Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 12, No. 4, pp. 801–811, October 2016

PROPERTIES AND SPLITTING METHOD FOR THE p-ELASTIC NET*

Najla Altwaijry, Souhail Chebbi, and Hong-Kun Xu†

Abstract: The lasso of Tibshirani is a popular model for variable selection. The elastic net of Zou and Hastie applies Tikhonov's regularization to the lasso to break some limitations of the lasso in the case where the number of predictors is much bigger than the number of observations, or where a group of variables have pairwise high correlations. We generalize the elastic net by replacing Tikhonov's regularization with a more general ℓ_p-norm regularization which we refer to as the p-elastic net. One difficulty for dealing with the p-elastic net lies in the fact that the ℓ_p-norm raised to the pth power fails to have a Lipschitz continuous gradient. We discuss some fundamental properties of the p-elastic net, and moreover, provide a splitting proximal algorithm for solving the p-elastic net.

Key words: *lasso, elastic net, proximal algorithm, inverse problem, regularization.*

Mathematics Subject Classification: *47J20, 47J25, 49J40, 49N45, 65J20, 65K10.*

1 Introduction

The lasso of Tibshirani [15] is a popular model for variable selections. It recently also became a fundamental model for the compressed sensing of recovering a highly undersampled signal [1, 2, 7].

The lasso amounts to the minimization problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1, \tag{1.1}$$

where A is an $m\times n$ (real) matrix, $b\in\mathbb{R}^m$, and $\lambda>0$ is a tuning parameter.

A feature of the lasso is the fact that the ℓ_1 norm can promote sparsity of the signal to be recovered. As a matter of fact, if the measurement matrix A satisfies certain property (such as the restricted isometry property [2]), the lasso actually recovers the sparsest signal. The lasso is also known as the basis pursuit denoising by Chen, et al [3] and has many variants in statistical sciences (see [9, 16, 20] and the references therein).

In spite of its great success, the lasso has its limitations [21] in the scenarios where the number of predictors is much bigger than the number of observations, or where a group

*This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-MAT2658-02).

†Corresponding author.

of variables have pairwise high correlations, as depicted in [21]. To better treat the above-mentioned scenarios, Zou and Hastie [21] introduced the elastic net which applies Tikhonov's technique of ℓ_2-norm regularization to the lasso. More precisely, the elastic net (EN) is the minimization problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{2}\|x\|_2^2, \tag{1.2}$$

where $\lambda > 0$ and $\gamma > 0$ are two regularization parameters.

EN not only promotes sparsity due to the ℓ_1-norm penalization, but also finds a minimal ℓ_2-norm solution due to the ℓ_2-penalization so that the coefficients of the solution are not permitted too large [17]. It is shown [21] that EN often outperforms the lasso, and in [17, p. 1045], Tropp called for further study of EN.

In the present paper we will generalize EN by replacing the Euclidean norm with a general ℓ_p norm, which we call the p-elastic net (p-EN). Namely, p-EN refers to the minimization problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{p}\|x\|_p^p, \tag{1.3}$$

where $\lambda > 0$, $\gamma > 0$, $p \in (1,\infty)$, and $\|\cdot\|_p$ stands for the ℓ_p-norm. It is clear that 2-EN is precisely EN. Since the objective function of p-EN (1.3) is strictly convex, there exists a unique solution which is denoted $x_{\lambda,\gamma}$. The purpose of this paper is twofold. Firstly, we will study properties of $x_{\lambda,\gamma}$ such as continuity and behavior as $\lambda \to 0$ and $\gamma \to 0$; secondly, we will provide a splitting proximal algorithm for solving p-EN (1.3).

2 Preliminaries

Recall that, for $1 \le p < \infty$, the ℓ_p norm of a vector $x \in \mathbb{R}^n$ is defined by

$$\|x\|_p = \left(\sum_{j=1}^n |x_j|^p\right)^{\frac{1}{p}}.$$

Recall also that a function $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ is said to be convex if

$$\varphi((1-\lambda)x + \lambda y) \le (1-\lambda)\varphi(x) + \lambda\varphi(y) \tag{2.1}$$

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 (2.1) holds for all $x \ne y$ 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 $\liminf_{y\to x} \varphi(y) \ge \varphi(x)$ for all $x \in \mathbb{R}^n$.

