INEXACT PROXIMAL OPERATORS FOR ℓ_p-QUASINORM MINIMIZATION

Cian O'Brien   *Mark D. Plumbley*

Centre for Vision, Speech and Signal Processing. University of Surrey. United Kingdom.

ABSTRACT

Proximal methods are an important tool in signal processing applications, where many problems can be characterized by the minimization of an expression involving a smooth fitting term and a convex regularization term – for example the classic ℓ_1-Lasso. Such problems can be solved using the relevant proximal operator. Here we consider the use of proximal operators for the ℓ_p-quasinorm where $0 \leq p \leq 1$. Rather than seek a closed form solution, we develop an iterative algorithm using a Majorization-Minimization procedure which results in an inexact operator. Experiments on image denoising show that for $p \leq 1$ the algorithm is effective in the high-noise scenario, outperforming the Lasso despite the inexactness of the proximal step.

***Index Terms*—** Proximal Methods, Compressed Sensing, Sparse Recovery, Majorization-Minimization.

1. INTRODUCTION

Sparse methods have proven to be an effective tool in many signal processing applications ranging from compressed sensing to image denoising. A common aspect of such problems is the requirement for solving an optimization problem, usually involving a smooth fitting term plus a regularization term which encourages some structure on the solution (typically sparsity). A popular choice is the ℓ_1-norm penalty on the solution which encourages the solution vector to be sparse (i.e. contain few non-zero entries). The ℓ_1-norm acts as a convex surrogate to the "true" sparsity penalty which counts the number of non-zero entries, making the optimization problem tractable.

Proximal methods are a class of algorithms which can solve problems involving non-smooth terms such as ℓ_1 regularization [1]. For the ℓ_1-norm, this results in the Iterative Shrinkage and Thresholding Algorithm (ISTA) [2]. Relatively less works have focused on the use of general ℓ_p-norm when p is less than one [3]. However, this case is closer to the true sparse penalty and may lead to improved results in sparse problems. A proximal method has been derived for the special cases $p \in \{0, \frac{1}{2}, \frac{2}{3}\}$ [4], which was further developed to the general case by Chen et al [5]. Reweighted ℓ_1 methods have also been shown to be related to the problem of ℓ_p-norm minimization [6] [7][8].

In this work we develop a simple algorithm for ℓ_p regularized problems using a majorization-maximization technique. By relaxing the need for an exact solution to the proximal problem, we derive an iterative procedure which results in an approximate or inexact proximal operator that works well in practice. Unlike reweighted ℓ_1 approaches, our algorithm is generic and can be used as part of any algorithm requiring the proximal operator such as (accelerated) proximal gradient or the Alternating Direction Method of Multipliers (ADMM) [9] and works for any $p \in (0, 1)$.

2. PROXIMAL METHODS

Consider the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) + \lambda g(\mathbf{x}) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and smooth and $g : \mathbb{R}^n \to \mathbb{R}$ is convex and possibly non-differentiable. Equation (1) represents a wide class of problems found in signal processing and machine learning. Generally, f is taken to be a smooth fitting term (for example the Euclidean distance) which penalizes the difference between the estimated signal and the desired signal. The second function g often assumes the role of a regularizer which promotes some desired structure in the solution. A common choice for g is the ℓ_2-norm which penalizes large values in the solution $\mathbf{x}^\star$, or the ℓ_1-norm which promotes sparsity. The problem of estimating $\mathbf{x}^\star$ under an ℓ_1 penalty is known as the Lasso and has seen widespread use in areas such as feature selection, denoising and compressed sensing.

Proximal methods are a class of algorithms for solving problems with the form specified by (1). Given a convex function g, the proximal operator associated with g is

$$\mathrm{prox}_{\lambda g}(\mathbf{x}) \triangleq \arg\min_{\mathbf{u}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 + \lambda g(\mathbf{u}) \right\}. \tag{2}$$

The theory of proximal operators includes projection operators onto convex sets and can deal with a wide range of regularization terms.

