Fixed Point Theory, 24(2023), No. 2, 661-674
DOI: 10.24193/fpt-ro.2023.2.13
http://www.math.ubbcluj.ro/~nodeacj/sfptcj.html

PRIMAL-DUAL FIXED POINT METHODS FOR REGULARIZED LEAST-SQUARES PROBLEMS

GUOHUI LIU* AND HONG-KUN XU**

*College of Mathematics and Information,
Henan Normal University, Xinxiang, 453007, China
E-mail: liuguohui1028@yeah.net

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

Abstract. We will study primal-dual fixed point methods for the least-squares problem regularized by l_p-norms with $p \in [1, 2]$. Our methods and results extend some of Ribeiro and Richtarik [9] and Silva, et al [10] where the case of $p = 2$ (i.e, the ridge regression) is studied. The case of $p = 1$ corresponds to the lasso [11] and the general case of $p \in [1, 2]$ corresponds to the iterative shrinkage/thresholding algorithm (ISTA) of Daubechies, et al [5]. We will apply the proximal-gradient methods to prove convergence of our primal-dual fixed point methods for the general l_p-regularization, and also for the elastic net problem [14].

Key Words and Phrases: Prime-dual method, fixed point method, lasso, elastic net, regularization.

2020 Mathematics Subject Classification: 47H10, 47J25, 90C25, 90C46.

1. Introduction

The ridge regression problem (RRP) is the least-squares problem regularized by the Euclidean 2-norm $\|\cdot\|_2$ (also known as Tikhonov regularization), that is,

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

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

The (Fenchel) dual problem of the primal problem (1.1) is also a RRP given by

$$\max_{z\in\mathbb{R}^m} D(z) := -\frac{1}{2\lambda}\|A^\top z\|_2^2 + \langle z, b\rangle - \frac{1}{2}\|z\|_2^2. \tag{1.2}$$

Here $A^\top$ is the transpose of A.

Recently, Ribeiro and Richtarik [9], and Silva, et al [10] introduced primal-dual fixed point methods to the primal and dual problems (1.1) and (1.2) by coupling the primal and dual variables x and z in the product space $\mathbb{R}^{n+m}$. [Note: our formulation

Supported in part by a Natural Science Foundation of China grant number U1811461 and by an Australian Research Council/Discovery Project DP200100124.

of RRP (1.1) is a slightly rescaled version of those of [9, Eq. (1), page 343] and Silva, et al [10, Eq. (2), page 1942], so is the corresponding dual problem (1.2); there is, however, no essential difference.] Let us briefly review the methods of [9, 10]. Consider the problem of minimizing the difference of the primal and dual objective functions over the product space $\mathbb{R}^{n+m}$:

$$\begin{aligned}\min_{w\in\mathbb{R}^{n+m}} \varphi(w) : &= P(x) - D(z) \\ &= \frac{1}{2}\|Ax-b\|_2^2 + \frac{\lambda}{2}\|x\|_2^2 + \frac{1}{2\lambda}\|A^\top z\|_2^2 - \langle z, b\rangle + \frac{1}{2}\|z\|_2^2, \qquad (1.3)\end{aligned}$$

where $w = (x^\top, z^\top)^\top \in \mathbb{R}^{n+m}$ (we will always write $w = (x, z)$ hereafter). This is a strongly convex, quadratic minimization and hence has a unique solution. Since

$$\nabla\varphi(w) = \begin{bmatrix} \nabla_x P(x) \\ -\nabla_z D(z) \end{bmatrix} = \begin{bmatrix} A^\top(Ax-b)+\lambda x \\ \lambda^{-1}AA^\top z - b + z \end{bmatrix},$$

the optimality condition of $\nabla\varphi(w) = 0$ turns out to be the equivalent fixed point equation

$$w = Mw + \bar{w}, \qquad (1.4)$$

where

$$M = \begin{bmatrix} -\lambda^{-1}A^\top A & 0 \\ 0 & -\lambda^{-1}AA^\top \end{bmatrix}, \quad \bar{w} = \begin{bmatrix} \lambda^{-1}A^\top b \\ b \end{bmatrix}.$$

A more general (equivalent) fixed point equation is the following

$$w = (1-\theta)w + \theta(Mw + \bar{w}), \qquad (1.5)$$

where $\theta \in (0, 2]$. When $\theta = 1$, (1.5) is reduced to (1.4).

The main primal-dual fixed point methods introduced in [9, 10] are of the form:

$$w_{k+1} = (1-\theta)w_k + \theta(Mw_k + \bar{w}) = Gw_k + \theta\bar{w}, \qquad (1.6)$$

with $G \equiv G(\theta) := (1-\theta)I + \theta M$. Observe that the fixed point problem (1.5) is a linear problem and the convergence of the linear fixed point method (1.6) depends on the property that the spectral radius $\rho(\theta)$ of the matrix $G(\theta)$ is strictly less than one with a suitably chosen parameter θ; see [9, Theorems 3.4 and 3.5] and [10, Theorems 2 and 4].