The subdifferential operator of a convex function φ is defined as the operator $\partial\varphi$ given by

$$\partial\varphi(x) = \{\xi \in \mathbb{R}^n : \varphi(y) \ge \varphi(x) + \langle \xi, y-x\rangle, \quad y \in \mathbb{R}^n\}, \quad x \in \mathbb{R}^n. \tag{2.2}$$

The inequality in (2.2) 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

$$(\partial\varphi(x))_j = \begin{cases} \mathrm{sgn}(x_j), & \text{if } x_j \ne 0, \\ \xi_j \in [-1,1], & \text{if } x_j = 0, \end{cases} \quad 1 \le j \le n. \tag{2.3}$$

Here sgn is the sign function; that is, for $a \in \mathbb{R}$,

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

Assume $p \in (1, \infty)$. Then $(\mathbb{R}^n, \|\cdot\|_p)$ is uniformly smooth and its duality map from $(\mathbb{R}^n, \|\cdot\|_p)$ to $(\mathbb{R}^n, \|\cdot\|_q)$ with $q = p/(p-1)$, $J_p(\cdot) = \nabla(\frac{1}{p}\|\cdot\|_p^p)$, possesses the following properties:

(J1) $\langle x, J_p x\rangle = \|x\|_p^p$ and $\|J_p x\|_q = \|x\|_p^{p-1}$ for all $x \in \mathbb{R}^n$, where $q = p/(p-1)$. In fact, we have the representation for J_p as follows:

$$(J_p x)_j = x_j |x_j|^{p-2}, \quad j = 1, 2, \cdots, n.$$

(J2) There exists a constant $c_p > 0$ such that [18]:

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

Consider the constrained minimization problem

$$\min_{x \in C} \varphi(x), \tag{2.4}$$

where C is a closed convex subset of $\mathbb{R}^n$. The following propositions are known in any standard optimization textbook.

Proposition 2.1. *Suppose that $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$ is proper, lower-semicontinuous, convex, and finite on C.*

(i) *If φ is strictly convex, then (2.4) admits at most one solution.*

(ii) *If φ satisfies the coercivity condition:*

$$x \in C, \quad \|x\| \to \infty \quad \Longrightarrow \quad \varphi(x) \to \infty,$$

then there exists at least one solution to (2.4). Therefore, if φ is both strictly convex and coercive, there exists one and only one solution to (2.4).

In particular, the lasso (1.1) is solvable and p-EN (1.3) is uniquely solvable.

Proposition 2.2. *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lower-semicontinuous and convex. The optimality condition for a point $z \in C$ to be a solution to (2.4) is that there holds the variational inequality*

$$\langle \xi, x - z\rangle \geq 0, \quad \forall x \in C,$$

where $\xi \in \partial\varphi(z)$; equivalently, z solves the inclusion

$$0 \in \partial\varphi(z) + N_C(z),$$

where $N_C(z)$ is the normal cone to C at z. Namely, $N_C(z) = \{y \in \mathbb{R}^n : \langle y, w - z\rangle \leq 0, \ \forall w \in C\}$. Consequently, $x \in \mathbb{R}^n$ solves the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} \varphi(x) \tag{2.5}$$

if and only if

$$0 \in \partial\varphi(x). \tag{2.6}$$

Lemma 2.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous, convex function such that*

$$S_f := \arg\min_{x\in\mathbb{R}^n} f(x) \neq \emptyset. \tag{2.7}$$

Let $p \in [1, \infty)$ and consider the regularized minimization

$$\min_{x\in\mathbb{R}^n} f(x) + \frac{\gamma}{p}\|x\|_p^p, \tag{2.8}$$

where $\gamma > 0$ is a regularization parameter. Let $\bar{S}_\gamma$ be the solution set of (2.8). Then $\{\bar{S}_\gamma\}_{\gamma>0}$ is bounded. Moreover, if $p > 1$, then $\bar{S}_\gamma$ consists of exactly one point denoted by $\bar{x}_\gamma$ which is convergent, as $\gamma \to 0$, to $\bar{x}^\dagger := \arg\min_{x\in S_f} \|x\|_p$. If $p = 1$, then every cluster point of $\{\bar{S}_\gamma\}$ as $\gamma \to 0$ is a point of S_f of minimal ℓ_1 norm.

