Chin. Ann. Math.
35B(3), 2014, 501–518
DOI: 10.1007/s11401-014-0829-9

Chinese Annals of Mathematics, Series B
© The Editorial Office of CAM and
Springer-Verlag Berlin Heidelberg 2014

Properties and Iterative Methods for the Lasso and Its Variants*

Hong-Kun XU[1]

Abstract The lasso of Tibshirani (1996) is a least-squares problem regularized by the ℓ_1 norm. Due to the sparseness promoting property of the ℓ_1 norm, the lasso has been received much attention in recent years. In this paper some basic properties of the lasso and two variants of it are exploited. Moreover, the proximal method and its variants such as the relaxed proximal algorithm and a dual method for solving the lasso by iterative algorithms are presented.

Keywords Lasso, Elastic net, Smooth-lasso, ℓ_1 regularization, Sparsity, Proximal method, Dual method, Projection, Thresholding
2010 MR Subject Classification 47J06, 47J25, 49J40, 49N45, 65J20, 65J22

1 Introduction

The lasso, abbreviation of "the least absolute shrinkage and selection operator", was introduced by Tibshirani [16] in 1996, and is formulated as the minimization problem

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 \quad \text{subject to } \|x\|_1 \le t, \tag{1.1}$$

where A is an $m \times n$ (real) matrix, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $t \ge 0$ is a tuning parameter. An equivalent formulation of (1.1) is the following regularized minimization problem:

$$\min_x \ \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1, \tag{1.2}$$

where $\gamma \ge 0$ is a regularization parameter.

The lasso has been received much attention due to the involvement of the ℓ_1 norm which promotes sparsity, phenomenon of many practical problems arising from image/signal processing, machine learning, and so on. As a matter of fact, in imaging science, an image of interest is to be recovered from a set of linear measurements taken randomly. Mathematically this can be modeled by the linear system

$$Ax = b, \tag{1.3}$$

Manuscript received June 9, 2013.
[1] Department of Applied Mathematics, National Sun Yat-Sen University, Kaohsiung 80424, Taiwan; Department of Mathematics, King Abdulaziz University, P. O. Box 80203, Jeddah 21589, Saudi Arabia. E-mail: xuhk@math.nsysu.edu.tw
*This work was supported by NSC 102-2115-M-110-001-MY3.

where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$ is an input, and $x \in \mathbb{R}^n$ represents the image of interest to be recovered. Since the dimension of x is high (i.e., n is big), and since the number of measurements is restricted due to some constraints (i.e., m is small compared to n), the system (1.3) is underdetermined (indeed, $m \ll n$), and therefore, admits infinitely many solutions (if any).

If a certain appropriate sparsity condition is imposed, then even a unique solution of the underdetermined linear system (1.3) can be achieved. This is the crucial role played by sparsity which has brought a revolution in the imaging science and which is pioneered by Donoho, Candes, Tao, Romberg, and others (see [1–4, 7]), which has given birth to a new research field-compressed sensing. Sparsity is based on the observation that, upon a cleverly chosen basis, such as a Fourier or wavelet basis, it will often be the case that an image/signal of interest is concentrated on a small subset of the basis, hence it can be well approximated by vectors with a small number of nonzero coefficients. Conversely, an image/signal is of interest if it can be described with a small number of the basis vectors. An uninteresting image/signal is actually viewed as noise as commented in [5].

Sparsity is described by the quasi ℓ_0-"norm":

$$\|x\|_0 := \#\{j : x_j \neq 0\}$$

for $x = (x_j)^t \in \mathbb{R}^n$. For a given integer $k \geq 0$, $x \in \mathbb{R}^n$ is said to be k-sparse if $\|x\|_0 \leq k$.

The sparse recovery problem is stated as finding the sparsest vector x with $Ax = b$. Put in another way, the problem is

$$\min \|x\|_0 \quad \text{subject to } Ax = b. \tag{1.4}$$

The minimization (1.4) is numerically unstable and combinatorial NP-hard, and therefore, not an ideal way to recover the sparse signal x.

The ℓ_1 norm plays an intermediation role between the ℓ_0 norm and the ℓ_2 norm since it shares the sensitivity of the ℓ_0 norm to sparsity and the convexity of the unit ball that the ℓ_2 norm has. However, minimizing the ℓ_0 norm is NP-complete and untractable, while minimizing the ℓ_2 norm, though much easier and efficient, requires too many measurements to guarantee an accurate recovery. Therefore, minimizing the ℓ_1 norm is ideal as it combines the parsimony of ℓ_0 and the computational efficiency of ℓ_2 (see [5]).

The theory of compressed sensing surprisingly guarantees that the NP-hard combinatorial minimization (1.4) can be exactly reconstructed under certain conditions by solving a convex polynomial-time ℓ_1-minimization.

It is known (see [1, 3]) that if a sufficiently sparse x_0 exists such that $Ax_0 = b$, then the basis pursuit (BP for short)

$$\min \|x\|_1 \quad \text{subject to } Ax = b, \tag{1.5}$$

will find it; indeed (1.5) can be recast as a linear program.

When measurements take errors (which is often the case), the exact system (1.3) turns out to be inexact:

$$Ax = b + e. \tag{1.6}$$

In this case, one looks for a vector with minimum ℓ_1 norm and within some error range, that is, the minimization problem

$$\min \|x\|_1 \quad \text{subject to } \|Ax - b\|_2 \le \varepsilon \tag{1.7}$$

or the equivalent ℓ_1 regularized minimization (1.2) will be taken into consideration.