We are interested in proximal operators for the so-called ℓ_p-norms, which include the Lasso as a special case ($p = 1$):

$$g(\mathbf{x}) \;=\; \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}. \tag{3}$$

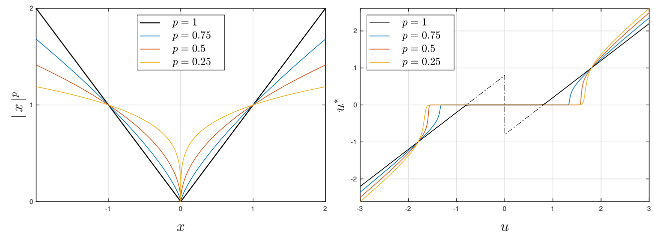

Fig. 1. The ℓ_p penalty for several values of p (left) and their associated approximate shrinkage functions (right). The dashed line is the exact ℓ_1 shrinkage function $u^{(t)} = \text{sign}(u^{(t)})\big(|u^{(t)}| - \lambda\big)$.

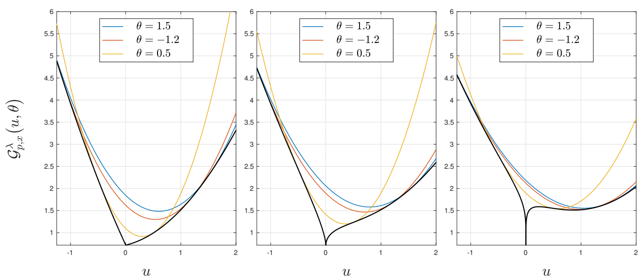

Fig. 2. Majorizing auxiliary functions for $h^\lambda_{p,x}$ with $p = 1$ (left), $p = 0.6$ (middle) and $p = 0.2$ (right). The black line is the graph of $h^\lambda_{p,x}$ for $x = 1.2$ and $\lambda = 1.5$.

Generally, the ℓ_p-norm is only defined for $p \geq 1$. Since it is easier to work without the fraction in the exponent, in the following we take $\|\mathbf{x}\|_p$ to mean (3) raised to the p-th power. For $p \geq 1$ the ℓ_p-norm will be convex (see for example the left panel of Fig. 1, black line). Given the proximal operator $\text{prox}_{\lambda g}(\,\cdot\,)$ for g, the problem given by equation (1) can be solved using proximal gradient descent via the iteration

$$\mathbf{x}^{(t+1)} := \text{prox}_{\lambda g}\big(\mathbf{x}^{(t)} - \eta_t \nabla f_{\mathbf{x}^{(t)}}\big) \tag{4}$$

which for $g(\,\cdot\,) = \|\cdot\|_p$ and $p = 1$ results in the widely used Iterative Shrinkage and Thresholding algorithm.

3. AN INEXACT PROXIMAL OPERATOR

The ℓ_1-norm is the tightest convex approximation to the "true" sparsity penalty which counts the number of non-zero entries in a vector and for this reason there exist many algorithms for ℓ_1 minimization. For $0 \leq p \leq 1$ the function $\|\cdot\|_p$ fails to be a norm, instead defining a quasinorm. In this case $\|\cdot\|_p$ is no longer convex (see left panel in Fig. 1). However, for $0 \leq p \leq 1$ the corresponding p-quasinorm is a closer approximation to the true sparsity penalty.

In this section we propose a method of approximating the solution to (6) in the $0 \leq p \leq 1$ case using a *Majorization-Minimization (MM)* scheme. We first define a function $h^\lambda_{p,x}$: $\mathbb{R} \to \mathbb{R}$

$$h^\lambda_{p,x}(u) \triangleq \frac{1}{2}(u - x)^2 + \lambda |u|^p. \tag{5}$$

Then since $\|\cdot\|_p$ is separable,

$$\mathbf{x}^\star = \text{prox}_{\lambda\|\cdot\|_p}(\mathbf{x}) \iff \forall i,\ x_i^\star = \arg\min_{u\in\mathbb{R}} h^\lambda_{p,x_i}(u) \tag{6}$$

and in order to solve the proximal problem (2), we seek the minimizers of h^λ_{p,x_i} for $i = 1, \ldots, n$.