Proof. We can derive that, for each $\bar{x} \in S_f$ and $\gamma > 0$,

$$\begin{aligned} f(\bar{x}) + \frac{\gamma}{p}\|\bar{x}_\gamma\|_p^p &\le f(\bar{x}_\gamma) + \frac{\gamma}{p}\|\bar{x}_\gamma\|_p^p \\ &\le f(\bar{x}) + \frac{\gamma}{p}\|\bar{x}\|_p^p. \end{aligned}$$

It turns out that

$$\|\bar{x}_\gamma\|_p \le \|\bar{x}\|_p. \tag{2.9}$$

In particular, for all $\lambda > 0$,

$$\|\bar{x}_\gamma\|_p \le \|\bar{x}^\dagger\|_p. \tag{2.10}$$

This shows that $\{\bar{x}_\gamma\}$ is bounded. Now if $\{\lambda_k\}$ is a null sequence such that $\lambda_k \to x^*$ as $k \to \infty$, then $x^* \in S_f$ and from (2.10) it easily follows that $\|x^*\|_p \le \|\bar{x}^\dagger\|_p$. Now if $p > 1$, the uniqueness of the ℓ_p norm minimal element of S_f implies that $x^* = \bar{x}^\dagger$; consequently, $\bar{x}_\gamma \to \bar{x}^\dagger$ as $\gamma \to 0$. If $p = 1$, then (2.10) implies that $\|x^*\|_1 \le \|\bar{x}^\dagger\|_1 = \min_{x\in S_f} \|x\|_1$, hence, $x^* \in S_f$ assumes minimal ℓ_1 norm of S_f. □

3 The p-Elastic Net

We consider a natural generalization of the elastic net where the Euclidean norm is replaced with the general ℓ_p-norm, which is, therefore, called the p-elastic net (p-EN). Another interpretation of the p-EN the lasso regularized by the ℓ_p norm. More precisely, the p-EN refers to the following optimization:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{p}\|x\|_p^p, \tag{3.1}$$

where $1 < p < \infty$ is a fixed positive number. It is evident that 2-EN is precisely EN.

It is quite natural to extend EN to p-EN from the mathematical point of view. The introduction of p-EN is also inspired by Tropp [17].

Let $p \in (1, \infty)$ be fixed and set

$$\varphi_{\lambda,\gamma}^p(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{p}\|x\|_p^p, \tag{3.2}$$

where $\lambda > 0$ and $\gamma > 0$ are fixed regularization parameters. It is evident that $\varphi_{\lambda,\gamma}^p$ is continuous and convex. It is also easy to find that the subdifferential operator of $\varphi_{\lambda,\gamma}^p$ is given by

$$\partial\varphi_{\lambda,\gamma}^p(x) = A^t(Ax - b) + \lambda\partial\|x\|_1 + \gamma J_p(x). \tag{3.3}$$

Since $\varphi^p_{\lambda,\gamma}$ is continuous, convex, coercive, by Proposition 2.1(ii), there exists a unique minimizer of $\varphi^p_{\lambda,\gamma}$. Namely, p-EN (3.1) has a unique solution which is denoted $x_{\lambda,\gamma}$. We next discuss some properties of $x_{\lambda,\gamma}$.

We use S_λ to denote the set of solutions of the lasso (1.1). Note that S_λ is closed, convex, and nonempty. Hence, S_λ contains a unique point, denoted $x^\dagger_\lambda$, assuming the minimal ℓ_p norm, that is,

$$\|x^\dagger_\lambda\|_p = \min\{\|x\|_p : x \in S_\lambda\}. \tag{3.4}$$

More properties of S_λ can be found in [19].

