Fixed Point Theory, 21(2020), No. 2, 805-818
DOI: 10.24193/fpt-ro.2020.2.57
http://www.math.ubbcluj.ro/~nodeacj/sfptcj.html

PROPERTIES AND ITERATIVE METHODS FOR THE ELASTIC NET WITH ℓ_p-NORM ERRORS

LILING WEI* AND HONG-KUN XU**

*School of Science, Hangzhou Dianzi University, Hangzhou, 310018, China
E-mail: wll.1225@foxmail.com

**School of Science, Hangzhou Dianzi University, Hangzhou, 310018, China
E-mail: xuhk@hdu.edu.cn (Corresponding author)

Abstract. The p-elastic net (p-EN) with $1 < p < \infty$ is introduced to recover a sparse signal $x \in \mathbb{R}^n$ from $m\,(< n)$ linear measurements with noise. The p-EN, which extends the elastic net of Zou and Hastie [23] and was implicitly suggested by Tropp [16], amounts to minimizing the objective function $(1/p)\|Ax - b\|_p^p + \lambda\|x\|_1 + (\mu/2)\|x\|_2^2$ over $x \in \mathbb{R}^n$, where A is the measurement matrix, b is the observation, and $\lambda > 0$, $\mu > 0$ are regularization parameters. Some basic geometric properties of the p-EN such as how the solution curve of the minimization depends on the parameters λ and μ are investigated. Moreover, iterative algorithms such as the proximal-gradient algorithm and the Frank-Wolfe algorithm are studied for solving the p-EN.

Key Words and Phrases: Lasso, compressed sensing, elastic net, ℓ_p-norm error, proximal gradient, Frank-Wolfe.
2010 Mathematics Subject Classification: 49J20, 47J06, 47J25, 47H10, 49N45.

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

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

where e represents noises.

A key issue is in which way the errors $e = Ax - b$ 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] (see also [10]) for variable selections (in group lasso [22] as well), and also used in compressed sensing [4, 5, 6, 9] to recover the sparsest signal x if the measurement matrix A satisfies the restricted isometry property [3] (which will not be formulated here).

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

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

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 $p\in[1,\infty)$.

The ℓ_1 norm case is studied in [17]. 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, 21].]

In this paper we will study the elastic net with ℓ_p-norm errors. More precisely, we will study the optimization below, which we call the elastic net with ℓ_p-norm errors (p-EN for short):

$$\min_{x\in\mathbb{R}^n} \left(\frac{1}{p}\|Ax-b\|_p^p+\lambda\|x\|_1+\frac{\mu}{2}\|x\|_2^2\right) \tag{1.7}$$

We will present certain basic properties of the p-EN and also some iterative methods that can be used to solve it. The extension from EN to p-EN is nontrivial, due to the fact that EN corresponds to optimization methods in Hilbert spaces (the Euclidean norm $\|\cdot\|_2$ is used throughout), while p-EN corresponds to optimization methods in Banach spaces (the space $\mathbb{R}^n$ equipped with ℓ_p-norm $\|\cdot\|_p$ with $p\neq 2$ is no longer Hilbertian). As a consequence, some methods which work for EN would fail to work for p-EN and we have to manipulate cleverly with the generalized duality map J_p which maps $\mathbb{R}^n$ equipped with ℓ_p-norm $\|\cdot\|_p$ to $\mathbb{R}^n$ equipped with ℓ_q-norm $\|\cdot\|_q$, with $q=p/(p-1)$ for $p\in(1,\infty)$. Banach space techniques are needed in our approach to p-EN in the rest of this paper.

2. Preliminaries

We use $\langle\cdot,\cdot\rangle$ to denote the dot product on $\mathbb{R}^n$; namely if $x=(x_1,\cdots,x_n)^t\in\mathbb{R}^n$ and $y=(y_1,\cdots,y_n)^t\in\mathbb{R}^n$ (here ^{t} means transpose), then

$$\langle x,y\rangle=\sum_{i=1}^n x_iy_i.$$

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

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

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

2.1. **Duality Maps.** Assume $p\in(1,\infty)$ and let $q=p/(p-1)$ be the conjugate of p. Recall that the (generalized) duality map J_p maps $(\mathbb{R}^n,\|\cdot\|_p)$ to its dual space $(\mathbb{R}^n,\|\cdot\|_q)$ with the properties:

$$\langle x,J_px\rangle=\|x\|_p^p=\|x\|_p\cdot\|J_qx\|_q \text{ and } \|J_px\|_q=\|x\|_p^{p-1}\tag{2.1}$$

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:

$$(J_px)_i=|x_i|^{p-1}\mathrm{sgn}(x_i),\quad i=1,2,\cdots,n.\tag{2.2}$$

Here $\mathrm{sgn}(t)$ is the sign function of $t\in\mathbb{R}$; namely,

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

Moreover, it is known that 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, Corollary 1]

$$\langle J_px-J_py,x-y\rangle\ge c_p\|x-y\|_p^p,\quad x,y\in\mathbb{R}^n.\tag{2.3}$$