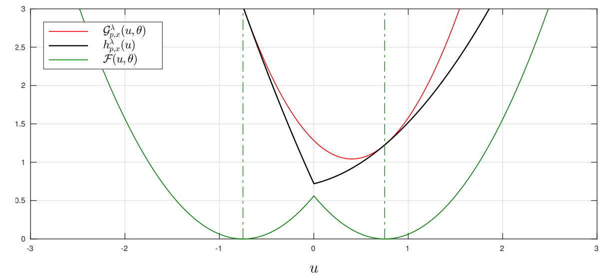

Fig. 3. Difference function $\mathcal{F} = \mathcal{G}^\lambda_{p,x} - h^\lambda_{p,x}$ (see Proposition 3.1). Here $\mathcal{F}$ obtains its minima at $u = \pm\theta = \pm 0.75$ (dashed lines).

3.1. Optimization via Auxiliary Functions

Given a function $h : \mathbb{R} \to \mathbb{R}$ to be minimized, a majorizing auxiliary function $\mathcal{G} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is such that

- $h(u) = \mathcal{G}(u, u) \quad \forall u \in \mathbb{R}$
- $h(u) \leq \mathcal{G}(u, \theta) \quad \forall u \in \mathbb{R}, \forall \theta \in \mathbb{R}.$

The idea behind the MM scheme is that minimization of h can be replaced with iterative minization of $\mathcal{G}$ via

$$u^{(t+1)} := \arg\min_u \ \mathcal{G}(u, u^{(t)}). \tag{7}$$

The two conditions above then guarantee a monotone sequence h so long as $\mathcal{G}(u^{(t+1)}, u^{(t)}) \leq \mathcal{G}(u^{(t)}, u^{(t)})$.

For the ℓ_p-quasinorm proximal operator, we define the following function $\mathcal{G}^\lambda_{p,x} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$\begin{aligned}\mathcal{G}^\lambda_{p,x}(u,\theta) \triangleq\ & \frac{1}{2}\Big((u - x)^2 + \lambda p |\,\theta\,|^{p-2}(u - \theta)^2\Big) + \\ & \lambda\Big(\ \text{sign}(\theta) p u\,|\,\theta\,|^{p-1} - (p-1)|\,\theta\,|^p\Big).\end{aligned} \tag{8}$$

Proposition 3.1. *For fixed* $\mathbf{x} \in \mathbb{R}^n$, $0 \leq p \leq 1$ *and* $\lambda \geq 0$, *the function* $\mathcal{G}^\lambda_{p,x}(u,\theta)$ *defined by (8) is a majorizing auxiliary function for* $h^\lambda_{p,x}(u)$.

Proof. The first condition follows by substituting $\theta = u$ in (8) and using the fact that $\text{sign}(u)u = |u|$.

For the second condition, let $\mathcal{F}(u,\theta) : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be the function given by the difference

$$\mathcal{F}(u,\theta) \triangleq \mathcal{G}^{\lambda}_{p,x}(u,\theta) \,-\, h^{\lambda}_{p,x}(u). \tag{9}$$

First note that $\mathcal{F}$ is symmetric in u and θ, i.e.

$$\mathcal{F}(u,\theta) \,=\, \mathcal{F}(-u,\theta) = \mathcal{F}(u,-\theta), \tag{10}$$

so without loss of generality we can work in the positive half-plane $u \in (0,+\infty)$ by replacing x with $|x|$. For positive θ and u we have

$$\frac{\partial \mathcal{F}(u,\theta)}{\partial u} \,=\, \frac{1}{2}\lambda(p-2)p|u|^{p-3}(\theta-u)(\theta+u) \tag{11}$$

and since $0 \le p \le 1$, the factor $(p-2)$ is negative and the overall quantity is positive or negative depending on the sign of the factor $(\theta-u)$. In general:

$$\begin{aligned} |u| \in \big(0,|\theta|\big) \;&\Longrightarrow\; \frac{\partial \mathcal{F}(u,\theta)}{\partial u} \,<\, 0,\\ |u| \in \big(|\theta|,+\infty\big) \;&\Longrightarrow\; \frac{\partial \mathcal{F}(u,\theta)}{\partial u} \,>\, 0,\\ u=\theta \;&\Longrightarrow\; \frac{\partial \mathcal{F}(u,\theta)}{\partial u} \,=\, 0. \end{aligned} \tag{12}$$

Therefore on $(0,+\infty)\times\mathbb{R}$ and $(-\infty,0)\times\mathbb{R}$, the function $\mathcal{F}(u,\theta)$ is convex and equals zero at its minimum. Furthermore, $\mathcal{F}(u,\theta)$ is symmetric in u about 0 for all $\theta\in\mathbb{R}$. $\square$