Let x_γ, depending on p, be the unique solution to the minimization:

$$\min_{\mathbb{R}^n} \varphi^p_\gamma(x) := \frac{1}{2}\|Ax-b\|_2^2 + \gamma\frac{1}{p}\|x\|_p^p. \tag{3.5}$$

We further assume that the least-squares problem

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 \tag{3.6}$$

is solvable and use S to denote the solution set of (3.6). That is,

$$S := \arg\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 \neq \emptyset. \tag{3.7}$$

Lemma 3.1. *Assume (3.7). Then $\{x_{\lambda,\gamma}\}$ is bounded. Indeed, there exists a constant $c>0$ such that*

$$\sup_{\substack{\lambda>0\\ \gamma>0}} \|x_{\lambda,\gamma}\|_p \le c \inf_{\bar{x}\in S} \|\bar{x}\|_1. \tag{3.8}$$

Proof. Let $c > 0$ satisfy

$$\|x\|_p \le c\|x\|_1 \quad \text{for all } x \in \mathbb{R}^n. \tag{3.9}$$

Apply Lemma 2.3 to the case where $f(x) := \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$ to get

$$\|x_{\lambda,\gamma}\|_p \le \|x_\lambda\|_p \quad \text{for all } x_\lambda \in S_\lambda. \tag{3.10}$$

Apply again Lemma 2.3 to the case where $f(x) := \frac{1}{2}\|Ax-b\|_2^2$ and $p=1$ to get

$$\|x_\lambda\|_1 \le \|\bar{x}\|_1 \quad \text{for all } \bar{x} \in S. \tag{3.11}$$

Combining (3.9)-(3.11) yields (3.8). ☐

Proposition 3.2. *We have*

(i) *$x_{\lambda,\gamma}$ is a continuous function of (λ,γ) over the region $\{(\lambda,\gamma) : \lambda > 0, \gamma > 0\}$ and uniformly continuous over the region $\{(\lambda,\gamma) : \lambda > 0, \gamma \ge \gamma_0\}$ for each fixed $\gamma_0 > 0$.*

(ii) *As $\gamma \to 0$ (for each fixed $\lambda > 0$), $x_{\lambda,\gamma} \to x^\dagger_\lambda$ as defined in (3.4); moreover, as $\lambda \to 0$, every cluster point of $x^\dagger_\lambda$ is an ℓ_1-minimal norm solution of the least-squares problem (3.6), i.e., a point in the set $\arg\min_{x\in S}\|x\|_1$.*

(iii) *As $\lambda \to 0$ (for each fixed $\gamma > 0$), $x_{\lambda,\gamma} \to \hat{x}_\gamma$. Moreover, as $\gamma \to 0$, $\hat{x}_\gamma \to \hat{x}$ which is the ℓ_p minimal norm solution of (3.6), that is, $\hat{x} = \arg\min_{x\in S}\|x\|_p$.*

Proof. By (3.3), the optimality condition

$$0 \in \partial\varphi^p_{\lambda,\gamma}(x_{\lambda,\gamma})$$

turns out to be

$$-\frac{1}{\lambda}\left(A^t(Ax_{\lambda,\gamma}-b)+\gamma J_p(x_{\lambda,\gamma})\right) \in \partial\|x_{\lambda,\gamma}\|_1. \tag{3.12}$$

Apply the subdifferential inequality to get

$$\lambda\|x\|_1 \ge \lambda\|x_{\lambda,\gamma}\|_1 - \langle A^t(Ax_{\lambda,\gamma}-b)+\gamma J_p(x_{\lambda,\gamma}), x-x_{\lambda,\gamma}\rangle \tag{3.13}$$

for $x \in \mathbb{R}^n$. It follows that, for $\lambda' > 0$ and $\gamma' > 0$,

$$\lambda\|x_{\lambda',\gamma'}\|_1 \ge \lambda\|x_{\lambda,\gamma}\|_1 - \langle A^t(Ax_{\lambda,\gamma}-b)+\gamma J_p(x_{\lambda,\gamma}), x_{\lambda',\gamma'}-x_{\lambda,\gamma}\rangle. \tag{3.14}$$