The aim of the present paper is to extend the primal-dual fixed point methods of [9, 10] to a general p-norm regularized least-squares problem; namely, the problem

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\|Ax-b\|_2^2 + \frac{\lambda}{p}\|x\|_p^p, \qquad (1.7)$$

where A is an $m \times n$ real matrix, $b \in \mathbb{R}^m$, $\lambda > 0$ is a regularization parameter, and $p \in [1, 2]$. [Here and throughout the rest of the paper, we use $F(x)$, instead of $P(x)$, to denote the objective function of the primal problem.] Note that when $p = 2$, (1.7) returns to RRP (1.1); when $p = 1$, (1.7) turns out to be the lasso [11]:

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1. \qquad (1.8)$$

Note also that the p-norm regularized least-squares problem (1.7) was first introduced in [5, 6]) for iterative shrinkage/thresholding algorithms (ISTA) for linear inverse problems with a sparsity constraint; see also [3] for application in sparse recovery of signals.

Let $F^*(z)$, which will be worked out in section 3, denote the dual function of the primal function $F(x)$ of (1.7). The main contribution of this paper is to extend the idea of [9, 10] for RRP (1.1) to (1.7). Such an extension is nontrivial because the optimality condition of the primal-dual function of RRP (1.1) is a linear fixed point problem (see (1.4) and (1.5)); while that of the primal-dual function of (1.7) is nonlinear. The latter means that more sophisticated tools (such as proximal mappings) must be employed.

More precisely, we will couple the primal and dual variables x and z and consider the minimization problem in the product space $\mathbb{R}^{n+m}$

$$\min_{(x,z)\in\mathbb{R}^{n+m}} \varphi(x,z) := F(x) - F^*(z). \tag{1.9}$$

Write $w = (x, z)$ as a general point in $\mathbb{R}^{n+m}$. Our strategy is to convert the optimality condition $\nabla\varphi(w) = 0$ to a fixed point problem of some nonlinear mapping T to which we apply appropriate fixed point methods to generate a sequence that will be convergent to a fixed point of T (hence a solution of (1.9)).

This strategy will also be applied to the elastic net (EN) [14] which is the optimization problem

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{2}\|x\|_2^2 \tag{1.10}$$

where λ, $\gamma > 0$ are regularization parameters.

The structure of the paper is as follows. In the next section we include basic concepts and tools such as conjugate functions and proximal mappings, and their properties. The main results will be presented and proved in section 3.

2. Preliminaries

Let H be a Hilbert space with inner product $\langle,\cdot,\cdot\rangle$ and norm $\|\cdot\|$. and let $\Gamma_0(H)$ be the space of proper, lower semicontinuous and convex functions from H to the extended real line $\overline{\mathbb{R}} := (-\infty, \infty]$.

2.1. Conjugate and Subdifferential.

Let $f \in \Gamma_0(H)$. The Fenchel conjugate of f is defined as

$$f^*(x^*) := \sup\{\langle x^*, x\rangle - f(x) : x \in H\}, \quad x^* \in H.$$

The following properties are pertinent to our argument:

(i) $(\lambda f)^*(x^*) = \lambda f^*(x^*/\lambda)$ for $\lambda > 0$.

(ii) If $f(x) = \frac{1}{p}\|x\|_p^p$ for some $p \in (1, \infty)$, then $f^*(x^*) = \frac{1}{p'}\|x^*\|_{p'}^{p'}$, where $p' = p/(p-1)$. In particular, if $f(x) = \frac{1}{2}\|x\|_2^2$, then $f^*(x^*) = \frac{1}{2}\|x^*\|_2^2$.

(iii) If $f(x) = \frac{1}{p}\|x - y\|_p^p$ for some $p \in (1, \infty)$ and fixed $y \in H$, then $f^*(x^*) = \langle x^*, y\rangle + \frac{1}{p'}\|x^*\|_{p'}^{p'}$, where $p' = p/(p-1)$.

The following theorem is helpful in finding dual problems.

Theorem 2.1. *Suppose H_1 and H_2 are Hilbert spaces and $A : H_1 \to H_2$ is a bounded linear operator. Let $f \in \Gamma_0(H_1)$ and $g \in \Gamma_0(H_2)$ and consider the primal problem*

$$\min_{x\in H_1} F(x) := f(x) + g(Ax). \tag{2.1}$$

Suppose $F \in \Gamma_0(H_1)$. Then the dual problem of (2.1) *is*

$$\max_{z\in H_2} F^*(z) := -f^*(A^*z) - g^*(-z). \tag{2.2}$$

Here A^* is the adjoint of A. If, in addition, the function $H_2 \ni z \mapsto g(Ax_0 - z)$ is continuous (where $x_0 \in H_1$ is such that $F(x_0) < \infty$) and F is coercive (i.e., $F(x) \to \infty$ as $\|x\| \to \infty$), then the prime and dual problems share the same optimal value, that is, $\min_{x\in H_1} F(x) = \max_{z\in H_2} F^*(z)$.

Recall that a point $\xi \in H$ is said to be a subgradient of a function $f \in \Gamma_0(H)$ at a point $x \in \mathrm{dom}(f)$ if

$$f(y) \geq f(x) + \langle \xi, y - x\rangle$$

for all $y \in H$. The set of all subgradients at x is denoted as $\partial f(x)$. The mapping ∂f is then referred to as the subdifferential (mapping) of f. For instance, if we take $f(x) = |x|$ for $x \in \mathbb{R}$, then f is nondifferentiable at $x = 0$. It is however subdifferentiable at $x = 0$ with the subdifferential $\partial f(0) = [-1, 1]$. For more details of conjugate functions and subdifferential, the reader is referred to the monographs [1, 2].

