Linear and Nonlinear Analysis
Yokohama Publishers
ISSN 2188-8167

Volume 4, Number 1, 2018, 135–144

PROPERTIES OF ℓ_p-NORM ERRORS IN SIGNAL RECOVERY

HONG-KUN XU

ABSTRACT. We use the ℓ_p-norm ($1 < p < \infty$) to measure the errors in signal processing. This requires to minimize the ℓ_1-norm regularized pth power of the errors and thus carries the difficulty that the gradient fails to be Lipschitz continuous (when $p \neq 2$), which further makes the proximal gradient algorithm inapplicable. In this paper we present several useful properties of the ℓ_p-norm errors. We also discuss iterative algorithms that can be used to find solutions of the ℓ_1 regularized problems.

1. Introduction

In signal processing theory, a signal $x \in \mathbb{R}^n$ of interest is sampled $m > 1$ times linearly and then recovered from the linear (exact) system

$$Ax = b. \tag{1.1}$$

Here $A \in \mathbb{R}^{m\times n}$ is an $m \times n$ matrix and $b \in \mathbb{R}^m$ is the observation. In compressed sensing [6, 9], $m \ll n$ and a sparse signal x is intended to be recovered. However, samples (or measurements) are taken with noises; in other words, the signal x is to be recovered from the perturbed linear (inexact) system

$$Ax + e = b, \tag{1.2}$$

where e represents noises.

A key issue is in which way the errors $e = b - Ax$ are measured. The most popular way is using the least-squares (i.e., the ℓ_2-norm) to measure the errors [12, 15, 23]:

$$\|e\|_2 = \|Ax - b\|_2. \tag{1.3}$$

This leads to the ℓ_1-norm regularized least-squares minimization problem (for recovering a sparse signal)

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1, \tag{1.4}$$

where $\lambda > 0$ is a regularization parameter. This is equivalent to the lasso of Tibshirani [15] for variable selections (in group lasso [22] as well), and also used in compressed sensing [4–6, 9] to recover the sparsest signal x if the measurement matrix A satisfies the restricted isometry property [3].

2010 *Mathematics Subject Classification.* 49J20, 47J06, 47J25, 49N45.

Key words and phrases. Least-squares, lasso, signal recovery, ℓ_p-norm error, proximal gradient, Frank-Wolfe.

Similarly, the elastic net (EN) of Zou and Hastie [23], i.e., the minimization

$$\min_{x\in\mathbb{R}^n}\left(\frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1+\frac{\gamma}{2}\|x\|_2^2\right) \tag{1.5}$$

is also induced from the ℓ_2-norm errors (1.3). A generalization of EN to p-elastic net (p-EN) can be found in [1].

However, Tropp [16, page 1045] pointed out that "One can imagine situations where the ℓ_2 norm is not the most appropriate way to measure the error in approximating the input signal." He further suggested that it may be more effective to use the convex program $\min \|b-Ax\|_p+\lambda\|x\|_1$, where $p\in[1,\infty]$. To be consistent, we will raise the pth power to the ℓ_p-norm error (so that when $p=2$, our problem exactly reduces to the lasso) and consider the ℓ_1-regularized least pth powered optimization problem

$$\min_{x\in\mathbb{R}^n}\ \frac{1}{p}\|Ax-b\|_p^p+\lambda\|x\|_1 \tag{1.6}$$

for $p\in[1,\infty)$ and

$$\min_{x\in\mathbb{R}^n}\ \|Ax-b\|_\infty+\lambda\|x\|_1. \tag{1.7}$$

The ℓ_1 norm case is studied in [17] and the ℓ_∞ norm case (1.7) in [10], respectively. We will in this paper focus on the ℓ_p norm case for $p\in(1,\infty)$. [Note that ℓ_p-norm regularization is also popularly utilized [1,8,20].]

In this paper we will discuss certain basic properties of the ℓ_p-norm error problem (1.6). We also briefly discuss iterative methods for solving it, including the proximal gradient algorithm and the generalized Frank-Wolfe algorithm.