Interchanging γ and γ', and δ and δ' yields

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

Now adding up (3.14) and (3.15) obtains

$$\begin{aligned}
&(\lambda'-\lambda)(\|x_{\lambda,\gamma}\|_1-\|x_{\lambda',\gamma'}\|_1)\\
&\ge \|Ax_{\lambda,\gamma}-Ax_{\lambda',\gamma'}\|_2^2 + \langle \gamma J_p(x_{\lambda,\gamma})-\gamma' J_p(x_{\lambda',\gamma'}), x_{\lambda,\gamma}-x_{\lambda',\gamma'}\rangle \qquad (3.16)\\
&\ge \|Ax_{\lambda,\gamma}-Ax_{\lambda',\gamma'}\|_2^2 + (\gamma-\gamma')\langle J_p(x_{\lambda,\gamma}), x_{\lambda,\gamma}-x_{\lambda',\gamma'}\rangle\\
&\quad + \gamma'\langle J_p(x_{\lambda,\gamma})-J_p(x_{\lambda',\gamma'}), x_{\lambda,\gamma}-x_{\lambda',\gamma'}\rangle\\
&\ge \|Ax_{\lambda,\gamma}-Ax_{\lambda',\gamma'}\|_2^2 + (\gamma-\gamma')\langle J_p(x_{\lambda,\gamma}), x_{\lambda,\gamma}-x_{\lambda',\gamma'}\rangle\\
&\quad + c_p\gamma'\|x_{\lambda,\gamma}-x_{\lambda',\gamma'}\|_p^p. \qquad (3.17)
\end{aligned}$$

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

$$\gamma'\|x_{\lambda,\gamma}-x_{\lambda',\gamma'}\|_p^p \le c(|\lambda-\lambda'|+|\gamma-\gamma'|). \tag{3.18}$$

This suffices to show that $x_{\lambda,\gamma}$ is continuous in the region $\{(\lambda,\gamma) : \lambda > 0, \gamma > 0\}$ and uniformly continuous over the region $\{(\lambda,\gamma) : \lambda > 0, \gamma \ge \gamma_0\}$ for each fixed $\gamma_0 > 0$.

Next, (ii) and (iii) are straightforward consequences of Lemma 2.3. For instance, applying Lemma 2.3 first to the case of $f(x) := \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$ and second to the case of $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ with $p = 1$ immediately yields (ii). (iii) is similarly proved. □

Remark 3.3. It is unclear if $x_{\lambda,\gamma}$ is uniformly continuous over the region $\{(\lambda,\gamma) : \lambda > 0, \gamma > 0\}$. It is also unclear if the double $\lim_{\lambda\to 0,\gamma\to 0} x_{\lambda,\gamma}$ exists or not. In addition, it is interesting to find the monotonicity of the functions $\|x_{\lambda,\gamma}\|_1$, $\|x_{\lambda,\gamma}\|_p$, and $\|Ax_{\lambda,\gamma}-b\|_2^2$ under the partial ordering: $(\lambda,\gamma) \preccurlyeq (\lambda',\gamma')$ if and only if $\lambda \le \lambda'$ and $\gamma \le \gamma'$.

In [19] it is shown that the solutions to the lasso (1.1) are trivial for large enough λ. Below we show that this property is inherited by the p-EN.