2.2. **Proximal Mappings.**

Definition 2.2. The proximal mapping of a function $f \in \Gamma_0(H)$ of level $\lambda > 0$ is defined by

$$\mathrm{prox}_{\lambda f}(x) := \arg\min_{v\in H} \left\{ f(v) + \frac{1}{2\lambda}\|v - x\|^2 \right\}, \quad x \in H. \tag{2.3}$$

The optimal value of (2.3) is known as the Moreau envelope, denoted f_λ. Namely,

$$f_\lambda(x) := \min_{v\in H} \left\{ f(v) + \frac{1}{2\lambda}\|v - x\|^2 \right\} = f(\mathrm{prox}_{\lambda f}(x)) + \frac{1}{2\lambda}\|\mathrm{prox}_{\lambda f}(x) - x\|^2. \tag{2.4}$$

We list some of the useful properties in the proximal operators.

Proposition 2.3. (cf. [4, 7, 13]) *Let $f \in \Gamma_0(H)$ and $\lambda \in (0, \infty)$.*

(i) *If C is a nonempty closed convex subset of H and $f = I_C$ is the indicator function of C, then the proximal mappings* $\mathrm{prox}_{\lambda f} = P_C$ *for all $\lambda > 0$, where P_C is the metric projection from H onto C, that is,*

$$P_C x = \arg\min_{y\in C} \|x - y\|^2, \quad x \in H.$$

(ii) $\mathrm{prox}_{\lambda f}$ *is firmly nonexpansive (hence nonexpansive). Recall that a mapping $T : H \to H$ is said to be firmly nonexpansive if*

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

and T is nonexpansive if $\|Tx - Ty\| \leq \|x - y\|$ for $x, y \in H$.

(iii) $\mathrm{prox}_{\lambda f} = (I + \lambda\partial f)^{-1} = J_\lambda^{\partial f}$*, the resolvent of the subdifferential ∂f of f.*

(iv) $y \in \partial f(x) \;\Leftrightarrow\; x = \operatorname{prox}_f(x+y)$.

(v) *For each* $x \in H$, $f(\operatorname{prox}_{\lambda f} x) \le f_\lambda(x) \le f(x)$, *and* $\lim_{\lambda \to 0} f_\lambda(x) = f(x)$.

(vi) *The Moreau envelope* f_λ *is Fréchet differentiable with gradient*

$$\nabla f_\lambda = \frac{1}{\lambda}(I - \operatorname{prox}_{\lambda f}).$$

A toy example of proximal mappings is the scalar soft-thresholding mapping (more examples can be found in [4]). This is the proximal mapping of the absolute value function in the one-dimensional real line:

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

The soft-thresholding mapping of the 1-norm of $\mathbb{R}^n$ is given componentwise by

$$(\operatorname{prox}_{\lambda\|\cdot\|}(x))_j = \operatorname{prox}_{\lambda|\cdot|}(x_j) = \operatorname{sgn}(x_j)\max\{|x_j| - \lambda, 0\}, \quad j = 1, 2, \cdots, n,$$

where $x = (x_1, \cdots, x_n)^\top \in \mathbb{R}^n$, and x_j stands for the j-th component of x.

2.3. **Proximal-Gradient Algorithm.** The proximal mappings can be used to minimize the sum of two convex functions:

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

where $f, g \in \Gamma_0(H)$. It is often the case where one of them is differentiable. The following is an equivalent fixed point formulation of (2.5).

Proposition 2.4. *Let* $f, g \in \Gamma_0(H)$. *Let* $x^* \in H$ *and* $\lambda > 0$. *Assume* f *is finite-valued and differentiable on* H. *Then* x^* *is a solution to (2.5) if and only if* x^* *solves the fixed point equation*

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

The fixed point equation (2.6) immediately yields the following fixed point algorithm which is also known as the proximal-gradient algorithm (PGA) for solving (2.5) as follows.

Initializing $x_0 \in H$ and iterating

$$x_{n+1} = (\operatorname{prox}_{\lambda_n g} \circ (I - \lambda_n \nabla f))x_n \tag{2.7}$$

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

Theorem 2.5. (cf. [4, 12]) *Let* $f, g \in \Gamma_0(H)$ *and assume (2.5) has a solution. Assume in addition that*

(i) ∇f *is L-Lipschitz continuous on* H: $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$ *for* $x, y \in H$.

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

Then the sequence (x_n) *generated by the proximal algorithm (2.7) converges weakly to a solution of (2.5).*

2.4. **Notation.** We adopt the following notation:

- $\mathbb{R}^n$ stands for the real Euclidean n-space with $n \geq 1$ integer.
- $A^\top$ stands for the transpose of real matrix A.
- $\|\cdot\|_p$ stands for the p-norm of $\mathbb{R}^n$, with $p \in [1,2]$, that is

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}, \quad x = (x_1, \cdots, x_n)^\top \in \mathbb{R}^n.$$

3. Prime-dual fixed point methods

3.1. **Prime-dual Fixed Point Method for Lasso.** Consider the primal lasso optimization problem

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1. \tag{3.1}$$

Here A is an $m\times n$ real matrix and $\lambda > 0$ is a regularization parameter. We can rewrite F as $F(x) = \lambda f(x) + g(Ax)$, where $f(x) := \|x\|_1$ for $x \in \mathbb{R}^n$ and $g(v) := \frac{1}{2}\|v-b\|^2$ for $v \in \mathbb{R}^m$.