2. Preliminaries

Let $p\in[1,\infty]$. Recall the ℓ_p norm on $\mathbb{R}^n$ is defined as

$$\|x\|_p=\left(\sum_{i=1}^n|x_i|^p\right)^{\frac{1}{p}}\quad(1\le p<\infty),$$

$$\|x\|_\infty=\max_{1\le i\le n}|x_i|.$$

Note that $(\mathbb{R}^n,\|\cdot\|_p)$ is a Banach space (not Hilbertian unless $p=2$).

2.1. **Duality Maps.** Assume $p\in(1,\infty)$. Recall that the duality map J_p is the (generalized) mapping J_p from $(\mathbb{R}^n,\|\cdot\|_p)$ to its dual space $(\mathbb{R}^n,\|\cdot\|_q)$, with $q=p/(p-1)$, such that

$$\langle x,J_px\rangle=\|x\|^p,\ \ \|J_px\|_q=\|x\|_p^{p-1}$$

for all $x\in\mathbb{R}^n$. [Note: J_p is the identity mapping when $p=2$.] It is known that $J_px=\nabla(\frac{1}{p}\|x\|_p^p)$ and has the expression:

$$(J_px)_i=x_i|x_i|^{p-2},\quad i=1,2,\dots,n.$$

Moreover, J_p is strongly monotone as stated below.

Lemma 2.1. *Assume $p \in (1, \infty)$. Then the duality map J_p is strongly monotone, namely, there exists a constant $c_p > 0$ such that [18]*

$$\langle J_p x - J_p y, x - y\rangle \geq c_p \|x - y\|_p^p, \quad x, y \in \mathbb{R}^n. \tag{2.1}$$

2.2. **Convex Functions and Subdifferential.** Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ be an extended real-valued function. We say that φ is convex [14] if

$$\varphi((1-\lambda)x + \lambda y) \leq (1-\lambda)\varphi(x) + \lambda\varphi(y) \tag{2.2}$$

for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}^n$. We say that φ is strictly convex if the strict inequality in (2.2) holds for all $x \neq y$ and $\lambda \in (0, 1)$ and that φ is proper if there exists at least one $x \in \mathbb{R}^n$ such that $\varphi(x)$ is finite. Recall that φ is said to be lower semicontinuous if $\liminf_{y\to x} \varphi(y) \geq \varphi(x)$ for all $x \in \mathbb{R}^n$. As standard, the symbol $\Gamma_0(\mathbb{R}^n)$ stands for the class of all proper, lower semicontinuous (l.s.c.), convex functions from $\mathbb{R}^n$ to $\overline{\mathbb{R}}$.

The subdifferential of $\varphi \in \Gamma_0(\mathbb{R}^n)$ is the operator $\partial\varphi$ defined by

$$\partial\varphi(x) = \{\xi \in \mathbb{R}^n : \varphi(y) \geq \varphi(x) + \langle \xi, y - x\rangle, \quad y \in \mathbb{R}^n\}, \quad x \in \mathbb{R}^n. \tag{2.3}$$

The inequality in (2.3) is referred to as the subdifferential inequality of φ at x. We say that f is subdifferentiable at x if $\partial\varphi(x)$ is nonempty. It is well-known that for an everywhere finite-valued convex function φ on $\mathbb{R}^n$, φ is everywhere subdifferentiable.

Examples: (i) If $\varphi(x) = |x|$ for $x \in \mathbb{R}$, then $\partial\varphi(0) = [-1, 1]$; (ii) If $\varphi(x) = \|x\|_1$ for $x \in \mathbb{R}^n$, then $\partial\varphi(x)$ is given componentwise by

$$(\partial\varphi(x))_j = \begin{cases} \operatorname{sgn}(x_j), & \text{if } x_j \neq 0, \\ \xi_j, & \text{if } x_j = 0, \end{cases} \quad 1 \leq j \leq n, \tag{2.4}$$

where $\xi_j \in [-1, 1]$ is any number, and 'sgn' is the sign function, that is, for $a \in \mathbb{R}$,

$$\operatorname{sgn}(a) = \begin{cases} 1, & \text{if } a > 0, \\ 0, & \text{if } a = 0, \\ -1, & \text{if } a < 0. \end{cases}$$

[More details about convex analysis can be found in [14].]