Proposition 3.4. *If $\lambda > \|A^tb\|_\infty$, then $x_{\lambda,\gamma} = 0$ for all $\gamma \in (0,\infty)$.*

Proof. By the optimality condition (3.12) and setting

$$z_{\lambda,\gamma} := A^t(Ax_{\lambda,\gamma}-b)+\gamma J_p(x_{\lambda,\gamma}),$$

we have

$$\begin{array}{rcll} -(z_{\lambda,\gamma})_j &=& \lambda \cdot \operatorname{sgn}((x_{\lambda,\gamma})_j), & \text{if } (x_{\lambda,\gamma})_j \neq 0, \\ |(z_{\lambda,\gamma})_j| &\leq& \lambda, & \text{if } (x_{\lambda,\gamma})_j = 0. \end{array}$$

(Here $(z)_j$ is the jth component of a vector $z \in \mathbb{R}^n$.) Substituting $2x_{\lambda,\gamma}$ for x in the subdifferential inequality (3.13) yields

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

It follows that

$$\begin{aligned} \lambda\|x_{\lambda,\gamma}\|_1 &= -\langle z_{\lambda,\gamma}, x_{\lambda,\gamma}\rangle \\ &= -\langle Ax_\gamma - b, Ax_\gamma\rangle - \gamma\langle J_p(x_{\lambda,\gamma}), x_{\lambda,\gamma}\rangle \quad (3.19) \\ &= -\|Ax_{\lambda,\gamma}\|_2^2 + \langle x_{\lambda,\gamma}, A^t b\rangle - \gamma\|x_{\lambda,\gamma}\|_p^p \\ &\leq \langle x_{\lambda,\gamma}, A^t b\rangle \leq \|x_{\lambda,\gamma}\|_1 \|A^t b\|_\infty. \quad (3.20) \end{aligned}$$

Consequently, we must have $\lambda \leq \|A^t b\|_\infty$ from (3.20) should $x_{\lambda,\gamma} \neq 0$. This completes the proof. □

4 A Splitting Method

Splitting methods are popular in solving optimization problems with composite objective functions and monotone operator equations; see [5,6,8,10,13,14] and the references therein.

In this section we provide a splitting method for solving the p-elastic net (p-EN) which is recalled below:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{p}\|x\|_p^p \triangleq \varphi_{\lambda,\gamma}^p(x), \tag{4.1}$$

where $\lambda > 0$, $\gamma > 0$, and $1 < p < \infty$.

Observe that the minimization of p-EN (4.1) is split in the sense that the objective $\varphi_{\lambda,\gamma}^p$ is written as the sum of three simpler convex functions, the first and third being differentiable with gradients $A^t(Ax - b)$ and $\gamma J_p(x)$, respectively. Since J_p is not Lipschitz continuous unless $p = 2$ (see [18, Remark 2, p. 1133]), applying the gradient methods where Lipschitz continuity of gradients is required is therefore difficult. Consequently, we will try the proximal method to solve (4.1).

Consider the composite minimization:

$$\min_{x\in\mathbb{R}^n} f(x) + g(x), \tag{4.2}$$

where f, g are convex functions on $\mathbb{R}^n$.

The proximal algorithm [4,19] is a powerful method to solve (4.2). To state it we need the notion of proximal operators.

Definition 4.1 ([11,12])**.** The proximal operator of a convex function φ defined on $\mathbb{R}^n$ is defined as

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

The proximal algorithm generates a sequence $\{x_k\}$ via the iteration process:

$$x_{k+1} = (\mathrm{prox}_{\lambda_k g} \circ (I - \lambda_k \nabla f))x_k \tag{4.3}$$

where the initial guess x_0 is arbitrarily chosen in $\mathbb{R}^n$ and $\{\lambda_k\}$ is a sequence of positive real numbers. The convergence of this algorithm is given below.

Since the minimization problem (4.2) is equivalent to the following fixed point problem

$$x = (\mathrm{prox}_{\alpha g} \circ (I - \alpha \nabla f))x \tag{4.4}$$

for any $\alpha > 0$, the proximal algorithm (4.3) is a fixed point algorithm. However, its convergence requires Lipschitz continuity of the gradient of f, ∇f.

Theorem 4.2. *Assume (4.2) is solvable. Assume, in addition, that:*

(i) *f is differentiable and the gradient operator ∇f satisfies the Lipschitz continuity condition:*

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

where $L > 0$ is a constant,

(ii) $0 < \liminf_{k\to\infty} \lambda_k \le \limsup_{k\to\infty} \lambda_k < \dfrac{2}{L}$.