Observe that the conjugate of $f(x) = \|x\|_1$ is the indicator function of the closed unit l_∞-ball, that is,

$$f^*(x) = \begin{cases} 0, & \text{if } \|x\|_\infty \leq 1 \\ \infty, & \text{if } \|x\|_\infty > 1. \end{cases} \tag{3.2}$$

Applying Theorem 2.1, we get that the dual objective function of F is given by (for $z \in \mathbb{R}^m$),

$$\begin{aligned} F^*(z) &= -(\lambda f)^*(A^\top z) - g^*(-z) \\ &= -\lambda f^*(A^\top z/\lambda) - (\frac{1}{2}\|-z\|^2 + \langle -z, b\rangle). \end{aligned}$$

Combining with (3.2) yields

$$F^*(z) = \begin{cases} -\frac{1}{2}\|z\|^2 + \langle z, b\rangle, & \text{if } \|A^\top z\|_\infty \leq \lambda \\ -\infty, & \text{if } \|A^\top z\|_\infty > \lambda. \end{cases}$$

Consequently, the difference $F(x) - F^*(z)$ is given by

$$F(x) - F^*(z) = \begin{cases} \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1 + \frac{1}{2}\|z\|^2 - \langle z, b\rangle, & \text{if } \|A^\top z\|_\infty \leq \lambda \\ \infty, & \text{if } \|A^\top z\|_\infty > \lambda. \end{cases}$$

The dual problem turns out to be

$$\min_{(x,z)\in\mathbb{R}^{n+m}} F(x) - F^*(z) = \min_{\substack{x\in\mathbb{R}^n \\ \|A^\top z\| \leq \lambda}} \left\{ \frac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_1 + \frac{1}{2}\|z\|^2 - \langle z, b\rangle \right\}. \tag{3.3}$$

Let

$$K = \{z \in \mathbb{R}^m : \|A^\top z\|_\infty \leq \lambda\}$$

and use i_K to denote the indicator of K, namely, $i_K(z) = 0$ if $z \in K$ and ∞ if $z \notin K$. We rewrite $F(x) - F^*(z) = \eta(x,z) + \xi(x,z)$, where

$$\eta(x,z) = \frac{1}{2}\|Ax-b\|^2 + \frac{1}{2}\|z\|^2 - \langle z, b\rangle, \quad \xi(x,z) = \lambda\|x\|_1 + i_K(z).$$

We have, for $w = (x, z)$,

$$\nabla\eta(w) = \begin{bmatrix} A^\top(Ax - b) \\ z - b \end{bmatrix} \tag{3.4}$$

and

$$\operatorname{prox}_{\mu\xi}(w) = \begin{bmatrix} \operatorname{prox}_{\mu\lambda\|\cdot\|_1}(x) \\ P_K(z) \end{bmatrix}. \tag{3.5}$$

Consequently, the optimization problem (3.3) can be solved by the proximal gradient algorithm (2.7) which generates a sequence $\{w_k\}$ by the iteration process

$$w_{k+1} = \operatorname{prox}_{\mu\xi}(w_k - \mu\nabla\eta(w_k)), \quad k = 0, 1, \cdots . \tag{3.6}$$

Equivalently, putting $w_k = (x_k, z_k)$,

$$\begin{aligned} x_{k+1} &= \operatorname{prox}_{\mu\lambda\|\cdot\|_1}(x_k - \mu A^\top Ax_k + \mu A^\top b), \\ z_{k+1} &= P_K((1 - \mu)z_k + \mu b). \end{aligned} \tag{3.7}$$

The convergence of (3.6) is given below.

Theorem 3.1. *Let the stepsize μ be chosen such that $0 < \mu < \frac{2}{\|A\|_2^2 \vee 1}$. [Here we use the notation: $a \vee b = \max\{a, b\}$ for real numbers a and b.] Then the sequence (w_k) generated by the proximal gradient algorithm* (3.6) *converges to a solution of the prime-dual problem* (3.3).

Proof. We claim that η is L-smooth (i.e., $\nabla\eta$ is L-Lipschitz) with $L = \|A\|_2^2 \vee 1$. As a matter of fact, by (3.4) we have, for $w = (x, z), w' = (x', z') \in \mathbb{R}^{n+m}$,

$$\begin{aligned} \|\nabla\eta(w) - \nabla\eta(w')\|_2^2 &= \|A^\top A(x - x')\|_2^2 + \|z - z'\|_2^2 \\ &\le \|A^\top A\|_2^2\|x - x'\|_2^2 + \|z - z'\|_2^2 \\ &\le (\|A^\top A\|_2^2 \vee 1)(\|x - x'\|_2^2 + \|z - z'\|_2^2) \\ &= L^2\|w - w'\|_2^2. \end{aligned}$$

Here $L = \|A^\top A\|_2 \vee 1 = \|A\|_2^2 \vee 1$. Now the convergence of $\{w_k\}$ follows immediately from Theorem 2.5. □

3.2. l_p **regularization** $(1 < p \le 2)$**.** Let $p \in (1, 2]$ and consider the (primal) l_p regularized least-squares problem

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\|Ax - b\|^2 + \lambda\frac{1}{p}\|x\|_p^p, \tag{3.8}$$

where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_2$, and $p \in (1, 2]$. Set

$$f(x) = \frac{1}{p}\|x\|_p^p \quad (x \in \mathbb{R}^n), \quad g(v) = \frac{1}{2}\|v - b\|^2 \quad (v \in \mathbb{R}^m).$$