Having determined $\mathcal{G}^{\lambda}_{p,x}$ we can define an MM scheme by iteratively solving $u^\star = \arg\min_\theta \mathcal{G}^{\lambda}_{p,x}(u,\theta)$ and setting $\theta = u^\star$. Equating the derivative $\partial_\theta\, \mathcal{G}^{\lambda}_{p,x}(u,\theta)$ with zero, we can find the solution $\theta^\star$ which globally minimizes $\mathcal{G}^{\lambda}_{p,x}(u,\theta)$ and hence $h^{\lambda}_{p,x}(u) \ge h^{\lambda}_{p,x}(\theta^\star)$. For $\mathbf{x}\in\mathbb{R}^n$ and $0\le p\le 1$ this gives a non-linear shrinkage function

$$u \,:=\, \frac{x\theta^2}{p\lambda|\theta|^p+\theta^2} \tag{13}$$

Applying (13) iteratively will monotonically decrease the objective (5). Additionally, we need to check the value of the objective function at the extremal point $u=0$. Comparing $h(u)$ and $h(0)$, we find that

$$h(u) \,<\, h(0) \iff u\big(u-2x\big)+2\lambda|u|^p < 0. \tag{14}$$

Putting all of this together, we define an *inexact* or *approximate* ℓ_p-quasinorm proximal operator:

$$\Big[\mathrm{aprox}_{\lambda\|\cdot\|_{0\le p\le 1}}(\mathbf{x})\Big]_{i=1,\ldots,n} := \begin{cases} u^\star_i & \text{if} \quad u^\star_i\big(u^\star_i-2x_i\big)+2\lambda|u^\star_i|^p<0\\ 0 & \text{otherwise}\end{cases} \tag{15}$$

where $u^\star_i$ is found by iterating (13). The number of iterations used to find $u^\star$ can be adjusted to balance accuracy and speed. Algorithm 1 describes the full proximal operator calculation, while algorithm 2 demonstrates the use of the operator as part of a proximal gradient procedure to minimize a smooth function with ℓ_p-quasinorm regularization. The right panel of Fig 1 shows plots of the shrinkage function (13) after several iterations for different values of p.

Initial numerical experiments in compressed sensing showed that using the proposed approach with small p values alleviates the bias towards zero seen in ℓ_1 sparse coding. The choice of p introduces a tradeoff between *identifiability* and *bias*; smaller values result in less shrinkage of the coefficients, while larger values are better able to recover small signal values. The following result provides a bound in terms of p and λ below which signal components cannot be recovered using ℓ_p-minimization.

Proposition 3.2 (Identifiability)**.** *Let $\lambda>0$ and $0\le p\le 1$ be fixed. A necessary condition for exact recovery of $\mathbf{x}^\star\in\mathbb{R}^n$ is that for all $i\in\{j : x^\star_j \ne 0\}$ in its support*

$$|x^\star_i| \;\ge\; \lambda p \kappa^{\frac{p-1}{2-p}} + \kappa^{\frac{1}{2-p}} + \lambda[\nabla f_{\mathbf{x}^\star}]_i \tag{16}$$

where $\kappa := \lambda p(1-p)$.

Proof. A solution $\mathbf{x}$ of (1) satisfies

$$\mathbf{x} = \mathrm{aprox}_{\lambda\|\cdot\|_{0\le p\le 1}}\big(\mathbf{x}-\lambda\nabla f_{\mathbf{x}}\big) \tag{17}$$

and each x must be either a fixed point of $h^{\lambda}_{p,x}$ or else have $u=0$ (note that $u=0$ is not a fixed point in general). For $x>0$, fixed points of $h^{\lambda}_{p,x}$ occur at the solution to

$$\frac{d}{du}h^{\lambda}_{p,x}(u) \,=\, \lambda p u^{p-1}+u-x \,=\, 0. \tag{18}$$

This derivative is anti-symmetric and convex (resp. concave) for $u>0$ (resp. $u<0$). The minimum of this function is achieved at

$$\operatorname*{arg\,min}_{u>0}\left\{\frac{d}{du}h^{\lambda}_{p,x}(u)\right\} \,=\, \big(\lambda p(1-p)\big)^{\frac{1}{2-p}} \,=\, \kappa^{\frac{1}{2-p}}. \tag{19}$$