2.2. **Subdifferential of Convex Functions.** Let $h:\mathbb{R}^n\to\overline{\mathbb{R}}:=\mathbb{R}\cup\{\infty\}$ be an extended real-valued function. Recall that h is said to be convex [14] if

$$h((1-\lambda)x+\lambda y)\le(1-\lambda)h(x)+\lambda h(y)\tag{2.4}$$

for all $\lambda\in(0,1)$ and $x,y\in\mathbb{R}^n$. When the strict inequality in (2.4) holds for all $x\neq y$ and $\lambda\in(0,1)$, h is said to be strictly convex. As standard, we use $\Gamma_0(\mathbb{R}^n)$ to denote the class of all proper, lower semicontinuous (l.s.c.), convex functions from $\mathbb{R}^n$ to $\overline{\mathbb{R}}$.

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

$$\partial h(x)=\{\xi\in\mathbb{R}^n:h(y)\ge h(x)+\langle\xi,y-x\rangle,\quad y\in\mathbb{R}^n\},\quad x\in\mathbb{R}^n.\tag{2.5}$$

The inequality in (2.5) is referred to as the subdifferential inequality of φ at x. We say that f is subdifferentiable at x if $\partial h(x)$ is nonempty. It is well-known that for an everywhere finite-valued convex function h on $\mathbb{R}^n$, φ is everywhere subdifferentiable. [More details about convex analysis can be found in [14].]

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

$$(\partial h(x))_j = \begin{cases} \operatorname{sgn}(x_j), & \text{if } x_j \neq 0, \\ [-1,1], & \text{if } x_j = 0, \end{cases} \quad 1 \leq j \leq n. \tag{2.6}$$

2.3. **Proximal Mappings.** We need the notion of the proximal mapping of a proper l.s.c. convex function.

Definition 2.2. The proximal mapping of a convex function $h \in \Gamma_0(\mathbb{R}^n)$ of index $\lambda > 0$ is defined as [13]

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

It is not hard to find that if $h(x) = |x|$ (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:

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

where $y_i = \operatorname{prox}_{\lambda|\cdot|}(x_i) = \operatorname{sgn}(x_i)\max\{|x_i| - \lambda, 0\}$ for $1 \leq i \leq n$.

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

$$\|Tx - Ty\|^2 \leq \langle Tx - Ty, x - y\rangle, \quad x, y \in \mathbb{R}^n.$$

In particular, T is nonexpansive, i.e., $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^n$.

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

$$\min_{x \in H} h(x) := f(x) + g(x) \tag{2.7}$$

where $f, g \in \Gamma_0(\mathbb{R}^n)$.

The following equivalence of (2.7) 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 (2.7) if and only if x^* is a solution to the fixed point problem*

$$x^* = \operatorname{prox}_{\lambda g}(x^* - \lambda \nabla f(x^*)). \tag{2.8}$$

The proximal gradient algorithm for solving (2.7) is a fixed point algorithm defined as follows.

Initializing $x_0 \in H$ and iterating

$$x_{k+1} = \mathrm{prox}_{\lambda_k g}(x_k - \lambda_k \nabla f(x_k)), \tag{2.9}$$

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

We have the following convergence result.

Theorem 2.4. [7, 19] *Assume (2.7) is solvable and f has a Lipschitz continuous gradient:*

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad x, y \in \mathbb{R}^n. \tag{2.10}$$

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

$$0 < \liminf_{k\to\infty} \lambda_k \le \limsup_{k\to\infty} \lambda_k < \frac{2}{L}. \tag{2.11}$$

Then the sequence (x_k) converges to a solution of (2.7).

Lemma 2.5. *Let $1 \le a < \infty$ and $\varepsilon > 0$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous, convex function such that*

$$S_f := \arg\min_{x\in\mathbb{R}^n} f(x) \ne \emptyset. \tag{2.12}$$

Consider the ℓ_a-norm regularized minimization problem

$$\min_{x\in\mathbb{R}^n} f(x) + \frac{\varepsilon}{a}\|x\|_a^a, \tag{2.13}$$

Let S_ε^a be the solution set of (2.13). *Then, for each fixed $1 \le a < \infty$, $\{S_\varepsilon^a\}_{\varepsilon>0}$ is bounded. When $a > 1$, S_ε^a consists of exactly one point, which is denoted as x_ε^a.*

(i) *If $1 < a < \infty$, then $x_\varepsilon^a \to x_0^a$ (as $\varepsilon \to 0$), where x_0^a is the unique point in S_f assuming the minimal ℓ_a-norm; that is, $x_0^a = \arg\min\{\|z\|_a : z \in S_f\}$.*

(ii) *If $a = 1$, then $\lim_{\varepsilon\to 0} \|x_\varepsilon\|_1 = |S_f|_1 := \min_{z\in S_f} \|z\|_1$, where $x_\varepsilon \equiv x_\varepsilon^1 \in S_\varepsilon^1$ for $\varepsilon > 0$. Thus, each cluster point of (x_ε) (as $\varepsilon \to 0$) assumes minimal ℓ_1-norm in S_f.*