Candes, Romberg and Tao [2] showed that if a sufficiently sparse x_0 exists such that $b = Ax_0 + e$, for some small error term $\|e\|_2 \le \varepsilon$, then the solution x^* to (1.7) will be close to x_0, that is, $\|x^* - x_0\|_2 \le C\varepsilon$, where C is a constant.

In this paper we will exploit certain basic properties of the lasso (1.1) and iterative methods for solving it. The main iterative method will be the proximal algorithm which will also be used to solve variants of the lasso such as the elastic net (see [20]) and the smooth-lasso (see [10]).

2 Properties of the Lasso

Let $\gamma > 0$ and let

$$\varphi_\gamma(x) := \frac{1}{2}\|Ax - b\|_2^2 + \gamma\|x\|_1$$

be the objective function of the lasso (1.1). Observing that φ_γ is continuous, convex, and coercive (i.e., $\varphi_\gamma(x) \to \infty$ as $\|x\|_2 \to \infty$), we have that the lasso (1.1) has a closed convex nonempty solution set which is denoted as S_γ.

Proposition 2.1 *We have the following assertions:*

(i) *A and $\|\cdot\|_1$ take constant values on the solution set S_γ, that is, $Ax_\gamma = A\widehat{x}_\gamma$ and $\|x_\gamma\|_1 = \|\widehat{x}_\gamma\|_1$ for $x_\gamma, \widehat{x}_\gamma \in S_\gamma$. Consequently, the functions*

$$\rho(\gamma) := \|x_\gamma\|_1 \quad \text{and} \quad \eta(\gamma) := \frac{1}{2}\|Ax_\gamma - b\|_2^2 \quad (x_\gamma \in S_\gamma)$$

are well-defined for $\gamma > 0$ (not depending upon a particular choice $x_\gamma \in S_\gamma$). This is mentioned in [15].

(ii) *$\rho(\gamma)$ is decreasing and continuous in $\gamma > 0$.*

(iii) *$\eta(\gamma)$ is increasing in $\gamma > 0$.*

(iv) *Ax_γ is continuous in $\gamma > 0$.*

Proof For $x_\gamma \in S_\gamma$, we have the optimality condition

$$0 \in \partial\varphi_\gamma(x_\gamma) = A^t(Ax_\gamma - b) + \gamma\partial\|x_\gamma\|_1.$$

Here A^t is the transpose of A and ∂ stands for the subdifferential in the sense of convex analysis. Equivalently,

$$-\frac{1}{\gamma}A^t(Ax_\gamma - b) \in \partial\|x_\gamma\|_1.$$

It turns out by the subdifferential inequality that

$$\gamma\|x\|_1 \geq \gamma\|x_\gamma\|_1 - \langle A^t(Ax_\gamma - b), x - x_\gamma\rangle, \quad \forall x \in \mathbb{R}^n. \tag{2.1}$$

In particular, for $\widehat{x}_\gamma \in S_\gamma$,

$$\gamma\|\widehat{x}_\gamma\|_1 \geq \gamma\|x_\gamma\|_1 - \langle A^t(Ax_\gamma - b), \widehat{x}_\gamma - x_\gamma\rangle. \tag{2.2}$$

Interchange x_γ and $\widehat{x}_\gamma$ to get

$$\gamma\|x_\gamma\|_1 \geq \gamma\|\widehat{x}_\gamma\|_1 - \langle A^t(A\widehat{x}_\gamma - b), x_\gamma - \widehat{x}_\gamma\rangle. \tag{2.3}$$

Adding up (2.2) and (2.3) yields

$$0 \geq \langle A\widehat{x}_\gamma - Ax_\gamma, A\widehat{x}_\gamma - Ax_\gamma\rangle = \|A\widehat{x}_\gamma - Ax_\gamma\|_2^2.$$

Consequently, $A\widehat{x}_\gamma = Ax_\gamma$, and (2.2)–(2.3) imply that $\|\widehat{x}_\gamma\|_1 \geq \|x_\gamma\|_1$ and $\|x_\gamma\|_1 \geq \|\widehat{x}_\gamma\|_1$, respectively. Hence $\|\widehat{x}_\gamma\|_1 = \|x_\gamma\|_1$, and the functions

$$\rho(\gamma) := \|x_\gamma\|_1 \quad \text{and} \quad \eta(\gamma) := \frac{1}{2}\|Ax_\gamma - b\|_2^2 \quad (x_\gamma \in S_\gamma)$$

are well-defined for $\gamma > 0$.

It turns out from (2.1) that, for $x_\beta \in S_\beta$,

$$\gamma\|x_\beta\|_1 \geq \gamma\|x_\gamma\|_1 - \langle A^t(Ax_\gamma - b), x_\beta - x_\gamma\rangle. \tag{2.4}$$

Interchange γ and β, and x_γ and x_β to find

$$\beta\|x_\gamma\|_1 \geq \beta\|x_\beta\|_1 - \langle A^t(Ax_\beta - b), x_\gamma - x_\beta\rangle. \tag{2.5}$$

Adding up (2.4) and (2.5) obtains

$$(\gamma - \beta)(\|x_\beta\|_1 - \|x_\gamma\|_1) \geq \|Ax_\gamma - Ax_\beta\|_2^2. \tag{2.6}$$

We therefore find that if $\gamma > \beta$, then $\|x_\beta\|_1 \geq \|x_\gamma\|_1$, that is, $\rho(\gamma)$ is nonincreasing. (2.6) also shows that Ax_γ is continuous, so is $\eta(\gamma)$ for $\gamma > 0$.