2.3. **Proximal Mappings.**

Definition 2.2. Let H be a Hilbert space and let $\Gamma_0(H)$ be the space of convex functions in H that are proper, lower semicontinuous and convex. The proximal operator of φ of order $\lambda > 0$ is defined as [13]

$$\operatorname{prox}_{\lambda\varphi}(x) := \arg\min_{v\in H} \left\{ \varphi(v) + \frac{1}{2\lambda}\|v - x\|^2 \right\}, \quad x \in H.$$

It is not hard to find that if $\varphi(x) = |x|$ (for $x \in \mathbb{R}$) is the absolute value function, then

$$\operatorname{prox}_{\lambda|\cdot|}(x) = \operatorname{sgn}(x)\max\{|x| - \lambda, 0\}.$$

This can be extended to the ℓ_1-norm of $x \in \mathbb{R}^n$ as follows:

$$\operatorname{prox}_{\lambda\|\cdot\|}(x) = (y_1, \ldots, y_n)^\top$$

where $y_i = \operatorname{prox}_{\lambda|\cdot|}(x_i) = \operatorname{sgn}(x_i)\max\{|x_i| - \lambda, 0\}$ for $1 \leq i \leq n$, and the symbol $^\top$ means transpose.

It is also known [7] that proximal mappings are firmly nonexpansive, that is, if we set $T = \mathrm{prox}_{\lambda\varphi}(\cdot)$, where $\varphi \in \Gamma_0(H)$ and $\lambda > 0$, then

$$\|Tx - Ty\|^2 \le \langle Tx - Ty, x - y\rangle, \quad x, y \in H.$$

In particular, T is nonexpansive, i.e., $\|Tx - Ty\| \le \|x - y\|$ for all $x, y \in H$.

2.4. **Proximal-Gradient Algorithm.** Consider a composite optimization problem of the form in a Hilbert space H:

$$\min_{x\in H} \varphi(x) := f(x) + g(x) \tag{2.5}$$

where $f, g \in \Gamma_0(H)$.

The following equivalence of (2.5) to a fixed point problem is known (cf. [7, 19]).

Proposition 2.3. *Let $\lambda > 0$ and assume f is continuously differentiable. Then x^* is a solution to (2.5) if and only if x^* is a solution to the fixed point problem*

$$x^* = \mathrm{prox}_{\lambda g}(x^* - \lambda\nabla f(x^*)). \tag{2.6}$$

The proximal gradient algorithm for solving (2.5) is a fixed point algorithm defined as follows.

Initializing $x_0 \in H$ and iterating

$$x_{k+1} = \mathrm{prox}_{\lambda_k g}(x_k - \lambda_k\nabla f(x_k)), \tag{2.7}$$

where $\{\lambda_k\}$ is a sequence of positive real numbers.

We have the following convergence result.

Theorem 2.4 ([7, 19]). *Assume (2.5) is solvable and f has a Lipschitz continuous gradient:*

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad x, y \in H. \tag{2.8}$$

Assume, in addition, the stepsize sequence (λ_k) satisfies the condition:

$$0 < \liminf_{k\to\infty} \lambda_k \le \limsup_{k\to\infty} \lambda_k < \frac{2}{L}. \tag{2.9}$$

Then the sequence (x_k) converges weakly to a solution of (2.5).

3. Geometric Properties of ℓ_p-norm errors

Let $\lambda > 0$ and $1 < p < \infty$, and set

$$\varphi_\lambda(x) := \frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1, \quad x \in \mathbb{R}^n. \tag{3.1}$$

Let S_λ be the set of minimizers of φ_λ, i.e.,

$$S_\lambda = \arg\min_{x\in\mathbb{R}^n} \left(\frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1\right).$$

Since φ_λ is continuous, convex, and coercive (i.e., $\varphi_\lambda(x) \to \infty$ as $\|x\|_2 \to \infty$), we find that S_λ is closed, convex, and nonempty.