Proof. We have, for each $z \in S_f$,

$$f(z) + \frac{\varepsilon}{a}\|x_\varepsilon^a\|_a^a \le f(x_\varepsilon^a) + \frac{\varepsilon}{a}\|x_\varepsilon^a\|_a^a \le f(z) + \frac{\varepsilon}{a}\|z\|_a^a. \tag{2.14}$$

It turns out that

$$\|x_\varepsilon^a\|_a \le \|z\|_a \quad (\forall z \in S_f). \tag{2.15}$$

In particular,

$$\|x_\varepsilon^a\|_a \le \min_{z\in S_f} \|z\|_a =: |S_f|_a. \tag{2.16}$$

This verifies that the net $\{x_\varepsilon^a\}_{\varepsilon>0}$ is bounded.

Now assume $\{\varepsilon_k\}$ is a sequence such that $\varepsilon_k \to 0$ and $x_{\varepsilon_k}^a \to x^*$ as $k \to \infty$.

We distinguish two cases.

Case 1: $a > 1$. In this case, there exists a unique point $x_0^a \in S_f$ assuming the minimal ℓ_a-norm in S_f; that is, $\|x_0^a\|_a = \min_{z\in S_f} \|z\|_a = |S_f|_a$.

We observe that a direct consequence of (2.14) (letting $\varepsilon = \varepsilon_k \to 0$) is $f(z) = f(x^*)$ for $z \in S_f$; hence $x^* \in S_f$. Now it turns out from (2.15) that $\|x^*\|_a \le \|x_0^a\|_a$. This

must imply that $x^* = x_0^a$, due to the uniqueness of the minimal ℓ_a-norm element of S_f. This has proved that $x_\varepsilon^a \to x_0^a$ as $\varepsilon \to 0$.
Case 2: $a = 1$. In this case, elements of the minimal ℓ_1-norm of S_f may not be unique due to the fact that the ℓ_1-norm is not strictly convex. Let $x_\varepsilon \in S_\varepsilon^1$. By (2.15), we get $\|x_\varepsilon\|_1 \le \|z\|_1$ for every $z \in S_f$. This implies that $\|x_\varepsilon\|_1 \le |S_f|_1 = \min\{\|z\|_1 : z \in S_f\}$. Repeating the argument of Case 1, we immediately obtain $\|x_\varepsilon\|_1 \to |S_f|_1$ (as $\varepsilon \to 0$) since every cluster point of x_ε assumes the minimal ℓ_1-norm of S_f, i.e., in the set $\arg\min\{\|z\|_1 : z \in S_f\}$. This completes the proof.

3. Geometric properties

Let $1 < p < \infty$ and $\lambda > 0$, $\mu > 0$ be given. Set

$$\varphi_{\lambda,\mu}(x) := \frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1 + \frac{\mu}{2}\|x\|_2^2, \quad x \in \mathbb{R}^n. \tag{3.1}$$

Here A is an $m \times n$ matrix and $b \in \mathbb{R}^m$. The following optimization problem is known as the elastic net with ℓ_p-norm errors (p-NE for short):

$$\min_{x\in\mathbb{R}^n} \varphi_{\lambda,\mu}(x) = \frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1 + \frac{\mu}{2}\|x\|_2^2. \tag{3.2}$$

Since $\varphi_{\lambda,\mu}$ is continuous, strictly convex, and coercive (i.e., $\varphi_{\lambda,\mu}(x) \to \infty$ as $\|x\|_2 \to \infty$), $\varphi_{\lambda,\mu}$ has a unique minimizer, which is denoted as $x_{\lambda,\mu}$; that is,

$$x_{\lambda,\mu} = \arg\min_{x\in\mathbb{R}^n} \left(\frac{1}{p}\|Ax - b\|_p^p + \lambda\|x\|_1 + \frac{\mu}{2}\|x\|_2^2\right). \tag{3.3}$$

We now discuss some properties of the minimizer $x_{\lambda,\mu}$ as a function defined on the domain $D := \{(\lambda,\mu) : \lambda > 0,\ \mu > 0\}$. Observe that the subdifferential of $\varphi_{\lambda,\mu}$ is given by

$$\partial\varphi_{\lambda,\mu}(x) := A^t J_p(Ax - b) + \lambda\partial\|x\|_1 + \mu x. \tag{3.4}$$

Here A^t is the transpose of the matrix A and J_p is the generalized duality map of the ℓ_p norm as given in (2.2). This implies that the minimizer $x_{\lambda,\mu}$ satisfies the optimality condition $0 \in A^t J_p(Ax_{\lambda,\mu} - b) + \lambda\partial\|x_{\lambda,\mu}\|_1 + \mu x_{\lambda,\mu}$, or equivalently:

$$-\frac{1}{\lambda}\left(A^t J_p(Ax_{\lambda,\mu} - b) + \mu x_{\lambda,\mu}\right) \in \partial\|x_{\lambda,\mu}\|_1. \tag{3.5}$$

Define a function ρ on D by

$$\rho(\lambda,\mu) = \|x_{\lambda,\mu}\|_1 \tag{3.6}$$

where $x_{\lambda,\mu}$ is defined by (3.3).