To see that the function $\eta(\gamma) = \frac{1}{2}\|Ax_\gamma - b\|_2^2$ is increasing, we use the inequality (as $x_\gamma \in S_\gamma$)

$$\frac{1}{2}\|Ax_\gamma - b\|_2^2 + \gamma\|x_\gamma\|_1 \leq \frac{1}{2}\|Ax_\beta - b\|_2^2 + \gamma\|x_\beta\|_1$$

which implies that

$$\frac{1}{2}\|Ax_\gamma - b\|_2^2 \leq \frac{1}{2}\|Ax_\beta - b\|_2^2 + \gamma(\|x_\beta\|_1 - \|x_\gamma\|_1).$$

Now if $\beta > \gamma > 0$, then as $\|x_\beta\|_1 \le \|x_\gamma\|_1$, we immediately find that $\frac{1}{2}\|Ax_\gamma - b\|_2^2 \le \frac{1}{2}\|Ax_\beta - b\|_2^2$. Namely, $\eta(\gamma) \le \eta(\beta)$.

Finally we prove the continuity of $\rho(\gamma)$ for $\gamma > 0$. Rewrite (2.4) as

$$\gamma\|x_\beta\|_1 \ge \gamma\|x_\gamma\|_1 - \langle Ax_\gamma - b, Ax_\beta - Ax_\gamma\rangle, \quad \beta > 0, \ \gamma > 0. \tag{2.7}$$

Since Ax_γ is continuous in $\gamma > 0$, it follows from (2.7) that

$$\lim_{\beta\to\gamma+} \gamma\rho(\beta) = \lim_{\beta\to\gamma+} \gamma\|x_\beta\|_1 \ge \gamma\|x_\gamma\|_1 = \gamma\rho(\gamma).$$

Hence, $\rho(\gamma+) \ge \rho(\gamma)$ which in turns implies that $\rho(\gamma+) = \rho(\gamma)$ for ρ is nonincreasing. Hence, ρ is right-continuous in $\gamma > 0$.

Taking the limit in (2.7) as $\gamma \to \beta-$ yields $\rho(\beta) \ge \rho(\beta-)$ that again implies $\rho(\beta) = \rho(\beta-)$ due to the nonincreasingness of ρ. Hence, ρ is also left-continuous. Consequently, $\rho(\gamma)$ is continuous in $\gamma > 0$.

Proposition 2.2 *We have the following assertions:*

(i) $\lim\limits_{\gamma\to 0} \eta(\gamma) = \inf\limits_x \frac{1}{2}\|Ax - b\|_2^2$.

(ii) $\lim\limits_{\gamma\to 0} \rho(\gamma) = \min\limits_{x\in S} \|x\|_1$, *where* $S := \arg\min\limits_x \|Ax - b\|_2^2$ *is assumed to be nonempty.*

Proof (i) Taking the limit as $\gamma \to 0$ in the inequality

$$\frac{1}{2}\|Ax_\gamma - b\|_2^2 + \gamma\|x_\gamma\|_1 \le \frac{1}{2}\|Ax - b\|_2^2 + \gamma\|x\|_1, \quad \forall x \in \mathbb{R}^n$$

immediately yields the result in (i).

As for (ii), we claim that $\|x_\gamma\|_1 \le \|\widetilde{x}\|_1$ for any $\widetilde{x} \in S$. As a matter of fact,

$$\begin{aligned}\frac{1}{2}\|Ax_\gamma - b\|_2^2 + \gamma\|x_\gamma\|_1 &\le \frac{1}{2}\|A\widetilde{x} - b\|_2^2 + \gamma\|\widetilde{x}\|_1\\ &\le \frac{1}{2}\|Ax_\gamma - b\|_2^2 + \gamma\|\widetilde{x}\|_1.\end{aligned}$$

It turns out that $\|x_\gamma\|_1 \le \|\widetilde{x}\|_1$. In particular, $\|x_\gamma\|_1 \le \|x^\dagger\|_1$, where $x^\dagger$ is an ℓ_1 minimum-norm element of S, that is, $\|x^\dagger\|_1 = \min\limits_{x\in S} \|x\|_1$.

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

$$\begin{aligned}\frac{1}{2}\|A\widehat{x} - b\|_2^2 &= \lim_{k\to\infty} \frac{1}{2}\|Ax_{\gamma_k} - b\|_2^2\\ &= \lim_{k\to\infty} \frac{1}{2}\|Ax_{\gamma_k} - b\|_2^2 + \gamma_k\|x_{\gamma_k}\|_1\\ &\le \lim_{k\to\infty} \frac{1}{2}\|Ax - b\|_2^2 + \gamma_k\|x\|_1 = \frac{1}{2}\|Ax - b\|_2^2.\end{aligned}$$

It turns out that $\widehat{x}$ solves the least-squares problem $\min\limits_x \frac{1}{2}\|Ax - b\|_2^2$, that is, $\widehat{x} \in S$. Consequently,

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

This suffices to ensure that the conclusion of (ii) holds.