Then the sequence (x_k) generated by the proximal algorithm (4.3) converges to a solution of (4.2).

Remark 4.3. The implementability of the proximal algorithm (4.3) is determined by two factors: (i) the Lipschitz continuity of the gradient ∇f and (ii) the computability of the proximal operator $\mathrm{prox}_{\lambda g}$ (thus, the lasso (1.1) can be solved by the proximal algorithm (4.3)). Though the p-EN (4.1) is the sum of three terms, it can also be solved by the proximal algorithm (4.3) since it can be reformulated as the sum of two convex functions as shown in (4.5) below, with one term having a Lipschitz continuous gradient and another term having a computable proximal operator.

In order to apply the proximal algorithm (4.3) to the p-EN (4.1), we take

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, \quad g(x) = \lambda\|x\|_1 + \gamma\frac{1}{p}\|x\|_p^p = \sum_{j=1}^{n}\left(\lambda|x_j| + \frac{\gamma}{p}|x_j|^p\right). \tag{4.5}$$

Notice that $\nabla f(x) = A^t(Ax - b)$ is $\|A\|^2$-Lipschitz continuous, and for each $\lambda_k > 0$, the proximal operator $\mathrm{prox}_{\lambda_k g}(x)$ can be computed in the following decomposed way:

$$\mathrm{prox}_{\lambda_k g}(x) = \sum_{j=1}^{n} \mathrm{prox}_{g_j}(x_j), \tag{4.6}$$

where

$$g_j(x_j) = \lambda_k\lambda|x_j| + \frac{\lambda_k\gamma}{p}|x_j|^p.$$

The lemma below provides a way to compute the proximal operator prox_{g_j}.

Lemma 4.4. *Let $p \in (1,\infty)$ and set $\psi(t) = \alpha|t| + \beta|t|^p$ for $t \in \mathbb{R}$, with $\alpha > 0, \beta > 0$. Then*

$$\operatorname{prox}_\psi(t) = \begin{cases} 0, & \text{if } |t| \le \alpha, \\ s, & \text{if } |t| > \alpha, \end{cases} \tag{4.7}$$

where $s \in \mathbb{R}$ is the unique solution to the equation:

$$\alpha \cdot \operatorname{sgn}(s) + \beta p |s|^{p-1} \operatorname{sgn}(s) + s - t = 0. \tag{4.8}$$

Proof. By definition, we have

$$\operatorname{prox}_\psi(t) = \arg\min_{s\in\mathbb{R}} \left\{ \alpha|s| + \beta|s|^p + \frac{1}{2}(s-t)^2 \right\}.$$

It turns out that $s := \operatorname{prox}_\psi(t)$ is the unique solution to the inclusion

$$0 \in \alpha\partial|s| + \beta p|s|^{p-1}\operatorname{sgn}(s) + s - t. \tag{4.9}$$

By the formula (2.3), we find that in the case of $s = 0$, the last equation is reduced to $0 \in \alpha[-1,1] - t$. Hence, $\operatorname{prox}_\psi(t) = 0$ for $|t| \le \alpha$. While in the case of $|t| > \alpha$, (4.9) is reduced to (4.8) as $\partial|s| = \operatorname{sgn}(s)$ for $s \ne 0$. □

Applying the proximal algorithm (4.3) and Theorem 4.2 to p-EN (4.1) yields the following splitting algorithm and its convergence.

Theorem 4.5. *Define f and g by (4.5). Generate a sequence $\{x_k\}$ by the following proximal algorithm:*

$$x_{k+1} = \operatorname{prox}_{\lambda_k g}(x_k - \lambda_k A^t(Ax_k - b)), \tag{4.10}$$

where $\operatorname{prox}_{\lambda_k g}$ is the proximal operator of $\lambda_k g$ given by (4.6) and for each j,

$$\operatorname{prox}_{g_j}(x_j) = \begin{cases} 0, & \text{if } |x_j| \le \lambda_k\lambda, \\ s_j, & \text{if } |x_j| > \lambda_k\lambda, \end{cases} \tag{4.11}$$

where $s_j \in \mathbb{R}$ is the unique solution to the equation:

$$\lambda_k\lambda \cdot \operatorname{sgn}(s_j) + \lambda_k\gamma|s_j|^{p-1}\operatorname{sgn}(s) + s - t = 0. \tag{4.12}$$

Assume

$$0 < \liminf_{k\to\infty} \lambda_k \le \limsup_{k\to\infty} \lambda_k < \frac{2}{\|A\|_2^2}. \tag{4.13}$$

Then the sequence (x_k) generated by the proximal algorithm (4.10) converges to the solution of the p-EN (3.1).