Then we may rewrite F in the form

$$F(x) = \lambda f(x) + g(Ax), \quad x \in \mathbb{R}^n.$$

By Theorem 2.1, the dual problem of (3.8) is given as follows:

$$\max_{z\in\mathbb{R}^m} F^*(z) := -(\lambda f)^*(A^\top z) - g^*(-z). \tag{3.9}$$

Since $f^*(x) = \frac{1}{q}\|x\|_q^q$, where $x \in \mathbb{R}^n$ and $q = p/(p-1)$, and $g^*(z) = \langle z, b\rangle + \frac{1}{2}\|z\|^2$, where $z \in \mathbb{R}^m$, it follows from (3.9) that

$$\begin{aligned} F^*(z) &= -\lambda f^*(A^\top z/\lambda) - g^*(-z) \\ &= -\lambda \frac{1}{q}\|A^\top z/\lambda\|_q^q - (\langle -z, b\rangle + \frac{1}{2}\| - z\|^2) \\ &= -\lambda^{1-q}\frac{1}{q}\|A^\top z\|_q^q + \langle z, b\rangle - \frac{1}{2}\|z\|^2. \end{aligned}$$

Consequently,

$$\varphi(w) := F(x) - F^*(z) = \frac{1}{2}\|Ax - b\|^2 + \lambda\frac{1}{p}\|x\|_p^p + \lambda^{1-q}\frac{1}{q}\|A^\top z\|_q^q - \langle z, b\rangle + \frac{1}{2}\|z\|^2, \quad (3.10)$$

where $w = (x, z)$. We have

$$\nabla\varphi(w) = \begin{bmatrix} \nabla_x F(x) \\ -\nabla_z F^*(z) \end{bmatrix} = \begin{bmatrix} A^\top(Ax - b) + \lambda J_p(x) \\ \lambda^{1-q} A J_q(A^\top z) - b + z \end{bmatrix}$$

where $J_p(x) = \nabla(\frac{1}{p}\|x\|_p^p)$ is the (generalized) duality map from $(\mathbb{R}^n, \|\cdot\|_p)$ to the dual space $(\mathbb{R}^n, \|\cdot\|_q)$, that is, $J_p(x) = x^*$, with $\langle x, x^*\rangle = \|x\|_p^p$ and $\|x^*\|_q = \|x\|_p^{p-1}$.

Setting

$$\varphi_1(w) = \frac{1}{2}\|Ax - b\|^2 - \langle z, b\rangle + \frac{1}{2}\|z\|^2, \quad \varphi_2(w) = \lambda\frac{1}{p}\|x\|_p^p + \lambda^{1-q}\frac{1}{q}\|A^\top z\|_q^q$$

we have $\varphi(w) = \varphi_1(w) + \varphi_2(w)$ and the problem is reduced to the composite minimization problem

$$\min_{w=(x,z)\in\mathbb{R}^{n+m}} \varphi_1(w) + \varphi_2(w). \quad (3.11)$$

It is easy to compute

$$\nabla\varphi_1(w) = \begin{bmatrix} A^\top(Ax - b) \\ z - b \end{bmatrix}. \quad (3.12)$$

This is the same as (3.4), hence, $\nabla\varphi_1(w)$ is L-Lipschtiz with $L = \|A\|_2^2 \vee 1$.

The proximal-gradient algorithm (PGA) applied to the composite optimization (3.11) results in the following algorithm:

$$w_{k+1} = \mathrm{prox}_{\mu\varphi_2}(w_k - \mu\nabla\varphi_1(w_k)), \quad k = 0, 1, \cdots, \quad (3.13)$$

where the initial point $w_0 = (x_0, z_0) \in \mathbb{R}^{n+m}$ is chosen arbitrarily, and $\mu > 0$ is a stepsize to be selected appropriately.

By (3.12) we have

$$w_k - \mu\nabla\varphi_1(w_k) = \begin{bmatrix} x_k - \mu A^\top(Ax_k - b) \\ (1-\mu)z_k + \mu b \end{bmatrix}. \quad (3.14)$$

Regarding the proximal mapping of φ_2, we have

$$\mathrm{prox}_{\mu\varphi_2}(w) = \begin{bmatrix} \mathrm{prox}_{\mu h_1}(x) \\ \mathrm{prox}_{\mu h_2}(z) \end{bmatrix} \quad (3.15)$$

where

$$h_1(x) = \frac{\lambda}{p}\|x\|_p^p, \quad h_2(z) = \frac{\lambda^{1-q}}{q}\|A^\top z\|_q^q.$$

The algorithm (3.13) can be rewritten componentwise as

$$\begin{aligned} x_{k+1} &= \operatorname{prox}_{\mu h_1}(x_k - \mu A^\top(Ax_k - b)) \\ z_{k+1} &= \operatorname{prox}_{\mu h_2}((1-\mu)z_k + \mu b). \end{aligned} \tag{3.16}$$

The convergence of (3.13) is given below.