Proposition 2.3 *If $\gamma > \|A^t b\|_\infty$, then $x_\gamma = 0$.*

Proof The optimality condition

$$-A^t(Ax_\gamma - b) \in \gamma\partial\|x_\gamma\|_1$$

implies that

$$\begin{aligned}&-(A^t(Ax_\gamma - b))_j = \gamma \cdot \operatorname{sgn}[(x_\gamma)_j], \quad \text{if } (x_\gamma)_j \neq 0,\\ &|(A^t(Ax_\gamma - b))_j| \leq \gamma, \quad \text{if } (x_\gamma)_j = 0.\end{aligned}$$

Taking $x = 2x_\gamma$ in the subdifferential inequality (2.1) yields

$$\begin{aligned}\gamma\|x_\gamma\|_1 &\geq -\langle A^t(Ax_\gamma - b), x_\gamma\rangle\\ &= -\sum_{(x_\gamma)_j \neq 0} (A^t(Ax_\gamma - b))_j (x_\gamma)_j\\ &= \sum_{(x_\gamma)_j \neq 0} \gamma \cdot [\operatorname{sgn}(x_\gamma)]_j (x_\gamma)_j\\ &= \gamma \sum_{(x_\gamma)_j \neq 0} |(x_\gamma)_j| = \gamma\|x\|_1.\end{aligned}$$

It follows that

$$\begin{aligned}\gamma\|x_\gamma\|_1 &= -\langle A^t(Ax_\gamma - b), x_\gamma\rangle = -\langle Ax_\gamma - b, Ax_\gamma\rangle \qquad (2.8)\\ &= -\|Ax_\gamma - b\|^2 - \langle x_\gamma, A^t b\rangle\\ &\leq -\langle x_\gamma, A^t b\rangle \leq \|x_\gamma\|_1 \|A^t b\|_\infty. \qquad (2.9)\end{aligned}$$

It turns out from (2.9) that if $x_\gamma \neq 0$, we must have $\gamma \leq \|A^t b\|_\infty$.

Notice that (2.8) shows that $\rho(\lambda) = \|x_\gamma\|_1$ can be determined by Ax_γ. Hence we arrive at the following characterization of solutions of the lasso (1.2).

Proposition 2.4 *Let $\gamma > 0$ and $x_\gamma \in S_\gamma$. Then $\widehat{x} \in \mathbb{R}^n$ is a solution to the lasso* (1.2) *if and only if $A\widehat{x} = Ax_\gamma$ and $\|\widehat{x}\| \leq \|x_\gamma\|$. It turns out that*

$$S_\gamma = x_\gamma + N(A) \cap B_{\rho(\gamma)}, \qquad (2.10)$$

where $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ is the null space of A, and B_r denotes the closed ball centered at the origin and with radius of $r > 0$. This shows that if we can find one solution to the lasso (1.2)*, then all solutions are found by* (2.10)*.*

Proof If $A\widehat{x} = Ax_\gamma$, then from the relation

$$\begin{aligned}\varphi_\gamma(x_\gamma) &= \frac{1}{2}\|Ax_\gamma - b\|_2^2 + \gamma\|x_\gamma\|_1\\ &\leq \frac{1}{2}\|A\widehat{x} - b\|_2^2 + \gamma\|\widehat{x}\|_1\\ &= \frac{1}{2}\|Ax_\gamma - b\|_2^2 + \gamma\|\widehat{x}\|_1,\end{aligned}$$

we obtain $\|x_\gamma\|_1 \le \|\widehat{x}\|_1$. This together with the assumption of $\|\widehat{x}\|_1 \le \|x_\gamma\|_1$ yields that $\|\widehat{x}\|_1 = \|x_\gamma\|_1$ which in turns implies that $\varphi_\gamma(\widehat{x}) = \varphi_\gamma(x_\gamma)$ and hence $\widehat{x} \in S_\gamma$.

Remark 2.1 As the solution set S_γ may contain more than one point, it is unclear what kind of continuity of the set-valued function $\gamma \to S_\gamma$ one can get.

Definition 2.1 *The function from* $(0,\infty)$ *to* $[0,\infty)\times[0,\infty)$ *defined by*

$$\ell(\gamma) := (\rho(\gamma), \eta(\gamma)), \quad \gamma > 0 \tag{2.11}$$

is referred to as the L-curve associating with the lasso (1.2).

Proposition 2.5 *The L-curve* $\ell(\gamma)$ *is continuous on* $(0,\infty)$.

Remark 2.2 For more details on L-curves and properties about the ℓ_1 regularization, the reader is referred to [15].

3 Iterative Methods

In this section we discuss the proximal iterative methods for solving the lasso (1.1). The basics are Moreau's concept of proximal operators.

3.1 Proximal operators

Let H be a Hilbert space and let $\Gamma_0(H)$ be the space of convex functions in H that are proper, lower semicontinuous and convex.

Definition 3.1 (see [13–14]) *The proximal operator of* $\varphi \in \Gamma_0(H)$ *is defined by*

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

The proximal operator of φ *of order* $\lambda > 0$ *is defined as the proximal operator of* $\lambda\varphi$, *that is,*

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

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

Proposition 3.1 (see [6, 12]) *Let* $\varphi \in \Gamma_0(H)$ *and* $\lambda \in (0,\infty)$.

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

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

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

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

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

(iv) $y \in \partial\varphi(x) \Leftrightarrow x = \operatorname{prox}_\varphi(x + y)$.

The proximal operator can have a closed-form expression in some cases as shown in the examples below (see [6]).