We also consider the least-pth power problem:

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_p^p. \tag{3.7}$$

Let S_p denote the set of solutions of (3.7). Namely,

$$S_p = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_p^p. \tag{3.8}$$

Proposition 3.1. *Let $(\lambda,\mu) \in D$ and fix $1 < p < \infty$. Then $(x_{\lambda,\mu})_{(\lambda,\mu)\in D}$ is bounded if and only if S_p is nonempty.*

Proof. Assume $S_p \neq \emptyset$. Consider the ℓ_1-norm regularized optimization problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{p}\|Ax-b\|_p^p + \lambda\|x\|_1. \tag{3.9}$$

We use S_p^λ to denote the set of solutions of (3.9). That is,

$$S_p^\lambda = \arg\min_{x\in\mathbb{R}^n} \frac{1}{p}\|Ax-b\|_p^p + \lambda\|x\|_1. \tag{3.10}$$

Note that S_p^λ is always nonempty. Applying (2.16) to the case where

$$f(x) := (1/p)\|Ax-b\|_p^p,\ \varepsilon := \lambda \text{ and } a := 1,$$

we obtain that

$$\|x_\lambda\|_1 \le |S_p|_1 = \min_{z\in S_p} \|z\|_1, \quad x_\lambda \in S_p^\lambda. \tag{3.11}$$

Applying again (2.16) to the case where $f(x) := (1/p)\|Ax-b\|_p^p + \lambda\|x\|_1$, $\varepsilon := \mu$ and $a := 2$, together with (3.11), we obtain (observing the fact that $\|v\|_2 \le \|v\|_1$ for all $v \in \mathbb{R}^n$)

$$\|x_{\lambda,\mu}\|_2 \le |S_p^\lambda|_2 = \min_{z\in S_p^\lambda} \|z\|_2 \le \min_{z\in S_p^\lambda} \|z\|_1 = |S_p^\lambda|_1 \le |S_p|_1. \tag{3.12}$$

Hence, $(x_{\lambda,\mu})$ is bounded.

Conversely, assume $(x_{\lambda,\mu})$ is bounded. Taking positive sequences (λ_k) and (μ_k) with the properties: $\lambda_k \to 0$, $\mu_k \to 0$, and $x_{\lambda_k,\mu_k} \to \hat{x}$ (as $k \to \infty$). By the definition (3.3), we get

$$\frac{1}{p}\|Ax_{\lambda_k,\mu_k} - b\|_p^p + \lambda_k\|x_{\lambda_k,\mu_k}\|_1 + \frac{\mu_k}{2}\|x_{\lambda_k,\mu_k}\|_2^2 \le \frac{1}{p}\|Ax-b\|_p^p + \lambda_k\|x\|_1 + \frac{\mu_k}{2}\|x\|_2^2$$

for all $x \in \mathbb{R}^n$ and $k \ge 1$. Upon taking the limit as $k \to \infty$, we obtain

$$\frac{1}{p}\|A\hat{x}-b\|_p^p \le \frac{1}{p}\|Ax-b\|_p^p$$

for all $x \in \mathbb{R}^n$. It turns out that $\hat{x} \in S_p$ and thus $S_p \neq \emptyset$. The proof is complete.

Proposition 3.2. *Fix $1 < p < \infty$ and let $D = \{(\lambda,\mu) : \lambda > 0,\ \mu > 0\}$. Assume $S_p \neq \emptyset$. We have the following statements.*

(i) *$x_{\lambda,\mu}$ is a continuous function of $(\lambda,\mu) \in D$ and uniformly continuous over the subregion $D_{\mu_0} := \{(\lambda,\mu) : \lambda > 0,\ \mu \ge \mu_0\}$ for each fixed $\mu_0 > 0$.*

(ii) *As $\mu \to 0$ (for each fixed $\lambda > 0$), $x_{\lambda,\mu} \to x_\lambda^\dagger$, the unique point in S_p^λ that has minimal ℓ_2-norm, i.e., $x_\lambda^\dagger = \arg\min\{\|z\|_2 : z \in S_p^\lambda\}$. Moreover, as $\lambda \to 0$, every cluster point of $x_\lambda^\dagger$ is a minimal ℓ_1-norm solution of the least-pth-power problem (3.7), i.e., a point in the set $\arg\min_{x\in S_p} \|x\|_1$.*

(iii) *As $\lambda \to 0$ (for each fixed $\mu > 0$), $x_{\lambda,\mu} \to \hat{x}_\mu$, where*

$$\hat{x}_\mu = \arg\min_{x\in\mathbb{R}^n} \left(\frac{1}{p}\|Ax-b\|_p^p + \frac{\mu}{2}\|x\|_2^2\right). \tag{3.13}$$

Moreover, as $\mu \to 0$, $\hat{x}_\mu \to \hat{x}$ which is the minimal ℓ_p-norm solution of (3.7), that is, $\hat{x} = \arg\min_{x\in S_p} \|x\|_p$.