Proposition 3.1. *Let $\lambda > 0$ and $1 < p < \infty$. We have the following statements.*

(i) *The matrix A and the norm $\|\cdot\|_1$ are constant on S_λ, that is, $Ax_\lambda = A\hat{x}_\lambda$ and $\|x_\lambda\|_1 = \|\hat{x}_\lambda\|_1$ for $x_\lambda, \hat{x}_\lambda \in S_\lambda$. Consequently, we can define the functions ρ and η by*

$$\rho(\lambda) := \|x_\lambda\|_1, \quad \eta(\lambda) := \frac{1}{p}\|Ax_\lambda - b\|_p^p \quad (x_\lambda \in S_\lambda). \tag{3.2}$$

(ii) *$\rho(\lambda)$ is decreasing and continuous in $\lambda > 0$.*
(iii) *$\eta(\lambda)$ is increasing in $\lambda > 0$.*
(iv) *Ax_λ is continuous in $\lambda > 0$.*

Proof. Take $x_\lambda \in S_\lambda$. Using the optimality condition

$$0 \in \partial\varphi_\lambda(x_\lambda) = A^\top J_p(Ax_\lambda - b) + \lambda\partial\|x_\lambda\|_1 \quad \text{or} \quad -\frac{1}{\lambda}A^\top(Ax_\lambda - b) \in \partial\|x_\lambda\|_1,$$

with $A^\top$ the transpose of A, we find that the subdifferential inequality turns out to be

$$\lambda\|x\|_1 \geq \lambda\|x_\lambda\|_1 - \langle J_p(Ax_\lambda - b), A(x - x_\lambda)\rangle, \quad \forall x \in \mathbb{R}^n. \tag{3.3}$$

In particular, we get, for $\hat{x}_\lambda \in S_\lambda$,

$$\lambda\|\hat{x}_\lambda\|_1 \geq \lambda\|x_\lambda\|_1 - \langle J_p(Ax_\lambda - b), A(\hat{x}_\lambda - x_\lambda)\rangle. \tag{3.4}$$

Interchanging x_λ and $\hat{x}_\lambda$ yields

$$\lambda\|x_\lambda\|_1 \geq \lambda\|\hat{x}_\lambda\|_1 - \langle J_p(A\hat{x}_\lambda - b), A(x_\lambda - \hat{x}_\lambda)\rangle. \tag{3.5}$$

Adding up (3.4) and (3.5) yields

$$0 \geq \langle J_p(Ax_\lambda - b) - J_p(A\hat{x}_\lambda - b), (Ax_\lambda - b) - (A\hat{x}_\lambda - b)\rangle \geq c_p\|Ax_\lambda - A\hat{x}_\lambda\|_p^p.$$

Consequently, $A\hat{x}_\lambda = Ax_\lambda$. Moreover, further using (3.4) and (3.5), we immediately get $\|\hat{x}_\lambda\|_1 = \|x_\lambda\|_1$. Therefore, the functions ρ and η defined by (3.2) are well-defined for $\lambda > 0$.

It turns out from (3.3) that, for $x_\beta \in S_\beta$ with $\beta > 0$,

$$\lambda\|x_\beta\|_1 \geq \lambda\|x_\lambda\|_1 - \langle J_p(Ax_\lambda - b), A(x_\beta - x_\lambda)\rangle. \tag{3.6}$$

Similarly, we have (or interchanging λ and β, and x_λ and x_β in (3.6)

$$\beta\|x_\lambda\|_1 \geq \beta\|x_\beta\|_1 - \langle J_p(Ax_\beta - b), A(x_\lambda - x_\beta)\rangle. \tag{3.7}$$

Adding up (3.6) and (3.7) obtains

$$(\lambda - \beta)(\|x_\beta\|_1 - \|x_\lambda\|_1) \geq \langle J_p(Ax_\lambda - b) - J_p(Ax_\beta - b)\rangle \geq c_p\|Ax_\lambda - Ax_\beta\|_p^p. \tag{3.8}$$

It immediately turns out that the function $\lambda \mapsto \|x_\lambda\|_1$ is nonincreasing: $\|x_\beta\|_1 \geq \|x_\lambda\|_1$ for $0 < \beta < \lambda$, namely, $\rho(\lambda)$ is nonincreasing. (3.8) also shows that Ax_γ is continuous, which implies the continuity of $\eta(\lambda)$ for $\lambda > 0$.