(a) If we take φ to be the norm of H, then

$$\operatorname{prox}_{\lambda\|\cdot\|}(x) = \begin{cases} \Big(1 - \dfrac{\lambda}{\|x\|}\Big)x, & \text{if } \|x\| > \lambda, \\ 0, & \text{if } \|x\| \le \lambda. \end{cases}$$

In particular, if $H = \mathbb{R}$, then the above operator is reduced to the scalar soft-thresholding operator:

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

(b) Let $\{e_n\}_{n=1}^\infty$ be an orthonormal basis of H and let $\{\omega_n\}$ be a sequence of real positive numbers. Define $\varphi \in \Gamma_0(H)$ by

$$\varphi(x) = \sum_{n=1}^\infty \omega_n|\langle x, e_n\rangle|.$$

Then $\operatorname{prox}_\varphi(x) = \sum\limits_{n=1}^\infty \alpha_n e_n$, where

$$\alpha_n = \operatorname{sgn}(\langle x, e_n\rangle)\max\{|\langle x, e_n\rangle| - \omega_n, 0\}.$$

Below is a restatement, in terms of proximal operators, of the resolvent identity of monotone operators.

Lemma 3.1 *The proximal identity*

$$\operatorname{prox}_{\lambda\varphi}x = \operatorname{prox}_{\mu\varphi}\Big(\frac{\mu}{\lambda}x + \Big(1 - \frac{\mu}{\lambda}\Big)\operatorname{prox}_{\lambda\varphi}x\Big) \tag{3.1}$$

holds for $\varphi \in \Gamma_0(H)$, $x \in H$, $\lambda > 0$ *and* $\mu > 0$.

3.2 Proximal algorithm

The proximal operators can be used to minimize the sum of two convex functions

$$\min_{x\in H} f(x) + g(x), \tag{3.2}$$

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 (3.2).

Proposition 3.2 *Let* $f, g \in \Gamma_0(H)$. *Let* $x^* \in H$ *and* $\lambda > 0$. *Assume that* f *is finite-valued and differentiable on* H. *Then* x^* *is a solution to* (3.2) *if and only if* x^* *solves the fixed point equation*

$$x^* = (\operatorname{prox}_{\lambda g} \circ (I - \lambda\nabla f))x^*. \tag{3.3}$$

Proof x^* is a solution to (3.2) if and only if

$$\begin{aligned}
& 0 \in \partial(f+g)x^* = \nabla f(x^*) + \partial g(x^*) \\
\Leftrightarrow\ & 0 \in x^* + \lambda\partial g(x^*) - (x^* - \lambda\nabla f(x^*)) \\
\Leftrightarrow\ & x^* - \lambda\nabla f(x^*) \in x^* + \lambda\nabla g(x^*) \\
\Leftrightarrow\ & x^* = (I + \lambda\partial g)^{-1}(x^* - \lambda\nabla f(x^*)) \\
\Leftrightarrow\ & x^* = (\operatorname{prox}_{\lambda g} \circ (I - \lambda\nabla f))x^*.
\end{aligned}$$

The fixed point equation (3.3) immediately yields the following fixed point algorithm which is also known as the proximal algorithm for solving (3.2) as follows.

Initialize $x_0 \in H$ and iterate

$$x_{n+1} = (\operatorname{prox}_{\lambda_n g} \circ (I - \lambda_n\nabla f))x_n, \tag{3.4}$$

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

Theorem 3.1 *Let $f, g \in \Gamma_0(H)$ and assume that* (3.2) *is consistent. Assume in addition that*

(i) *∇f is Lipschitz continuous on H:*

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad x, y \in H;$$

(ii) $0 < \liminf\limits_{n\to\infty}\lambda_n \le \limsup\limits_{n\to\infty}\lambda_n < \frac{2}{L}$.

Then the sequence (x_n) generated by the proximal algorithm (3.4) *converges weakly to a solution of* (3.2). *No strong convergence in general is guaranteed if* $\dim H = \infty$.

To prove Theorem 3.1, we need the concept of averaged mappings. Let $\alpha \in (0, 1)$. We say that a mapping $T : H \to H$ is an α-averaged mapping (α-av for short) if

$$T = (1 - \alpha)I + \alpha U, \quad \text{with } U : H \to H \text{ nonexpansive.}$$

It is known that projections and proximal operators are all $\frac{1}{2}$-av (equivalently, firmly nonexpansive). We will use the fact (see [18]) that under the assumption (i) of Theorem 3.1, the operator

$$V_\lambda \equiv \operatorname{prox}_{\lambda g} \circ (I - \lambda\nabla f) \tag{3.5}$$

is $\frac{2+\lambda L}{4}$-av for each $0 < \lambda < \frac{2}{L}$.

It is known (see [9]) that if T is averaged with fixed points, then for each $x \in H$, the iterates $T^n x$ converge weakly to a fixed point of T.

It is also known (see [9]) that if T is nonexpansive, then the graph of $I - T$ is demiclosed, namely, if $x_n \to x$ weakly and $(I - T)x_n \to y$ in norm, then it follows that $(I - T)x = y$. This is named the demiclosedness principle of nonexpansive mappings in Hilbert spaces.

Proof of Theorem 3.1 A sketch of the proof is given in [6]. Here we provide a slightly different proof by using the technique of averaged mappings (see [18]). Let S be the nonempty solution set of (3.2). For the sake of simplicity, we may assume, due to condition (ii), that

$$0 < a \le \lambda_n \le b < \frac{2}{L} \tag{3.6}$$

for all n, where a, b are constants.