(iv) *$\rho(\lambda,\mu) := \|x_{\lambda,\mu}\|_1$ is decreasing in λ for each given $\mu > 0$.*

(v) $\xi(\lambda,\mu) := \|x_{\lambda,\mu}\|_2$ *is decreasing in* μ *for each given* $\lambda > 0$.

Proof. (i) Using the optimality condition (3.5) and subdifferential inequality, we get

$$\lambda\|x\|_1 \geq \lambda\|x_{\lambda,\mu}\|_1 - \langle A^t J_p(Ax_{\lambda,\mu} - b) + \mu x_{\lambda,\mu}, x - x_{\lambda,\mu}\rangle \tag{3.14}$$

for $x \in \mathbb{R}^n$. It follows that, for $(\lambda',\mu') \in D$,

$$\lambda\|x_{\lambda',\mu'}\|_1 \geq \lambda\|x_{\lambda,\mu}\|_1 - \langle A^t J_p(Ax_{\lambda,\mu} - b) + \mu x_{\lambda,\mu}, x_{\lambda',\mu'} - x_{\lambda,\mu}\rangle. \tag{3.15}$$

Interchanging λ and λ', and μ and μ' yields

$$\lambda'\|x_{\lambda,\mu}\|_1 \geq \lambda'\|x_{\lambda',\mu'}\|_1 - \langle A^t J_p(Ax_{\lambda',\mu'} - b) + \mu' x_{\lambda',\mu'}, x_{\lambda,\mu} - x_{\lambda',\mu'}\rangle. \tag{3.16}$$

Adding up (3.15) and (3.16) obtains

$$\begin{aligned}
&(\lambda' - \lambda)(\|x_{\lambda,\mu}\|_1 - \|x_{\lambda',\mu'}\|_1) \\
&\geq \langle A^t J_p(Ax_{\lambda,\mu} - b) + \mu x_{\lambda,\mu} - (A^t J_p(Ax_{\lambda',\mu'} - b) + \mu' x_{\lambda',\mu'}), x_{\lambda,\mu} - x_{\lambda',\mu'}\rangle \\
&= \langle J_p(Ax_{\lambda,\mu} - b) - J_p(Ax_{\lambda',\mu'} - b), A(x_{\lambda,\mu} - b) - A(x_{\lambda',\mu'} - b)\rangle \\
&\quad + \langle \mu x_{\lambda,\mu} - \mu' x_{\lambda',\mu'}, x_{\lambda,\mu} - x_{\lambda',\mu'}\rangle.
\end{aligned}$$

By Lemma 2.1, we get

$$\begin{aligned}
&(\lambda' - \lambda)(\|x_{\lambda,\mu}\|_1 - \|x_{\lambda',\mu'}\|_1) \\
&\geq c_p\|Ax_{\lambda,\mu} - Ax_{\lambda',\mu'}\|_p^p + \langle \mu x_{\lambda,\mu} - \mu' x_{\lambda',\mu'}, x_{\lambda,\mu} - x_{\lambda',\mu'}\rangle \\
&= c_p\|Ax_{\lambda,\mu} - Ax_{\lambda',\mu'}\|_p^p + (\mu - \mu')\langle x_{\lambda,\mu}, x_{\lambda,\mu} - x_{\lambda',\mu'}\rangle + \mu'\|x_{\lambda,\mu} - x_{\lambda',\mu'}\|_2^2 \\
&\geq (\mu - \mu')\langle x_{\lambda,\mu}, x_{\lambda,\mu} - x_{\lambda',\mu'}\rangle + \mu'\|x_{\lambda,\mu} - x_{\lambda',\mu'}\|_2^2.
\end{aligned} \tag{3.17}$$

However, by Proposition 3.1, $\{x_{\lambda,\mu}\}$ is bounded. It thus follows from (3.17) that

$$\|x_{\lambda,\mu} - x_{\lambda',\mu'}\|_2^2 \leq \frac{c}{\mu'}(|\lambda - \lambda'| + |\mu - \mu'|) \tag{3.18}$$

for some constant $c > 0$. This shows that $x_{\lambda,\mu}$ is continuous in D and uniformly continuous in D_{μ_0} for each fixed $\mu_0 > 0$.

(ii) For each fixed $\lambda > 0$, $x_{\lambda,\mu} = \arg\min_{x\in\mathbb{R}^n} f(x) + (\mu/2)\|x\|_2^2$, where

$$f(x) := (1/p)\|Ax - b\|_p^p + \lambda\|x\|_1.$$

Applying Lemma 2.5, we obtain that, as $\mu \to 0$, $x_{\lambda,\mu} \to x_\lambda^\dagger := \arg\min_{z\in S_p^\lambda} \|z\|_2$. Applying Lemma 2.5(ii) to the case where $f(x) = (1/p)\|Ax - b\|_p^p$, we obtain that, as $\lambda \to 0$, $\|x_\lambda^\dagger\|_1 \to |S_p|_1$ and each cluster point of $(x_\lambda^\dagger)$ is of minimal ℓ_1-norm in the set S_p.