To see the increasingness of the function $\eta(\lambda)$, we notice that the fact $x_\lambda \in S_\lambda$ implies for $\beta > 0$

$$\frac{1}{p}\|Ax_\lambda - b\|_p^p + \lambda\|x_\lambda\|_1 \leq \frac{1}{p}\|Ax_\beta - b\|_p^p + \lambda\|x_\beta\|_1$$

which can be rewritten as

$$\frac{1}{p}\|Ax_\lambda - b\|_p^p \leq \frac{1}{p}\|Ax_\beta - b\|_p^p + \lambda(\|x_\beta\|_1 - \|x_\lambda\|_1).$$

Now if $\beta > \lambda > 0$, then as $\|x_\beta\|_1 \le \|x_\lambda\|_1$, we immediately get that $\frac{1}{p}\|Ax_\lambda - b\|_p^p \le \frac{1}{p}\|Ax_\beta - b\|_p^p$. Namely, $\eta(\lambda) \le \eta(\beta)$.

Finally to the continuity of $\rho(\lambda)$ for $\lambda > 0$, we assume $0 < \beta < \lambda$ and take the limit as $\beta \to \lambda$ in (3.6), arriving at (noticing the continuity of Ax_λ)

$$\lambda\rho(\lambda-) = \lambda \lim_{\beta\to\lambda-} \rho(\beta) \ge \lambda\rho(\lambda) - \lim_{\beta\to\lambda-} \langle J_p(Ax_\lambda - b), Ax_\beta - Ax_\lambda\rangle = \lambda\rho(\lambda).$$

Hence, $\rho(\lambda-) \ge \rho(\lambda)$. This suffices to imply the continuity of ρ at $\lambda > 0$ because of the nonincreasingness of ρ. □

Proposition 3.2. *Assume* $S := \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_p^p$ *is nonempty.*

(i) $\lim_{\lambda\to 0} \rho(\lambda) = \min_{x\in S} \|x\|_1$.
(ii) $\lim_{\lambda\to 0} \eta(\lambda) = \min_{x\in\mathbb{R}^n} \frac{1}{p}\|Ax - b\|_p^p$.

Proof. To prove (i), we first assert that $\|x_\lambda\|_1 \le \|\tilde{x}\|_1$ for any $\tilde{x} \in S$. As a matter of fact,

$$\begin{aligned} \tfrac{1}{p}\|Ax_\lambda - b\|_2^2 + \lambda\|x_\lambda\|_1 &\le \tfrac{1}{p}\|A\tilde{x} - b\|_p^p + \lambda\|\tilde{x}\|_1 \\ &\le \tfrac{1}{p}\|Ax_\lambda - b\|_p^p + \lambda\|\tilde{x}\|_1. \end{aligned}$$

It turns out that $\|x_\lambda\|_1 \le \|\tilde{x}\|_1$. In particular, $\|x_\lambda\|_1 \le \|x^\dagger\|_1$, where $x^\dagger$ is a minimum-norm element of S, that is, $\|x^\dagger\|_1 = \min_{x\in S} \|x\|_1$.

Assume $\lambda_k \to 0$ is such that $x_{\lambda_k} \to \hat{x}$. Then for any x,

$$\begin{aligned} \tfrac{1}{p}\|A\hat{x} - b\|_p^p &= \lim_{k\to\infty} \tfrac{1}{p}\|Ax_{\lambda_k} - b\|_p^p \\ &= \lim_{k\to\infty} \tfrac{1}{p}\|Ax_{\lambda_k} - b\|_p^p + \lambda_k\|x_{\lambda_k}\|_1 \\ &\le \lim_{k\to\infty} \tfrac{1}{p}\|Ax - b\|_p^p + \lambda_k\|x\|_1 = \tfrac{1}{p}\|Ax - b\|_p^p. \end{aligned}$$

It turns out that $\hat{x}$ solves the least pth-power problem $\min_x \frac{1}{p}\|Ax - b\|_p^p$, that is, $\hat{x} \in S$. Consequently,

$$\lim_{\lambda\to 0} \rho(\lambda) = \lim_{k\to\infty} \rho(\lambda_k) = \lim_{k\to\infty} \|x_{\lambda_k}\|_1 = \|\hat{x}\|_1 \le \|x^\dagger\|_1 = \min_{x\in S} \|x\|_1.$$

This suffices to imply that the conclusion of (i).