Substituting (19) into (18) gives

$$\lambda p\kappa^{\frac{p-1}{2-p}} + \kappa^{\frac{1}{2-p}} - x \tag{20}$$

and this function has zero-values if and only if

$$x \,\ge\, \lambda p\kappa^{\frac{p-1}{2-p}} + \kappa^{\frac{1}{2-p}}. \tag{21}$$

Therefore, if condition (21) is not met for $x>0$ then $h^{\lambda}_{p,x}$ has no fixed points. The inequality in (16) follows by taking $x = x^\star_i - \lambda[\nabla f_{\mathbf{x}^\star}]_i$. By symmetry, a similar argument holds for $x<0$. Exact recovery depends on this condition being met for every component, which completes the proof. $\square$

Algorithm 1 Inexact ℓ_p Proximal Operator

Input: Signal $\mathbf{x} \in \mathbb{R}^n$, $\lambda \geq 0$, $p \in [0,1]$ and iteration count $T > 0$.
  Initialize $u_i^{(1)} := 0$ for $i = 1, \ldots, n$.
  for $t = 1$ to $T - 1$
    Update each $u_i^{(t+1)}$ according to (13).
  end for
  for $i = 1$ to n
    Update each $u_i^{(T)}$ according to (15).
  end for

Algorithm 2 Inexact ℓ_p Proximal Gradient with Linesearch

Input: Objective $F = f + \lambda \|\cdot\|_p$, regularizer $\lambda \geq 0$, $p \in [0,1]$, decay term $c \in (0,1)$ and tolerance $\epsilon > 0$.
  Initialize $\mathbf{x} := \mathbf{0}$.
  while $|F(\mathbf{x}^{(t)}) - F(\mathbf{x}^{(t+1)})| \geq \epsilon$
    $\tilde{\mathbf{x}} := \mathbf{x}^{(t)}$.
    while $f(\tilde{\mathbf{x}}) \leq f(\mathbf{x}^{(t)}) + \nabla f_{\mathbf{x}^{(t)}}^\top (\tilde{\mathbf{x}} - \mathbf{x}^{(t)}) + \frac{1}{2\lambda} f(\tilde{\mathbf{x}} - \mathbf{x}^{(t)})$
      $\tilde{\mathbf{x}} := \text{aprox}_{\lambda \|\cdot\|_p}\left(\mathbf{x}^{(t)} - \lambda \nabla f_{\mathbf{x}^{(t)}}\right)$.
      $\lambda := c\lambda$.
    end while
    $\mathbf{x}^{(t+1)} := \tilde{\mathbf{x}}$.
  end while

4. DICTIONARY LEARNING FOR IMAGE DENOISING

To validate the proposed algorithm, it was tested as part of a dictionary-based image denoising system. Dictionary learning has proven an effect method for a variety of image processing applications [10][11][12]. Here our goal is not to present state-of-the-art denoising results, but to (i) validate the performance of the proposed algorithm and (ii) examine the effect of p on the denoising results.

Given an input image which has been corrupted by noise, dictionary learning proceeds by learning a collection of dictionary atoms which can be used to reconstruct the original signal. First, the input image is divided into a collection of m patches of size $\sqrt{n} \times \sqrt{n}$. The dictionary learning problem is then given by

$$\min_{\mathbf{D} \in \mathbb{S}^{n \times d}, \mathbf{x}_{i=1,\ldots,m}} \sum_{i=1}^{m} \|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda \|\mathbf{x}_i\|_p \tag{22}$$

where $\mathbf{D} \in \mathbb{R}^{n \times d}$ is a matrix whose columns consist of individual dictionary atoms, $\mathbb{S}^{n \times d}$ is the set of all $n \times d$ matrices whose columns live on the unit sphere in $\mathbb{R}^n$ and $\mathbf{x}_{i=1,\ldots,m}$ are the sparse coefficients for the m image patches $\mathbf{y}_{i=1,\ldots,m}$ (arranged as column vectors). After learning, each image patch is reconstructed as $\mathbf{D}\mathbf{z}_i$ and arranged to recover the full denoised image. Since the dictionary learning problem in nonconvex in all of its parameters jointly, we take the standard alternating update approach; for the sparse coding step, we use the inexact proximal gradient algorithm (2) to recover $\mathbf{x}_i$ for each $i \in \{1, \ldots, m\}$. Next, holding the codes fixed, we update the dictionary ([13])

