_0$ be the 256×256 Shepp-Logan phantom, and let A correspond to sampling the discrete Fourier transform (DFT) of $\mathbf{x}_0$ along varying numbers of radial lines. We let $\mathbf{y} = A\mathbf{x}_0$, and consider various examples of G in the equality-constrained version of (8), which can be obtained by incorporating the method of multipliers [17,18] into (8) itself. We let T be a discrete gradient operator, computed with simple forward differencing and periodic boundary conditions.

The case of G being the ℓ^1 norm (thus S being soft thresholding) first appeared in [19]. 18 lines of DFT samples are required for perfect reconstruction. The case of $G(\mathbf{w}) = \|\mathbf{w}\|_{1/2}^{1/2}$ first appeared in [20], with the proximal mapping having been worked out in [13]. With this G, 10 lines are required. This is reduced to 9 lines using $G_{-1/2}$ induced by (9); this was shown in [11]. In this work, we use the ADMM approach described in the Introduction, with $\alpha = 10^{-2}$ and $\lambda = \mu = 10^{10}$. We find 6 lines to be sufficient for perfect reconstruction, greatly improving upon previous results. These results are summarized in Fig. 2. Note that with 5 lines of samples, counting both real and imaginary parts (or equivalently, using the knowledge that the result is real-valued), we have 2830 samples, while $\nabla\mathbf{x}_0$ has 2952 nonzero pixels. Consequently, even a global minimizer using $G(\mathbf{w}) = \|\mathbf{w}\|_0$ cannot equal $\mathbf{x}_0$. Thus the result presented here would appear to be the best possible for this particular example.

We also consider the case of noisy data, by adding noise drawn from the standard normal distribution to each of the real and imaginary parts of $\mathbf{y}$. The resulting data has an SNR of 29.9 dB. Using $\lambda = 10^{4.4}$ and $\mu = 10^{4.7}$, we obtain a reconstruction of 22.4 dB with all features preserved, which is remarkable given the very low sampling (2.6%). For comparison, using G_1 (the ℓ^1 norm) and $G_{-1/2}$ instead give reconstructions of 2.2 dB and 21.3 dB, with some features lost. Results are in Fig. 3.

(a) $p = 1$, 2.2 dB (b) $p = -1/2$, 21.3 dB (c) smooth hard, 22.4 dB

Fig. 3. Reconstructions from noisy MRI data (SNR 29.9 dB). Only the result using G_{SH} preserves all features.

4. CONCLUSIONS

We described an approach for constructing penalty functions by specifying a shrinkage mapping, and requiring the penalty function have the shrinkage as its proximal mapping. Simple conditions on a shrinkage function are shown to lead to desirable properties for the induced penalty function. We gave a new example of a shrinkage mapping, a smooth approximation of hard thresholding, and showed numerical results that demonstrated that the induced penalty function allows reconstruction of sparse images from fewer data than ever before, while retaining reasonable robustness to noise.

5. APPENDIX

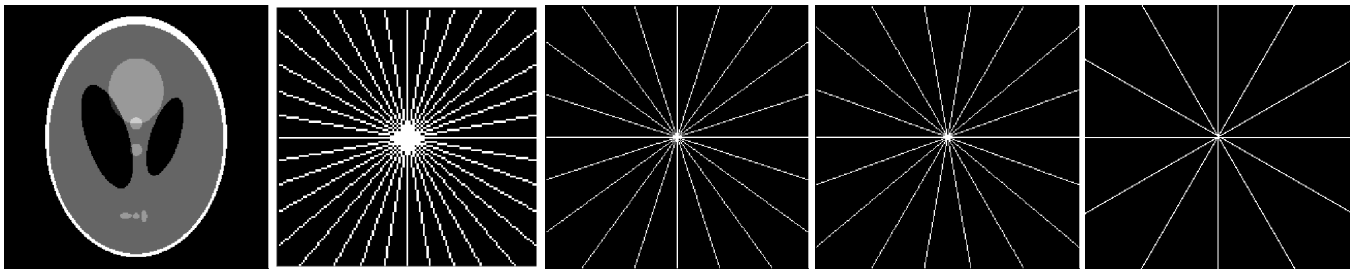
Proof of Theorem 1. We proceed slowly. First, we need to construct our componentwise penalty function g, and then show that it has the claimed properties. We will be employing tools from convex analysis, and a key ingredient is for s to be the derivative of a convex function (which will be more convenient to have defined on all of $\mathbb{R}$). To this end, extend s to a function on $\mathbb{R}$ by requiring $s(-x) = -s(x)$ so that s is odd. Then simply define $f(x) = \int_0^x s(t)\,dt$. Then we have $f' = s$, and thus f is a C^1 function. Now we bring in convex duality, in particular the Legendre-Fenchel transform. Applied to f, this produces the convex conjugate f^*, defined by:

$$f^*(w) = \sup_x wx - f(x). \tag{11}$$

See Fig. 4. We will use w for our dual variable and x for our primal variable. Given a w, we consider the line wx through the origin, and find the largest vertical difference (including sign) between this line and the graph of f. The value of this difference is $f^*(w)$. The place (or set of places) where the largest difference occurs is also important. We will use the following relations, from [21, Prop. 11.3]:

$$\arg\max_x wx - f(x) = \partial f^*(w) \tag{12}$$

$$\arg\max_w wx - f^*(w) = \partial f(x). \tag{13}$$



(a) test image (b) ℓ^1 norm: 18 lines (c) $\ell^{1/2}$ norm: 10 lines (d) $G_{-1/2}$: 9 lines (e) G_{SH}: 6 lines

Fig. 2. Reconstructions of the 256×256 Shepp-Logan phantom from samples of its Fourier transform (as in MRI), with different penalty functions composed with a discrete gradient. Shown are the sampling masks for radial sampling using the minimum number of lines necessary for perfect reconstruction, with the new shrinkage presented in this paper giving much better results, optimal for this particular example.

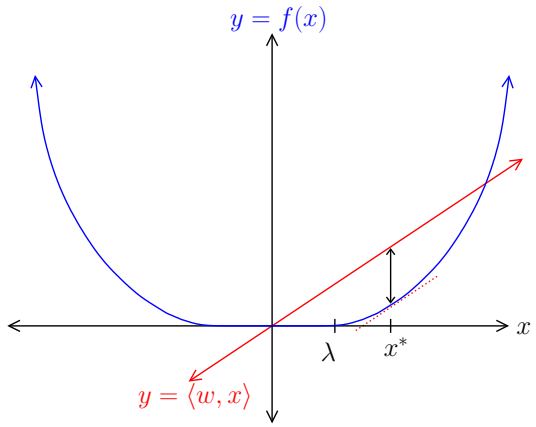


Fig. 4. An example of the function f. Its convex conjugate at w is the greatest height above the graph of f attained by a line through the origin with slope w.

The subdifferential ∂ reduces to the derivative for differentiable functions; otherwise, it is the set of slopes of "subtangents," or lines tangent and locally not crossing above the graph. For example, $\partial|x|(0) = [-1, 1]$. Note that (13) would follow from (12) if we replaced the final f by f^{**}, but our assumptions on f suffice to give us that $f^{**} = f$.

Now we can define $g(w) = (f^*(w) - w^2/2)/\lambda$. We need to show that it has s as a proximal function. We use (13):

$$\begin{aligned} s(x) = f'(x) &= \arg\max_w wx - f^*(w) \\ &= \arg\max_w wx - \lambda g(w) - w^2/2 \\ &= \arg\min_w g(w) - \tfrac{1}{2\lambda}(w - x)^2, \end{aligned} \tag{14}$$

where the last line is obtained by subtracting $x^2/2$ (which doesn't affect the maximizing w) and dividing by $-\lambda$. Then if we define $G(\mathbf{w}) = \sum_i g(w_i)$, we will have

$$\arg\min_{\mathbf{w}} G(\mathbf{w}) + \tfrac{1}{2\lambda}\|\mathbf{w} - \mathbf{x}\|_2^2 = S(\mathbf{x}), \tag{15}$$

since the optimization problem on the left is separable, meaning its solution can be computed separately for each component. Thus we have established that G is a penalty function having S for a proximal mapping.