(iii) Applying Lemma 2.5 to the case where $f(x) = (1/p)\|Ax - b\|_p^p + (\mu/2)\|x\|_2^2$, we immediately find that $x_{\lambda,\mu}$ converges, as $\lambda \to 0$, to $\hat{x}_\mu$ defined by (3.13). Again by Lemma 2.5(ii), we obtain that $\hat{x}_\mu$ converges, as $\mu \to 0$, to the minimal ℓ_p-norm element of S_p.

(iv) Using the subdifferential inequality (3.14), we get

$$\lambda(\|x_{\lambda',\mu}\|_1 - \|x_{\lambda,\mu}\|_1) \geq \langle A^t J_p(Ax_{\lambda,\mu} - b) + \mu x_{\lambda,\mu}, x_{\lambda,\mu} - x_{\lambda',\mu}\rangle. \tag{3.19}$$

Interchange λ and λ' from (3.19) to get

$$\lambda'(\|x_{\lambda,\mu}\|_1 - \|x_{\lambda',\mu}\|_1) \geq \langle A^t J_p(Ax_{\lambda',\mu} - b) + \mu x_{\lambda',\mu}, x_{\lambda',\mu} - x_{\lambda,\mu}\rangle. \tag{3.20}$$

Adding (3.19) and (3.20) up yields

$$\begin{aligned}&(\lambda-\lambda')(\|x_{\lambda',\mu}\|_1-\|x_{\lambda,\mu}\|_1)\\ &\geq \langle J_p(Ax_{\lambda,\mu}-b)-J_p(Ax_{\lambda',\mu}-b), A(x_{\lambda,\mu}-b)-A(x_{\lambda',\mu}-b)\rangle+\mu\|x_{\lambda,\mu}-x_{\lambda',\mu}\|^2\\ &\geq c_p\|Ax_{\lambda,\mu}-Ax_{\lambda',\mu}\|_p^p+\mu\|x_{\lambda,\mu}-x_{\lambda',\mu}\|^2\geq 0.\end{aligned}$$

This immediately implies that $\|x_{\lambda',\mu}\|_1 \geq \|x_{\lambda,\mu}\|_1$ whenever $\lambda \geq \lambda'$. That is, $\rho(\cdot,\mu)$ is nonincreasing for each fixed $\mu > 0$.

(v) Similarly to (3.19) and (3.20) we have for $\mu > 0$ and $\mu' > 0$,

$$\lambda(\|x_{\lambda,\mu'}\|_1-\|x_{\lambda,\mu}\|_1)\geq \langle A^tJ_p(Ax_{\lambda,\mu}-b)+\mu x_{\lambda,\mu}, x_{\lambda,\mu}-x_{\lambda,\mu'}\rangle$$

and

$$\lambda(\|x_{\lambda,\mu}\|_1-\|x_{\lambda,\mu'}\|_1)\geq \langle A^tJ_p(Ax_{\lambda,\mu'}-b)+\mu' x_{\lambda,\mu'}, x_{\lambda,\mu'}-x_{\lambda,\mu}\rangle.$$

Adding up the last two inequalities yields

$$\begin{aligned}0&\geq \langle J_p(Ax_{\lambda,\mu}-b)-J_p(Ax_{\lambda,\mu'}-b), A(x_{\lambda,\mu}-b)-A(x_{\lambda,\mu'}-b)\rangle\\ &\quad+\langle \mu x_{\lambda,\mu}-\mu' x_{\lambda,\mu'}, x_{\lambda,\mu}-x_{\lambda,\mu'}\rangle\\ &\geq c_p\|Ax_{\lambda,\mu}-Ax_{\lambda,\mu'}\|_p^p+(\mu-\mu')\langle x_{\lambda,\mu}, x_{\lambda,\mu}-x_{\lambda',\mu}\rangle+\mu'\|x_{\lambda,\mu}-x_{\lambda,\mu'}\|_2^2\\ &= c_p\|Ax_{\lambda,\mu}-Ax_{\lambda,\mu'}\|_p^p+(\mu-\mu')(\|x_{\lambda,\mu}\|_2^2-\langle x_{\lambda,\mu}, x_{\lambda',\mu}\rangle)+\mu'\|x_{\lambda,\mu}-x_{\lambda,\mu'}\|_2^2\\ &\geq (\mu-\mu')(\|x_{\lambda,\mu}\|_2^2-\langle x_{\lambda,\mu}, x_{\lambda',\mu}\rangle).\end{aligned}$$

It turns out that if $\mu > \mu'$, then we must have $\|x_{\lambda,\mu}\|_2^2 - \langle x_{\lambda,\mu}, x_{\lambda',\mu}\rangle \leq 0$. Since

$$\langle x_{\lambda,\mu}, x_{\lambda',\mu}\rangle \leq \|x_{\lambda,\mu}\|_2\cdot\|x_{\lambda',\mu}\|_2$$

by the Cauchy-Schwartz inequality, we obtain that $\|x_{\lambda,\mu}\|_2 \leq \|x_{\lambda',\mu}\|_2$.
Namely, $\xi(\lambda,\cdot)$ is nonincreasing for fixed $\lambda > 0$. The proof is complete.