We follow the proof of [18, Theorem 4.1] (see also [11]). As V_λ given in (3.5) is $\frac{2+\lambda L}{4}$-av, we can rewrite

$$V_\lambda \equiv \mathrm{prox}_{\lambda g} \circ (I - \lambda \nabla f) = \Big(1 - \frac{2+\lambda L}{4}\Big) I + \frac{2+\lambda L}{4} T_\lambda, \tag{3.7}$$

where T_λ is nonexpansive, that is, $\|T_\lambda x - T_\lambda y\| \le \|x - y\|,\ x, y \in H$. Putting

$$V_n = V_{\lambda_n}, \quad T_n = T_{\lambda_n}, \quad \gamma_n = \frac{2+\lambda_n L}{4},$$

we then get

$$x_{n+1} = V_n x_n = (1-\gamma_n)x_n + \gamma_n T_n x_n. \tag{3.8}$$

It turns out that, for $x^* \in S = \mathrm{Fix}(V_\gamma)$ for all $\gamma > 0$,

$$\begin{aligned} \|x_{n+1} - x^*\|^2 &= \|(1-\gamma_n)(x_n - x^*) + \gamma_n(T_n x_n - x^*)\|^2 \\ &= (1-\gamma_n)\|x_n - x^*\|^2 + \gamma_n\|T_n x_n - x^*\|^2 - \gamma_n(1-\gamma_n)\|x_n - T_n x_n\|^2 \\ &\le \|x_n - x^*\|^2 - \gamma_n(1-\gamma_n)\|x_n - T_n x_n\|^2 \\ &\le \|x_n - x^*\|^2 - \delta\|x_n - T_n x_n\|^2, \end{aligned} \tag{3.9}$$

where $\delta = \frac{(2+aL)(2-bL)}{16} > 0$. Consequently, we get $\|x_{n+1} - x^*\| \le \|x_n - x^*\|$ for all n. In particular, (x_n) is bounded and moreover,

$$\lim_{n\to\infty} \|x_n - x^*\| \quad \text{exists for every } x^* \in S. \tag{3.10}$$

We also find from (3.9) that

$$\|x_n - T_n x_n\|^2 \le \frac{1}{\delta}(\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2).$$

This implies that $\|x_n - T_n x_n\| \to 0$ which, together with (3.8), in turns implies that

$$\lim_{n\to\infty} \|x_{n+1} - x_n\| = \lim_{n\to\infty} \|x_n - V_n x_n\| = 0. \tag{3.11}$$

We next prove that

$$\omega_w(x_n) \subset S. \tag{3.12}$$

Here $\omega_w(x_n)$ is the set of all weak cluster points of (x_n). Note that (3.10) and (3.12) together guarantee that (x_n) converges weakly to a point in S and then the proof is complete. To see

(3.12) we proceed as follows. Take $\widehat{x} \in \omega_w(x_n)$ and assume that $\{x_{n_j}\}$ is a subsequence of $\{x_n\}$ weakly converging to $\widehat{x}$. Hence by (3.11), $x_{n_j+1} \to \widehat{x}$ weakly as well. Without loss of generality, We may assume $\lambda_{n_j} \to \lambda$. Then $0 < \lambda < \frac{2}{L}$, due to (3.6). Setting $T = \text{prox}_{\lambda g}(I - \lambda \nabla f)$, then T is nonexpansive. Setting

$$y_j = x_{n_j} - \lambda_{n_j} \nabla f(x_{n_j}), \quad z_j = x_{n_j} - \lambda \nabla f(x_{n_j}),$$

we get $x_{n_j+1} = V_{n_j} x_{n_j} = \text{prox}_{\lambda_{n_j} g} y_j$. Using the proximal identity of Lemma 3.1, we deduce that

$$\begin{aligned}
\|x_{n_j+1} - Tx_{n_j}\| &= \|\text{prox}_{\lambda_{n_j} g} y_j - \text{prox}_{\lambda g} z_j\| \\
&= \left\|\text{prox}_{\lambda g}\Big(\frac{\lambda}{\lambda_{n_j}} y_j + \Big(1 - \frac{\lambda}{\lambda_{n_j}}\Big)\text{prox}_{\lambda_{n_j} g} y_j\Big) - \text{prox}_{\lambda g} z_j\right\| \\
&\leq \frac{\lambda}{\lambda_{n_j}}\|y_j - z_j\| + \left|1 - \frac{\lambda}{\lambda_{n_j}}\right|\|x_{n_j+1} - z_j\| \\
&\leq \frac{\lambda}{\lambda_{n_j}}|\lambda_{n_j} - \lambda|\|\nabla f(x_{n_j})\| + \frac{|\lambda_{n_j} - \lambda|}{\lambda_{n_j}}\|x_{n_j+1} - z_j\|.
\end{aligned}$$

As (x_n) is bounded, ∇f is Lipschitz continuous (hence $\{\nabla f(x_n)\}$ is bounded), and $\lambda_{n_j} \to \lambda$, we immediately derive from the last relation that $\|x_{n_j+1} - Tx_{n_j}\| \to 0$. As a result, we find

$$\|x_{n_j} - Tx_{n_j}\| \leq \|x_{n_j} - x_{n_j+1}\| + \|x_{n_j+1} - Tx_{n_j}\| \to 0.$$

Now the demiclosedness of the nonexpansive mapping $I - T$ implies that $(I - T)\widehat{x} = 0$. Namely, $\widehat{x} \in \text{Fix}(T) = S$. Therefore, (3.12) is proven.