Proof. By Lemma 4.4, this is a straightforward application of Theorem 4.2 to the case where f and g are given by (4.5). □

Remark 4.6. If we choose

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \gamma\frac{1}{p}\|x\|_p^p, \quad g(x) = \lambda\|x\|_1, \tag{4.14}$$

then f is differentiable with the gradient

$$\nabla f(x) = A^t(Ax - b) + \gamma J_p(x) \tag{4.15}$$

for $x \in \mathbb{R}^n$. The proximal algorithm (4.3) then yields the following algorithm:

$$x_{k+1} = \operatorname{prox}_{\lambda_k\|\cdot\|_1}(x_k - \lambda_k(A^t(Ax_k - b) + \gamma J_p(x_k))). \tag{4.16}$$

However, since J_p is not Lipschitz continuous (except for $p = 2$), Theorem 4.5 does not apply to this choice of f. It is therefore interesting to know whether the sequence $\{x_k\}$ generated by the algorithm (4.16) is convergent under the condition (4.13).

Acknowledgment

The authors are grateful to the referees for their suggestions and comments which improved the presentation of this manuscript.

References

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

[2] E. Candes, J. Romberg and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure Applied Math.* 59 (2006) 1207–1223.

[3] S. Chen, D. Donoho and M. Saunders, Atomic decomposition by basis pursuit, *SIAM J. Scientific Comp.* 20 (1998) 33–61.

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

[5] Y. Dang, F. Meng and J. Sun, Convergence of an algorithm for the split common fixed-point of asymptomatic quasi-nonexpansive operators, *Pacific J. Optim.* 10 (2014) 453–460.

[6] Y. Dang, F. Meng and J. Sun, An iterative algorithm for split common fixed-point problem for demicontractive mappings, in the Collection of ICOTA2012, Springer (to appear).

[7] D. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* 52 (2006) 1289–1306.

[8] J. Douglas and H.H. Rachford, On the numerical solution ofthe heat conduction problem in 2 and 3 space variables, *Trans. Amer. Math. Soc.* 82 (1956) 421–439.

[9] 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.

[10] P.L. Lions and B. Mercier, Splitting algorithms for the sum of two nonlinear operators, *SIAM J. Numer. Anal.* 16 (1979) 964–979.

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

[12] J.-J. Moreau, Proximite et dualite dans un espace hilbertien, *Bull. Soc. Math. France* 93 (1965) 272–299.

[13] G.B. Passty, Ergodic convergence to a zero of the sum of monotone operators in Hilbert spaces, *J. Math. Anal. Appl.* 72 (1979) 383–390.

[14] D.H. Peaceman and H.H. Rachford, The numerical solution ol parabolic elliptic differential equations, *J. Soc. Indust. Appl. Math.* 3 (1955) 28–41.

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

[16] R. Tibshirani, M. Saunders, R. Rosset, J. Zhu and K. Knight, Sparsity and smoothness via the fused lasso, *J. Royal Statist. Soc.* Ser. B 67 (2005) 91–108.

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

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

[21] 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 10 November 2013
revised 13 August 2014
accepted for publication 25 August 2014

Najla Altwaijry
College of Science, King Saud University
Box 2455, Riyadh 11451, Saudi Arabia
E-mail address: najla@ksu.edu.sa

Souhail Chebbi
Department of Mathematics, College of Science
King Saud University, Box 2455, Riyadh 11451
Saudi Arabia
E-mail address: schebbi@ksu.edu.sa

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