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

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

$$\Delta_p := \sup_{(\lambda,\mu)\in D}\|A^t(J_p(Ax_{\lambda,\mu})-J_p(Ax_{\lambda,\mu}-b))\|_\infty. \tag{3.21}$$

If $\lambda > \Delta_p$, then $x_{\lambda,\mu} = 0$ for all $\mu \in (0,\infty)$.

Remark 3.4. Since $(x_{\lambda,\mu})_{(\lambda,\mu)\in D}$ is bounded, Δ_p is finite. Also, since by (3.12), $\|x_{\lambda,\mu}\|_2 \leq |S_p|_1$ for $(\lambda,\mu)\in D$, we can replace the Δ_p in Proposition 3.3 with $\tilde{\Delta}_p$ which is defined as

$$\tilde{\Delta}_p := \sup_{\|x\|_2\leq|S_p|_1}\|A^t(J_p(Ax)-J_p(Ax-b))\|_\infty\ (\geq \Delta_p). \tag{3.22}$$

Proof of Proposition 3.3. Setting

$$z_{\lambda,\mu} = A^tJ_p(Ax_{\lambda,\mu}-b)+\mu x_{\lambda,\mu},$$

we can rewrite the optimality condition (3.5) as

$$-\frac{1}{\lambda}z_{\lambda,\mu} \in \partial\|x_{\lambda,\mu}\|_1$$

and the subdifferential equality (3.14) turns out to be

$$\lambda\|x\|_1 \geq \lambda\|x_{\lambda,\mu}\|_1 - \langle z_{\lambda,\mu}, x - x_{\lambda,\mu}\rangle \tag{3.23}$$

for $x \in \mathbb{R}^n$. Noticing

$$\begin{array}{lll} -(z_{\lambda,\mu})_i & = \lambda \cdot \operatorname{sgn}[(x_{\lambda,\mu})_i], & \text{if } (x_{\lambda,\mu})_i \neq 0, \\ |(z_{\lambda,\mu})_i| & \leq \lambda, & \text{if } (x_{\lambda,\mu})_i = 0. \end{array}$$

and taking $x = 2x_{\lambda,\mu}$ in (3.23) yields

$$\begin{aligned} \lambda\|x_{\lambda,\mu}\|_1 \geq -\langle z_{\lambda,\mu}, x_{\lambda,\mu}\rangle &= -\sum_{(x_{\lambda,\mu})_i \neq 0} (z_{\lambda,\mu})_i \ (x_{\lambda,\mu})_i \\ &= \lambda \sum_{(x_{\lambda,\mu})_i \neq 0} \operatorname{sgn}[(x_{\lambda,\mu})_i] \ (x_{\lambda,\mu})_i \\ &= \lambda \sum_{(x_{\lambda,\mu})_i \neq 0} |(x_{\lambda,\mu})_i| = \lambda\|x_{\lambda,\mu}\|_1. \end{aligned}$$

Consequently, we must have

$$\begin{aligned} \lambda\|x_{\lambda,\mu}\|_1 &= -\langle z_{\lambda,\mu}, x_{\lambda,\mu}\rangle \\ &= -\langle A^t J_p(Ax_{\lambda,\mu} - b) + \mu x_{\lambda,\mu}, x_{\lambda,\mu}\rangle \\ &= -\langle J_p(Ax_{\lambda,\mu} - b), Ax_{\lambda,\mu}\rangle - \mu\langle x_{\lambda,\mu}, x_{\lambda,\mu}\rangle \\ &= \langle J_p(Ax_{\lambda,\mu}) - J_p(Ax_{\lambda,\mu} - b), Ax_{\lambda,\mu}\rangle - \langle J_p(Ax_{\lambda,\mu}), Ax_{\lambda,\mu}\rangle - \mu\|x_{\lambda,\mu}\|_2^2 \\ &= \langle A^t(J_p(Ax_{\lambda,\mu}) - J_p(Ax_{\lambda,\mu} - b)), x_{\lambda,\mu}\rangle - \|Ax_{\lambda,\mu}\|_p^p - \mu\|x_{\lambda,\mu}\|_2^2 \\ &\leq \langle A^t(J_p(Ax_{\lambda,\mu}) - J_p(Ax_{\lambda,\mu} - b)), x_{\lambda,\mu}\rangle \\ &\leq \|x_{\lambda,\mu}\|_1\|A^t(J_p(Ax_{\lambda,\mu}) - J_p(Ax_{\lambda,\mu} - b))\|_\infty \\ &\leq \Delta_p \cdot \|x_{\lambda,\mu}\|_1. \end{aligned}$$

This implies that if $x_{\lambda,\mu} \neq 0$, we must have $\lambda \leq \Delta_p$. Consequently, if $\lambda > \Delta_p$, we necessarily have $x_{\lambda,\mu} = 0$. This completes the proof.