Theorem 3.2. *Suppose the stepsize μ is chosen so that $0 < \mu < \frac{2}{\|A\|_2^2 \vee 1}$. Then the sequence (w_k) generated by the algorithm* (3.13) *converges to a solution of* (3.11)*.*

The implementation of the algorithm depends on the evaluations of the proximal mappings $\operatorname{prox}_{\mu h_1}$ and $\operatorname{prox}_{\mu h_2}$. The special case of $p = 2$ recovers the primal ridge regression problem (1.1) and its dual problem (1.2). In this case, we have

$$h_1(x) = \lambda \frac{1}{2}\|x\|_2^2 \text{ and } h_2(z) = \frac{1}{2\lambda}\|A^\top z\|_2^2.$$

It then turns out that $\operatorname{prox}_{\mu h_1}(x) = \frac{1}{1+\lambda\mu}x$ and $\operatorname{prox}_{\mu h_2}(z) = (I + \frac{\mu}{\lambda}AA^\top)^{-1}z$. Moreover, the algorithm (3.16) is reduced to the algorithm:

$$\begin{aligned} x_{k+1} &= \tfrac{1}{1+\lambda\mu}(x_k - \mu A^\top(Ax_k - b)) \\ z_{k+1} &= (I + \tfrac{\mu}{\lambda}AA^\top)^{-1}((1-\mu)z_k + \mu b). \end{aligned} \tag{3.17}$$

By Theorem 3.2, we obtain that the sequence $\{w_k\}$ generated by the algorithm (3.17) converges to a solution of the primal-dual ridge regression problem (1.3) provided the stepsize μ is chosen so that $0 < \mu < \frac{2}{\|A\|_2^2 \vee 1}$ and the regularization parameter $\lambda > 0$ is chosen arbitrarily fixed. This convergence result differs from those of [9, 10], due to our different approach.

For a general $p \in (1, 2]$, let us discuss the proximal mapping of the functional $h(x) := \mu\|x\|_p^p$ on $\mathbb{R}^n$, which is

$$\begin{aligned} \operatorname{prox}_{\mu h}(x) &= \arg\min_{u \in \mathbb{R}^n} \left(\|u\|_p^p + \frac{1}{2\mu}\|u - x\|_2^2 \right) \\ &= \arg\min_{u \in \mathbb{R}^n} \sum_{i=1}^n \left(|u_i|^p + \frac{1}{2\mu}(u_i - x_i)^2 \right). \end{aligned}$$

It turns out that the i-th component of $\operatorname{prox}_{\mu h}(x)$ is given by

$$z_i := (\operatorname{prox}_{\mu h}(x))_i = \arg\min_{u_i \in \mathbb{R}} \left(|u_i|^p + \frac{1}{2\mu}(u_i - x_i)^2 \right).$$

The unique solution z_i to the last equation is determined by the first-order optimality condition:

$$\mu p |z_i|^{p-1}\operatorname{sgn}(z_i) + z_i - x_i = 0. \tag{3.18}$$

This equation has no closed solution formula, in general; it does, however, for some particular values of p [8]. For instance, if $p = 2$, then $z_i = (1 + 2\mu)^{-1}x_i$ for each i, and $\operatorname{prox}_{\mu h}(x) = (1 + 2\mu)^{-1}x$ as mentioned previously. If $p = 4/3$, equation (3.18) turns out to be $\frac{4}{3}\mu|z_i|^{1/3}\mathrm{sgn}(z_i)+z_i-x_i=0$, which is reduced to the algebraic equation of order three (via the substitution $s_i:=|z_i|^{1/3}$):

$$\frac{4}{3}\mu s_i+s_i^3-|x_i|=0. \tag{3.19}$$

The only real solution to the equation (3.19) is given by the following formula:

$$s_i=\sqrt[3]{\frac{|x_i|}{2}+\sqrt{\frac{|x_i|^2}{4}+\frac{64}{729}\mu^3}}+\sqrt[3]{\frac{|x_i|}{2}-\sqrt{\frac{|x_i|^2}{4}+\frac{64}{729}\mu^3}}.$$

3.3. **Primal-dual Fixed Point Method for Elastic Net.** Zou and Hastie [14] introduced the elastic net (EN) which is the minimization problem

$$\min_{x\in\mathbb{R}^n} F(x):=\frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1+\gamma\frac{1}{2}\|x\|_2^2. \tag{3.20}$$

Evidently, EN has a unique solution, due to the strict convexity of the 2-norm $\|x\|_2$.

To find the dual problem of (3.20),we rewrite F as the sum $F(x)=f(x)+g(Ax)$, where

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

We now compute the conjugate of f. By definition, we have for $x\in\mathbb{R}^n$,

$$\begin{aligned}f^*(x)&=\sup_{u\in\mathbb{R}^n}(\langle x,u\rangle-f(u))\\&=\sup_{u\in\mathbb{R}^n}(\langle x,u\rangle-\lambda\|u\|_1-\gamma\frac{1}{2}\|u\|_2^2)\\&=-\lambda\min_{u\in\mathbb{R}^n}\{-\frac{1}{\lambda}\langle x,u\rangle+\|u\|_1+\frac{\gamma}{2\lambda}\|u\|_2^2\}\\&=-\lambda\min_{u\in\mathbb{R}^n}\{\|u\|_1+\frac{1}{2(\lambda/\gamma)}\|u-\frac{1}{\gamma}x\|_2^2\}+\frac{1}{2\gamma}\|x\|_2^2.\end{aligned} \tag{3.21}$$