To prove (ii) we first notice the boundedness of (x_λ). Next by taking the limit as $\lambda \to 0$ in the inequality

$$\frac{1}{p}\|Ax_\lambda - b\|_p^p + \lambda\|x_\lambda\|_1 \le \frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1, \quad \forall x \in \mathbb{R}^n$$

we obtain

$$\lim_{\eta\to 0} \eta(\lambda) \le \frac{1}{p}\|Ax - b\|_p^p, \quad \forall x \in \mathbb{R}^n.$$

The result in (ii) follows immediately. □

The following result shows that if $\lambda > 0$ is sufficiently big, then the minimization (1.6) has trivial solutions only.

Proposition 3.3. *Assume* $S = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_p^p$ *is nonempty and set*

$$\Delta_p := \sup_{\lambda>0} \|A^\top(J_p(Ax_\lambda) - J_p(Ax_\lambda - b))\|_\infty < \infty. \tag{3.9}$$

If $\lambda > \Delta_p$, *then* $x_\lambda = 0$.

Proof. The optimality condition

$$-A^\top J_p(Ax_\lambda - b) \in \lambda\partial\|x_\lambda\|_1$$

implies that

$$\begin{array}{rcl} -(A^\top(J_p(Ax_\lambda - b)))_i & = & \lambda \cdot \mathrm{sgn}[(x_\lambda)_i], \quad \text{if } (x_\lambda)_i \neq 0, \\ |(A^\top(J_p(Ax_\lambda - b)))_i| & \leq & \lambda, \quad \text{if } (x_\lambda)_i = 0. \end{array}$$

Taking $x = 2x_\lambda$ in the subdifferential inequality (3.3) yields

$$\begin{array}{rcl} \lambda\|x_\lambda\|_1 & \geq & -\langle A^\top J_p(Ax_\lambda - b), x_\lambda\rangle \\ & = & -\sum_{(x_\lambda)_i\neq 0}(A^\top(J_p(Ax_\lambda - b)))_i(x_\lambda)_i \\ & = & \sum_{(x_\lambda)_i\neq 0}\lambda \cdot [\mathrm{sgn}(x_\lambda)]_i(x_\lambda)_i \\ & = & \lambda\sum_{(x_\lambda)_i\neq 0}|(x_\lambda)_i| = \lambda\|x_\lambda\|_1. \end{array}$$

Consequently, we must have

$$\begin{array}{rcl} \lambda\|x_\lambda\|_1 & = & -\langle A^\top J_p(Ax_\gamma - b), x_\lambda\rangle = -\langle J_p(Ax_\lambda) - b, Ax_\lambda\rangle \\ & = & \langle J_p(Ax_\lambda) - J_p(Ax_\lambda - b), Ax_\lambda\rangle - \|Ax_\lambda\|_p^p \\ & \leq & \langle A^\top(J_p(Ax_\lambda) - J_p(Ax_\lambda - b)), x_\lambda\rangle \\ & \leq & \|x_\lambda\|_1\|A^\top(J_p(Ax_\lambda) - J_p(Ax_\lambda - b))\|_\infty \\ & \leq & \Delta_p\|x_\lambda\|_1. \end{array}$$

This implies that if $x_\lambda \neq 0$, we must have $\lambda \leq \Delta_p$. This finishes the proof. □

Remark 3.4. When $p = 2$, the duality map $J_p = I$ and $\Delta_2 = \|A^\top b\|_\infty$. Thus $x_\lambda = 0$ whenever $\lambda > \|A^\top b\|_\infty$. This recovers [19, Proposition 2.3]

Proposition 3.5. *Let $\lambda > 0$ and $x_\lambda \in S_\lambda$. Then $\hat{x} \in \mathbb{R}^n$ is a solution of the lasso (1.4) if and only if $A\hat{x} = Ax_\lambda$ and $\|\hat{x}\|_1 \leq \|x_\lambda\|_1$. It turns out that*

$$S_\lambda = x_\lambda + N(A) \cap B_{\rho(\lambda)}, \tag{3.10}$$

where $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ is the null space of A and B_r denotes the closed ball centered at the origin and with radius of $r > 0$. This shows that if we can find one solution to the lasso (1.4), then all solutions are found by (3.10).