$$\mathbf{D}^{(t+1)} := \mathcal{P}_{\mathbb{S}^{n \times d}}\left(\mathbf{Y}\mathbf{X}^{(t)\top}\left(\mathbf{X}^{(t)}\mathbf{X}^{(t)\top}\right)^{-1}\right) \tag{23}$$

where $\mathcal{P}_{\mathbb{S}^{n \times d}}(\,\cdot\,) : \mathbb{R}^{n \times d} \to \mathbb{S}^{n \times d}$ projects the columns of its argument into $\mathbb{S}^{n \times d}$. This process is then iterated until convergence. For the input signal we used the Barbara test image corrupted with varying amounts of Gaussian noise with variance $\sigma \in \{0.2, 0.3\}$ and a patch size of 8×8. The dictionary size was set to 256 atoms, initialized with exemplars from the training set. We used $p \in \{0.4, 0.7\}$ as well as the Lasso ($p = 1$), which was solved using ISTA. Additionally, we test a continuation strategy where p was initialized to 1 and decayed at each iteration until a final value of 0.3 (last row in Table 1).

p	$\sigma = 0.2$		$\sigma = 0.3$	
	MSE	PSNR	MSE	PSNR
Lasso	3.99×10^{-2}	13.98	7.96×10^{-2}	10.98
0.7	$\mathbf{3.36} \times 10^{-2}$	**14.72**	$\mathbf{7.62} \times 10^{-2}$	**11.17**
0.4	3.88×10^{-2}	14.10	7.84×10^{-2}	11.05
Decay	4.06×10^{-2}	13.91	8.87×10^{-2}	10.51

Table 1. Image denoising results for several noise and p-values.

The results are summarized in Table 1 where we record the performance in terms of Peak Signal to Noise Ratio (PSNR) and Mean Squared Error (MSE). We see that values of $p < 1$ lead to improved performance, particularly in the high noise scenario.

5. CONCLUSION

We have presented an algorithm for solving ℓ_p-norm regularized problems in signal processing. The proposed approach is generic, in the sense that it can be used as part of any algorithm requiring the ℓ_p proximal operator. We find that values of $p < 1$ result in less bias in the solution. The approach was validated on a small dictionary learning and image denoising experiment, where we find that tuning the value of p can improve the performance over the ℓ_1 case.

6. ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-PEOPLE-2013-ITN) under Grant Agreement no. 607290 SpaRTaN.

7. REFERENCES

[1] Neal Parikh, Stephen Boyd, et al., "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[2] Ingrid Daubechies, Michel Defrise, and Christine De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[3] Rick Chartrand and Valentina Staneva, "Restricted isometry properties and nonconvex compressive sensing," *Inverse Problems*, vol. 24, no. 3, pp. 035020, 2008.

[4] Thomas Blumensath, Mehrdad Yaghoobi, and Mike E Davies, "Iterative hard thresholding and l0 regularisation," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 3, pp. III–877.

[5] Feishe Chen, Lixin Shen, and Bruce W Suter, "Computing the proximity operator of the lp norm with $0 < p < 1$," *IET Signal Processing*, vol. 10, no. 5, pp. 557–565, 2016.

[6] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5, pp. 877–905, 2008.

[7] Feishe Chen, *Composite Minimization: Proximity Algorithms and Their Applications*, Ph.D. thesis, Syracuse University, 2015.

[8] Ming-Jun Lai, Yangyang Xu, and Wotao Yin, "Improved iteratively reweighted least squares for unconstrained smoothed \ell_q minimization," *SIAM Journal on Numerical Analysis*, vol. 51, no. 2, pp. 927–957, 2013.

[9] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[10] Ling Shao, Ruomei Yan, Xuelong Li, and Yan Liu, "From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1001–1013, 2014.

[11] Ruomei Yan, Ling Shao, and Yan Liu, "Nonlocal hierarchical dictionary learning using wavelets for image denoising," *IEEE transactions on image processing*, vol. 22, no. 12, pp. 4689–4698, 2013.

[12] Michal Aharon, Michael Elad, and Alfred Bruckstein, "k-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[13] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. IEEE, 1999, vol. 5, pp. 2443–2446.