Now we need to establish the claimed properties of g. The evenness of g follows from the evenness of f, which follows from s being odd. If $w > 0$, it is evident from Fig. 4 that the maximizer of $wx - f(x)$ is unique, so by (12), $\partial f^*(w)$ is a singleton. This means that $f^*(w)$ is differentiable on $(0, \infty)$, hence so is g. This also gives us continuity of g on $(0, \infty)$, but it is clear from the definitions and Fig. 4 that $f^*(0) = 0$ and $f^*(w) \to 0$ as $w \to 0$. Hence g is continuous at 0 as well. And since $\arg\max_x 0x - f(x) = [-\lambda, \lambda] = \partial f^*(0)$, we have $\partial g(0) = [-1, 1]$.

To show that g is nondecreasing on $(0, \infty)$, we show that $g'(w) = ((f^*)'(w) - w)/\lambda$ is nonnegative there. From (12), we have $(f^*)'(w) = \arg\max_x wx - f(x)$, and need to show that this is at least w. The maximizing x^* satisfies $w = f'(x^*) = s(x^*) \le x^*$, establishing the claim.

Now we suppose $x - s(x)$ is nonincreasing on $[\lambda, \infty)$. We will show that g' is nonincreasing on $(0, \infty)$, which will establish the concavity of g. As before, we have $\lambda g'(w) + w = x^*$ with $s(x^*) = w$. Then $\lambda g'(w) = x^* - s(x^*)$, which is a nonincreasing function of x^*. Since x^* is clearly a nonincreasing function of w (see Fig. 4), we have that g' is nonincreasing.

Lastly, we show that $g(v+w) \le g(v)+g(w)$. If v or w are zero, this is trivial. If $vw < 0$, then $|v + w| < \max\{|v|, |w|\}$. Then

$$g(v + w) = g(|v + w|) < g(\max\{|v|, |w|\}) < g(v) + g(w), \tag{16}$$

the last simply because g is nonnegative. Finally, assume $vw > 0$, and without loss of generality assume that v and w are positive. Since g' is nonincreasing, we have $\int_0^w g'(v + t) - g'(t)\,dt \le 0$. Evaluating the integrals directly makes the desired result fall out. □

6. REFERENCES

[1] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, 2004.

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, pp. 183–202, 2009.

[3] J. Yang, W. Yin, Y. Zhang, and Y. Wang, "A fast algorithm for edge-preserving variational multichannel image restoration," *SIAM J. Imaging Sci.*, vol. 2, pp. 569–592, 2009.

[4] T. Goldstein and S. Osher, "The split Bregman method for L1 regularized problems," *SIAM J. Imaging Sci.*, vol. 2, pp. 323–343, 2009.

[5] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, pp. 265–274, 2009.

[6] B. Dong and Y. Zhang, "An efficient algorithm for l0 minimization in wavelet frame based image restoration," *J. Sci. Comput.*, vol. 54, pp. 350–368, 2013.

[7] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Math. Doklady*, vol. 27, pp. 372–376, 1983.

[8] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comp. Math. Appl.*, 1976.

[9] 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*, 1975.

[10] R. Chartrand, "Generalized shrinkage and penalty functions," in *IEEE Glob. Conf. Signal Inform. Process.*, 2013.

[11] ——, "Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data," in *IEEE Int. Symp. Biomed. Imaging*, 2009.

[12] ——, "Nonconvex splitting for regularized low-rank + sparse decomposition," *IEEE Trans. Signal Process.*, vol. 60, pp. 5810–5819, 2012.

[13] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Neural Information and Processing Systems*, 2009.

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

[15] S. Voronin and R. Chartrand, "A new generalized thresholding algorithm for inverse problems with sparsity constraints," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013.

[16] R. Chartrand and B. Wohlberg, "A nonconvex ADMM algorithm for group sparsity with sparse groups," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013.

[17] M. R. Hestenes, "Multiplier and gradient methods," *J. Optim. Theory Appl.*, vol. 4, pp. 303–320, 1969.

[18] M. Powell, "A method for nonlinear constraints in minimization problems," in *Optimization*, R. Fletcher, Ed. Academic Press, New York, 1969, pp. 283–298.

[19] 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, pp. 489–509, 2006.

[20] R. Chartrand, "Exact reconstructions of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, pp. 707–710, 2007.

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