Proof. If $A\hat{x} = Ax_\lambda$, then from the relations

$$\begin{array}{rcl} \varphi_\lambda(x_\lambda) & = & \dfrac{1}{p}\|Ax_\lambda - b\|_p^p + \lambda\|x_\lambda\|_1 \\ & \leq & \dfrac{1}{p}\|A\hat{x} - b\|_p^p + \lambda\|\hat{x}\|_1 \\ & = & \dfrac{1}{p}\|Ax_\lambda - b\|_p^p + \lambda\|\hat{x}\|_1, \end{array}$$

we obtain $\|x_\lambda\|_1 \leq \|\hat{x}\|_1$. This together with the assumption that $\|\hat{x}\|_1 \leq \|x_\lambda\|_1$ yields that $\|\hat{x}\|_1 = \|x_\lambda\|_1$ which in turns implies that $\varphi_\lambda(\hat{x}) = \varphi_\lambda(x_\lambda)$ and hence $\hat{x} \in S_\lambda$. □

4. Iterative Methods

Taking $f(x) = \frac{1}{p}\|Ax-b\|_p^p$ and $g(x) = \lambda\|x\|_1$, we rewrite (1.6) as (2.5). Notice that f is diffferentiable with gradient given by (assuming $p \in (1,\infty)$)

$$\nabla f(x) = A^\top J_p(Ax-b). \tag{4.1}$$

4.1. **Proximal-gradient algorithm.** Applying the proximal gradient algorithm (2.7) to (1.6), we get a sequence (x_k) given as follows:

$$x_{k+1} = \text{prox}_{\lambda_k\lambda\|\cdot\|_1}(x_k - \lambda_k A^\top J_p(Ax_k - b)), \tag{4.2}$$

where $x_0 \in \mathbb{R}^n$ is an initial guess and $\{\lambda_k\}$ is a sequence of positive real numbers. However, Theorem 2.4 does not apply to (4.2) because the gradient of f, ∇f, as given in (4.1), fails to be Lipschitz (except for the case of $p=2$). We therefore pose the following open question.

Question: Does the sequence (x_k) generated by the algorithm (4.2) converge to a solution of (1.6)?

4.2. **Generalized Frank-Wolfe Algorithm.** The Frank-Whole algorithm (FWA) [11] provides an iterative algorithm that does not require the gradient to be Lipschitz continuous, and is thus applicable to the optimization (1.6). In fact, a generalization of FWA, called generalized Frank-Whole algorithm (gFWA) [2,21], has recently been developed to treat the composite optimization (2.5). Let C be a closed bounded convex subset of $\mathbb{R}^n$. The gFWA generates a sequence (x_k) via the following iteration process:

$$\begin{cases} \bar{x}_k = \arg\min_{x\in C}\langle f'(x_k), x\rangle + g(x), & \text{(4.3a)} \\ x_{k+1} = x_k + \gamma_k(\bar{x}_k - x_k) & \text{(4.3b)} \end{cases}$$

where $x_0 \in C$ is an initial and $\gamma_k \in [0,1)$ is the stepsize of the kth iteration.

Theorem 4.1 ([21, Theorem 5.2]). *Consider the sequence $\{x_k\}$ generated by the generalized Frank-Wolfe algorithm (4.4). Assume the conditions below are satisfied:*

(i) *the Fréchet derivative f' is uniformly continuous over C;*

(ii) *the stepsizes $\{\gamma_k\} \subset (0,1]$ satisfy the open loop conditions:*

  (C1) $\lim_{k\to\infty}\gamma_k = 0$,

  (C2) $\sum_{k=0}^\infty \gamma_k = \infty$.

Then $\lim_{k\to\infty}\varphi(x_k) = \varphi^* := \inf_C \varphi$, *where* $\varphi = f + g$.

Now assume $S = \arg\min_{x\in\mathbb{R}^n}\|Ax-b\|_p^p$ is nonempty. Then by (3.9) we find that the solution x_λ of (1.6) is trivial (i.e., $x_\lambda = 0$) for all $\lambda > \tilde{\Delta}_p$, where

$$\tilde{\Delta}_p := \sup\{\|A^\top(J_p x - J_p y)\|_\infty : \|x\|_2, \|y\|_2 \le \|A\|_{1,2}|S|_1 + \|b\|_2\},$$

where $|S|_1 := \min\{\|z\|_1 : z \in S\}$ and $\|A\|_{1,2} := \sup\{\|Ax\|_2/\|x\|_1 : x \neq 0\}$ is the $(1,2)$ operator norm of A. It turns out that we can restrict the minimization problem (1.6) to the closed ball B_r for achieving nontrivial solutions. Here $r > 0$ is big enough (i.e. $r > \|A\|_{1,2}|S|_1 + \|b\|_2$). Hence, the gFWA (4.4) applies, where we take $f(x) = \frac{1}{p}\|Ax - b\|_p^p$ and $g(x) = \lambda\|x\|_1$. Note again $f'(x) = A^\top J_p(Ax - b)$. Consequently, the following result follows immediately from Theorem 4.1.