Recall that the soft-thresholding mapping of the norm $\|\cdot\|_1$ is defined as

$$S_\mu(x):=\mathrm{prox}_{\mu\|\cdot\|_1}(x)=\arg\min_{u\in\mathbb{R}^n}\left(\|u\|_1+\frac{1}{2\mu}\|u-x\|_2^2\right),\quad \mu>0,\ x\in\mathbb{R}^n. \tag{3.22}$$

Recall also that $S_\mu(x)$ is given componentwise by

$$(S_\mu(x))_j=\mathrm{sgn}(x_j)\max\{|x_j|-\mu,0\},\quad j=1,2,\cdots,n.$$

Moreau's envelope is the optimal value of the minimization problem in (3.22), that is,

$$M_\mu(x):=\min_{u\in\mathbb{R}^n}\left(\|u\|_1+\frac{1}{2\mu}\|u-x\|_2^2\right)=\|S_\mu(x)\|_1+\frac{1}{2\mu}\|S_\mu(x)-x\|_2^2. \tag{3.23}$$

It turns out from (3.21)

$$\begin{aligned}f^*(x)&=-\lambda M_{\lambda/\gamma}(x/\gamma)+\frac{1}{2\gamma}\|x\|_2^2\\&=-\lambda\|S_{\lambda/\gamma}(x/\gamma)\|_1-\frac{\gamma}{2}\|S_{\lambda/\gamma}(x/\gamma)-x/\gamma\|_2^2+\frac{1}{2\gamma}\|x\|_2^2.\end{aligned} \tag{3.24}$$

Further, by Theorem 2.1 and (3.24) we derive that, for $z \in \mathbb{R}^m$,

$$\begin{aligned} F^*(z) &= -f^*(A^\top z) - g^*(-z) \\ &= \lambda M_{\lambda/\gamma}(A^\top z/\gamma) - \frac{1}{2\gamma}\|A^\top z\|_2^2 + \langle z, b\rangle - \frac{1}{2}\|z\|_2^2 \end{aligned} \tag{3.25}$$

and

$$\begin{aligned} \varphi(w) :&= F(x) - F^*(z) \\ &= \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 + \gamma\frac{1}{2}\|x\|_2^2 - \lambda M_{\lambda/\gamma}(A^\top z/\gamma) \\ &\quad + \frac{1}{2\gamma}\|A^\top z\|_2^2 - \langle z, b\rangle + \frac{1}{2}\|z\|_2^2 \\ &=: h_1(w) + h_2(w), \end{aligned} \tag{3.26}$$

where

$$\begin{aligned} h_1(w) &= \frac{1}{2}\|Ax-b\|_2^2 + \gamma\frac{1}{2}\|x\|_2^2 - \lambda M_{\lambda/\gamma}(A^\top z/\gamma) \\ &\quad + \frac{1}{2\gamma}\|A^\top z\|_2^2 - \langle z, b\rangle + \frac{1}{2}\|z\|_2^2 && (3.27) \\ h_2(w) &= \lambda\|x\|_1. && (3.28) \end{aligned}$$

By Proposition 2.3(vi), we have, for $z \in \mathbb{R}^m$,

$$\nabla M_{\lambda/\gamma}(A^\top z/\gamma) = \frac{1}{\lambda}A\left(\frac{1}{\gamma}A^\top z - S_{\lambda/\gamma}(\frac{1}{\gamma}A^\top z)\right).$$

Thus h_1 is differentiable and its gradient is given by

$$\nabla h_1(w) = \begin{bmatrix} A^\top(Ax-b) + \gamma x \\ z + AS_{\lambda/\gamma}\left(\frac{1}{\gamma}A^\top z\right) - b \end{bmatrix}. \tag{3.29}$$

It follows that for $w, w' \in \mathbb{R}^{n+m}$,

$$\begin{aligned} \|\nabla h_1(w) - \nabla h_1(w')\|_2^2 &= \|(\gamma I + A^\top A)(x-x')\|_2^2 \\ &\quad + \|z - z' + A[S_{\lambda/\gamma}(\frac{1}{\gamma}A^\top z) - S_{\lambda/\gamma}(\frac{1}{\gamma}A^\top z')]\|_2^2. \end{aligned} \tag{3.30}$$

Since $S_{\lambda/\gamma}$ is nonexpansive, we get

$$\|S_{\lambda/\gamma}(\frac{1}{\gamma}A^\top z) - S_{\lambda/\gamma}(\frac{1}{\gamma}A^\top z')\|_2 \leq \frac{1}{\gamma}\|A^\top(z-z')\|_2 \leq \frac{1}{\gamma}\|A^\top\|_2\|z-z'\|_2.$$

Moreover, we derive from (3.30) that

$$\begin{aligned} \|\nabla h_1(w) - \nabla h_1(w')\|_2^2 &\leq \|\gamma I + A^\top A\|_2^2\|x-x'\|_2^2 + (1 + \frac{1}{\gamma}\|A\|_2\|A^\top\|_2)^2\|z-z'\|_2^2 \\ &\leq (\gamma + \|A^\top A\|_2)^2\|x-x'\|_2^2 + (1 + \frac{1}{\gamma}\|A\|_2\|A^\top\|_2)^2\|z-z'\|_2^2 \end{aligned}$$

Consequently, we get

$$\|\nabla h_1(w) - \nabla h_1(w')\|_2 \leq \max\left\{\gamma + \|A\|_2^2, 1 + \frac{1}{\gamma}\|A\|_2^2\right\} \|w - w'\|_2 = L\|w - w'\|_2. \tag{3.31}$$

Here

$$L = \max\left\{\gamma + \|A\|_2^2, 1 + \frac{1}{\gamma}\|A\|_2^2\right\}. \tag{3.32}$$

It is not hard to see that if $0 < \gamma < 1$, then $L = 1 + \frac{1}{\gamma}\|A\|_2^2$, and if $\gamma \geq 1$, then $L = \gamma + \|A\|_2^2$.