Remark 3.5. When $p = 2$, the duality map $J_p = I$ and $\Delta_2 = \|A^t b\|_\infty$. Thus $x_{\lambda,\mu} = 0$ whenever $\lambda > \|A^t b\|_\infty$. This particularly recovers [19, Proposition 2.3].

4. Iterative methods

Taking $f(x) = (1/p)\|Ax - b\|_p^p + (\mu/2)\|x\|_2^2$ and $g(x) = \lambda\|x\|_1$, we rewrite (3.2) as the composite optimization (2.7). Notice that f is differentiable with gradient given by (assuming $p \in (1, \infty)$)

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

4.1. **Proximal-gradient algorithm.** Applying the proximal gradient algorithm (2.9) to (3.2), we get a sequence (x_k) given as follows:

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

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 is not applicable 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: Does the sequence (x_k) generated by the algorithm (4.2) converge to the solution of (3.2)?

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, 20], has recently been developed to treat the composite optimization (2.7). Let C be a closed bounded convex subset of $\mathbb{R}^n$ and consider the constrained composite optimization problem

$$\min_{x \in \mathbb{R}^n} \varphi(x) := f(x) + g(x) \tag{4.3}$$

where f and g are convex.

The gFWA generates a sequence (x_k) via the following iteration process:

$$\begin{cases} \bar{x}_k &= \arg\min_{x \in C} \langle f'(x_k), x \rangle + g(x), \\ x_{k+1} &= x_k + \gamma_k(\bar{x}_k - x_k) \end{cases} \tag{4.4}$$

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

Theorem 4.1. ([20, 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 Proposition 3, the solution x_λ of (1.6) is trivial (i.e., $x_\lambda = 0$) for all $\lambda > \tilde{\Delta}_p$, where $\tilde{\Delta}_p$ is defined by (3.22). It turns out that we can restrict the minimization problem (1.6) to the closed ball B_r for achieving nontrivial solutions. Here $r = |S_p|_1$. Hence, the gFWA (4.4) applies, where we take

$$f(x) = \frac{1}{p}\|Ax - b\|_p^p + \frac{\mu}{2}\|x\|_2^2 \text{ and } g(x) = \lambda\|x\|_1.$$

Note again

$$f'(x) = A^t J_p(Ax - b) + \mu x.$$

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:*

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

Let (γ_k) *satisfy the open loop conditions (C1) and (C2). Then*

$$\lim_{k\to\infty} \varphi_{\lambda,\mu}(x_k) = \min_{\mathbb{R}^n} \varphi_{\lambda,\mu},$$

with $\varphi_{\lambda,\mu}$ *defined in (3.2).*

Acknowledgments. The authors were grateful to the anonymous referee for her/his helpful comments and suggestions which improved the presentation of this article. HKX was supported in part by NSFC project number U1811461 and ARC/Discoery project DP200100124.

References

[1] N. Altwaijry, S. Chebbi, H.K. Xu, *Properties and splitting methods for the p-elastic net*, Pacific J. Optim., **12**(2016), no. 4, 801-811.

[2] K. Bredies, D.A. Lorenz, 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, T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, **52**(2006), no. 2, 489-509.

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

[6] E.J. Candés, M.B. Wakin, *An introduction to compressive sampling*, IEEE Signal Processing Magazine, **25**(2008), no. 2, 21-30.

[7] P.L. Combettes, R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., **4**(2005), no. 4, 1168-1200.

[8] I. Daubechies, M. Defrise, 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), no. 4, 1289-1306.

[10] D.L. Donoho, M. Elad, *On the stability of basis pursuit in the presence of noise*, Signal Process., **86**(2006), no. 3, 511-532.

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

[12] M. Hebiri, 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), no. 3, 1030-1051.

[17] J. Wright, Y. Ma, *Dense error correction via* ℓ_1*-minimization*, IEEE Transactions on Information Theory, **56**(2010), no. 7, 3540-3560.

[18] H.K. Xu, *Inequalities in Banach spaces with applications*, Nonlinear Anal., **16**(1991), no. 12, 1127-1138.

[19] H.K. Xu, *Properties and iterative methods for the lasso and its variants*, Chin. Ann. Math. Ser. B, **35**(2014), no. 3, 501-518.

[20] H.K. Xu, *Convergence analysis of the Frank-Wolfe algorithm and its generalization in Banach spaces*, arXiv2043381.

[21] H.K. Xu, M.A. Alghamdi, N. Shahzad, *Regularization for the split feasibility problem*, J. Nonlinear Convex Anal., **17**(2016), no. 3, 513-525.

[22] M. Yuan, Y. Lin, *Model selection and estimation in regression with grouped variables*, J. Royal Statist. Soc. Ser. B, **68**(2006), 49-67.

[23] H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, J. Royal Statist. Soc. Ser. B, **67**(2005), 301-320.

Received: September 15, 2019; Accepted: November 10, 2019.