Theorem 4.2. *Let the sequence $\{x_k\}$ be generated by the generalized Frank-Wolfe algorithm:*

$$\begin{cases} \bar{x}_k = \arg\min_{x \in B_r} \langle A^\top J_p(Ax_k - b), x\rangle + \lambda\|x\|_1, & (4.4a) \\ x_{k+1} = x_k + \gamma_k(\bar{x}_k - x_k) & (4.4b) \end{cases}$$

Assume (γ_k) satisfies the above conditions (C1) and (C2). Then $\lim_{k\to\infty} \varphi_\lambda(x_k) = \min_{\mathbb{R}^n} \varphi_\lambda$, with φ_λ defined in (3.1).

References

[1] N. Altwaijry, S. Chebbi and H. K. Xu, *Properties and splitting methods for the p-elastic net*, Pacific J. Optim. **12** (2016), 801–811.

[2] K. Bredies, D. A. Lorenz and P. Maass , *A generalized conditional gradient method and its connection to an iterative shrinkage method*, Comput. Optim. Appl. **42** (2009), 173–193.

[3] E. J. Candés, *The restricted isometry property and its implications for compressed sensing*, C. R. Acad. Sci. I **346** (2008), 589–592.

[4] E. J. Candés, J. Romberg and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory **52** (2006), 489–509.

[5] E. J. Candés, J. Romberg and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Applied Math. **LIX** (2006), 1207–1223.

[6] E. J. Candés and M. B. Wakin, *An introduction to compressive sampling*, IEEE Signal Processing Magazine **25** (2008), 21–30.

[7] P. L. Combettes and R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul. **4** (2005), 1168–1200.

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

[9] D. L. Donoho, *Compressed sensing*, IEEE Trans. Info. Theory **52** (2006), 1289–1306.

[10] D. L. Donoho and M. Elad, *On the stability of basis pursuit in the presence of noise*, Signal Process **86** (2006), 511–532.

[11] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly **3** (1956), 95–110.

[12] M. Hebiri and S. van de Geer, *The smooth-lasso and other $\ell_1 + \ell_2$-penalized methods*, Electron. J. Statist. **5** (2011), 1184–1226.

[13] J.-J. Moreau, *Proprietes des applications "prox"*, C. R. Acad. Sci. Paris Ser. A Math. **256** (1963), 1069–1071.

[14] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[15] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal Statist. Soc. Ser. B **58** (1996), 267–288.

[16] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, IEEE Transactions on Information Theory **52** (2006), 1030–1051.

[17] J. Wright and Y. Ma, *Dense error correction via ℓ_1-minimization*, IEEE Transactions on Information Theory **56** (2010), 3540–3560.

[18] H. K. Xu, *Inequalities in Banach spaces with applications*, Nonlinear Anal. **16** (1991), 1127–1138.

[19] H. K. Xu, *Properties and iterative methods for the lasso and its variants*, Chin. Ann. Math. **35B** (2014), 501–518.

[20] H. K. Xu, M. A. Alghamdi and N. Shahzad, *Regularization for the split feasibility problem*, J. Nonlinear Convex Anal. **17** (2016), 513–525.
[21] H. K. Xu, *Convergence analysis of the Frank-Wolfe algorithm and its generalization in Banach spaces*, arXiv2043381.
[22] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Royal Statist. Soc. Ser. B **68** (2006), 49–67.
[23] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Statist. Soc. Ser. B **67** (2005), 301–320.

Manuscript received 20 March 2018

H. K. Xu
Department of Mathematics, School of Science, Hangzhou Dianzi University, Hangzhou 310018, China
E-mail address: xuhk@hdu.edu.cn