3.3 The relaxed proximal algorithm

The relaxed proximal algorithm generates a sequence (x_n) by the following iteration process. Initialize $x_0 \in H$ and iterate

$$x_{n+1} = (1 - \alpha_n)x_n + \alpha_n(\text{prox}_{\lambda_n g} \circ (I - \lambda_n \nabla f))x_n, \tag{3.13}$$

where $\{\alpha_n\}$ is the sequence of relaxation parameters and $\{\lambda_n\}$ is a sequence of positive real numbers.

Theorem 3.2 *Let $f, g \in \Gamma_0(H)$ and assume* (3.2) *is consistent. Assume in addition that*
(i) ∇f *is Lipschitz continuous on* H:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad x, y \in H;$$

(ii) $0 < \liminf\limits_{n\to\infty} \lambda_n \leq \limsup\limits_{n\to\infty} \lambda_n < \frac{2}{L}$;
(iii) $0 < \liminf\limits_{n\to\infty} \alpha_n \leq \limsup\limits_{n\to\infty} \alpha_n < \frac{4}{2 + L\cdot\limsup\limits_{n\to\infty}\lambda_n}$.
Then the sequence (x_n) generated by the proximal algorithm (3.4) *converges weakly to a solution of* (3.2). *No strong convergence in general is guaranteed if* $\dim H = \infty$.

Proof Set

$$V_n = \mathrm{prox}_{\lambda_n g}(I - \lambda_n \nabla f).$$

Notice that V_n can be rewritten as

$$V_n = (1 - \gamma_n)I + \gamma_n T_n,$$

where $\gamma_n = \frac{2+\lambda_n L}{4}$ and T_n is nonexpansive. We can further rewrite x_{n+1} as

$$x_{n+1} = (1 - \alpha_n \gamma_n)x_n + \alpha_n \gamma_n T_n x_n. \tag{3.14}$$

Also observe from the assumption (iii) that

$$0 < \liminf_{n\to\infty} \alpha_n \gamma_n \le \limsup_{n\to\infty} \alpha_n \gamma_n < 1. \tag{3.15}$$

Repeating the proof of Theorem 3.1, we can see (details omitted) that the relations (3.10) and (3.12) remain valid and consequently, (x_n) converges weakly to a solution of (3.2).

If we take $\lambda_n \equiv \lambda \in \big(0, \frac{2}{L}\big)$, then the relaxation parameters α_n can be chosen from a larger pool; they are allowed to be close to zero. More precisely, we have the following theorem (the proof of which is omitted here).

Theorem 3.3 *Let $f, g \in \Gamma_0(H)$ and assume that* (3.2) *is consistent. Define the sequence (x_n) by the following relaxed proximal algorithm:*

$$x_{n+1} = (1 - \alpha_n)x_n + \alpha_n \mathrm{prox}_{\lambda g}(x_n - \lambda \nabla f(x_n)). \tag{3.16}$$

Suppose that

(a) *∇f satisfies the Lipschitz continuity condition* (i) *in Theorem* 3.2;

(b) *$0 < \lambda < \frac{2}{L}$ and $0 \le \alpha_n \le \frac{2+\lambda L}{4}$ for all n;*

(c) $\sum\limits_{n=1}^{\infty} \alpha_n \big(\frac{4}{2+\lambda L} - \alpha_n\big) = \infty.$

Then (x_n) converges weakly to a solution of (3.2).

3.4 Proximal algorithms applied to lasso

For the lasso (1.2), we take $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) = \gamma \|x\|_1$. Noticing that $\nabla f(x) = A^t(Ax - b)$ which is Lipschitz continuous with constant $L = \|A\|_2^2$, we find that the proximal algorithm (3.4) is reduced to the following algorithm for solving the lasso (1.2):

$$x_{k+1} = \mathrm{prox}_{\lambda_k \gamma \|\cdot\|_1}(I - \lambda_k(A^t(Ax - b)))x_k. \tag{3.17}$$

Here we have that for $\alpha > 0$ and $x = (x_j)^t \in \mathbb{R}^n$,

$$\mathrm{prox}_{\alpha\|\cdot\|_1} = (\mathrm{prox}_{\alpha|\cdot|}(x_1), \cdots, \mathrm{prox}_{\alpha|\cdot|}(x_n))^t$$

with $\mathrm{prox}_{\alpha|\cdot|}(\beta) = \mathrm{sgn}(\beta)\max\{|\beta| - \alpha, 0\}$ for $\beta \in \mathbb{R}$.

The convergence theorem of the general proximal algorithm (1.2) reads the following for the lasso (1.2).

Theorem 3.4 *Assume* $0 < \liminf_{k\to\infty} \lambda_k \le \limsup_{k\to\infty} \lambda_k < \frac{2}{\|A\|_2^2}$. *Then the sequence* (x_k) *generated by the proximal algorithm* (3.17) *converges to a solution of the lasso* (1.2).

Remark 3.1 The relaxed proximal algorithms (3.13) and (3.16) also apply to the lasso (1.2). We however do not elaborate on the details.