The proximal mapping of h_2 is given by the following formula.

$$\mathrm{prox}_{\mu h_2}(w) = \begin{bmatrix} \mathrm{prox}_{\mu\|\cdot\|_1}(x) \\ z \end{bmatrix} \tag{3.33}$$

for $\mu > 0$ and $w = (x, z) \in \mathbb{R}^{n+m}$.

From (3.26), the dual minimization problem is established as the minimization below:

$$\min_{w\in\mathbb{R}^{n+m}} \varphi(w) = F(x) - F^*(z) = h_1(w) + h_2(w), \tag{3.34}$$

where $h_1(w)$ and $h_2(w)$ are defined in (3.27) and (3.28), respectively. The proximal-gradient method for the composite optimization (2.5) is applicable, and we the following result.

Theorem 3.3. *Let $\{w_k\}$ be generated by the proximal gradient method:*

$$w_{k+1} = prox_{\mu h_2}(w_k - \mu\nabla h_1(w_k)), \quad k = 0, 1, \cdots, \tag{3.35}$$

where the stepsize $0 < \mu < \frac{2}{L}$ and L is the Lipschitz constant of ∇h_1 as defined in (3.32). *Then (w_k) converges to an optimal solution of* (3.34).

Notice that the PGA (3.35) can be split into x-iterates and z-iterates as follows:

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{\mu\|\cdot\|_1}[(1-\mu\gamma)x_k - \mu A^\top A x_k + \mu A^\top b] \\ z_{k+1} &= (1-\mu)z_k - \mu A S_{\lambda/\gamma}(A^\top z_k/\gamma) + \mu b. \end{aligned} \tag{3.36}$$

4. Conclusion

In this paper we have worked out certain primal-dual fixed point methods for the least-squares problem regularized by l_p-norms with $p \in [1, 2]$. The case of $p = 1$ corresponds to the lasso and the case of $p = 2$ to the ridge regression problem. The latter case, which has been studied by Ribeiro and Richtarik [9] and Silva, et al [10], is a quadratic and smooth optimization problem, and thus has linear optimality conditions. Their fixed point algorithms [9, 10] are governed by a linear operator. It turns out that the convergence of these algorithms are equivalent to the spectral radius of that operator being less than one (with respect to an appropriate norm). The other cases of $p \in [1, 2)$ correspond however to nonlinear, nonsmooth optimization problems, seemingly more complicated. In this case we have converted the coupled primal-dual problem to a composite minimization problem in the product space of the prime and dual variables, and then successfully applied the proximal-gradient method to solve the problem. We have proved convergence to an optimal solution of the iterates generated by several primal-dual proximal-gradient algorithms to cope with different situations arisen from different p's, including the elastic net [14].

It would be an interesting problem to extend the approaches by Ribeiro and Richtarik [9], Silva, et al [10], and ours in this article to more general optimization problems.

References

[1] V. Barbu, T. Precupanu, *Convexity and Optimization in Banach Spaces* (4th Ed.), Springer Dordrecht Heidelberg London New York, 2012. DOI 10.1007/978-94-007-2247-7

[2] H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (2nd Ed.), Springer Science + Business Media, 2011. DOI 10.1007/978-3-319-48311-5

[3] W. Cao, H.K. Xu, *An $l_1 - l_p$ DC regularization method for compressed sensing,* J. Nonlinear Convex Anal., **21**(2020), no. 9, 1889-1901.

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

[5] 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.

[6] I. Daubechies, M. Defrise, C. De Mol, *Sparsity-enforcing regularisation and ISTA revisited,* Inverse Problems, **32**(2016), 104001 (15 pp). doi:10.1088/0266-5611/32/10/104001

[7] C.A. Micchelli1, L. Shen, Y. Xu, *Proximity algorithms for image models: Denoising,* Inverse Problems, **27**(2011), 045009 (30 pp).

[8] B. Peng, H.K. Xu, *Proximal methods for reweighted l_Q-regularization of sparse signal recovery,* Appl. Math. Comput., **386**(2020), 125408. https://doi.org/10.1016/j.amc.2020.125408

[9] A.A. Ribeiro, P. Richtarik, *The complexity of primal-dual fixed point methods for ridge regression,* Linear Algebra and Its Applications, **556**(2018), 342-372.

[10] T.C. Silva, A.A. Ribeiro, G.A. Pericaro, *A new accelerated algorithm for ill-conditioned ridge regression problems,* Comp. Appl. Math., **37**(2018), 1941-1958. https://doi.org/10.1007/s40314-017-0430-4

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

[12] H.K. Xu, *Averaged mappings and the gradient-projection algorithm,* J. Optim. Theory Appl., **150**(2011), 360-378.

[13] H. K. Xu, *Properties and iterative methods for the lasso and its variants,* Chin. Ann. Math., **35B(3)** (2014), 501-518. DOI: 10.1007/s11401-014-0829-9

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

Received: May 10, 2022; Accepted: July 11, 2022.