3.5 A dual method

Write $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) = \|x\|_1$. Then the lasso (1.2) can be rewritten as the minimization

$$\min_{x\in\mathbb{R}^n} F(x) := f(x) + \gamma g(x). \tag{3.18}$$

The optimality condition of (3.18) is that $x^* \in \mathbb{R}^n$ solves (3.18) if and only of

$$0 \in \nabla f(x^*) + \gamma \partial g(x^*) \quad \text{or} \quad -\frac{1}{\gamma}\nabla f(x^*) \in \partial g(x^*).$$

This is equivalent, by the Young-Fenchel equality, to

$$x^* \in \partial g^*\Big(-\frac{1}{\gamma}\nabla f(x^*)\Big), \tag{3.19}$$

where g^* is the conjugate of g, that is,

$$g^*(v) := \sup_{u\in\mathbb{R}^n} \{\langle u, v\rangle - g(u)\}, \quad v \in \mathbb{R}^n. \tag{3.20}$$

Setting $y = x^* - \lambda\nabla f(x^*)$ with $\lambda > 0$, we rewrite (3.19) as

$$x^* \in \partial g^*\Big(\frac{1}{\lambda\gamma}(y - x^*)\Big). \tag{3.21}$$

Further setting $z = \frac{1}{\lambda\gamma}(y - x^*)$, we get

$$0 \in z - \frac{1}{\lambda\gamma}y + \frac{1}{\lambda\gamma}\partial g^*(z). \tag{3.22}$$

Note that (3.22) can be rewritten as

$$0 \in \partial\Big(\frac{1}{2}\Big\|z - \frac{1}{\lambda\gamma}y\Big\|_2^2 + \frac{1}{\lambda\gamma}g^*(z)\Big). \tag{3.23}$$

This is equivalent to the fact

$$z = \arg\min_{v\in\mathbb{R}^n}\Big(\frac{1}{2}\Big\|v - \frac{1}{\lambda\gamma}y\Big\|_2^2 + \frac{1}{\lambda\gamma}g^*(v)\Big). \tag{3.24}$$

Since the homogeneity of g (i.e., $g(\sigma x) = \sigma g(x)$ for all $\sigma \ge 0$ and $x \in \mathbb{R}^n$) implies that the conjugate of g, g^*, is the indicator of the set $K := \partial g(0)$:

$$g^*(v) = \begin{cases} 0, & \text{if } v \in K, \\ \infty, & \text{if } v \notin K, \end{cases} \tag{3.25}$$

it turns out that (3.24) is reduced equivalently to

$$z = \arg\min_{v\in K} \frac{1}{2}\Big\|v - \frac{1}{\lambda\gamma}y\Big\|_2^2 = P_K\Big(\frac{1}{\lambda\gamma}y\Big). \tag{3.26}$$

Here P_K is the projection from $\mathbb{R}^n$ to K. Now by the definition of z, (3.26) implies that

$$y - x^* = \lambda\gamma P_K\Big(\frac{1}{\lambda\gamma}y\Big) = P_{\lambda\gamma K}(y). \tag{3.27}$$

It follows from (3.27) that x^* is a solution of (3.18) if and only if x^* satisfies the fixed point equation

$$x^* = (I - P_{\lambda\gamma K})y = (I - P_{\lambda\gamma K})(I - \lambda\nabla f)x^*. \tag{3.28}$$

Consequently, we immediately get the following fixed point algorithm for the lasso (1.2):

$$x_{k+1} = (I - P_{\lambda_k\gamma K})(I - \lambda_k\nabla f)x_k. \tag{3.29}$$

Similarly to the proximal algorithm (3.4), we have the following convergence result for the algorithm (3.29) with the proof omitted.

Theorem 3.5 *Let* (λ_k) *satisfy the condition*

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

Then the sequence (x_k) *generated by the algorithm* (3.29) *converges to a solution of the lasso* (1.2).

Remark 3.2 Since $K = \partial\|x\|_1|_{x=0} = [-1,1]^n$, we see that for each positive number $\lambda > 0$, $P_{\lambda K}$ is the projection of the Euclidean space $\mathbb{R}^n$ to the ℓ_∞ ball with radius of λ, i.e., $\{x \in \mathbb{R}^n : \|x\|_\infty \le \lambda\}$. It is not hard to find that the algorithm (3.29) and the proximal algorithm (3.4) coincide when $g(x) = \|x\|_1$ because it is not hard to find that $\text{prox}_{\lambda\gamma K} = I - P_{\lambda\gamma K}$. Indeed if we set

$$v = \text{prox}_{\lambda\gamma K}(u) = \arg\min_{z\in\mathbb{R}^n}\Big\{g(z) + \frac{1}{2\lambda\gamma}\|z-u\|_2^2\Big\},$$

then we have

$$0 \in \partial g(v) + \frac{1}{\lambda\gamma}(v-u) \quad \text{or} \quad v \in \partial g^*\Big(\frac{1}{\lambda\gamma}(u-v)\Big).$$

This corresponds to (3.21) in the case where $v := x^*$ and $u := y$. It therefore turns out by (3.27) that $\text{prox}_{\lambda\gamma K}(u) = v = u - P_{\lambda\gamma K}u$.

4 Two Variants of the Lasso

The lasso (1.2) promotes sparsity. It is however ill-posed and regularization is needed to accommodate other purposes except for sparsity. Various variants of the lasso have therefore been proposed. Here we focus on the elastic net (see [20]) and S-lasso (see [10]). More variants, such as group lasso and sparse group lasso can be found in [8, 17, 19].

4.1 The elastic net

Zou and Hastie [20] used the ℓ_2 norm (i.e., the Tikhonov regularization) to regularize the lasso (1.2) and therefore introduced the concept of the elastic net (EN for short) which is the minimization problem

$$\min_{x\in\mathbb{R}^n} \varphi_{\gamma,\delta}(x) := \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1 + \delta\frac{1}{2}\|x\|_2^2. \tag{4.1}$$

The advantage of EN lies in its unique solvability (due to the strict convexity if the ℓ_2 norm). Let $x_{\gamma,\delta}$ denote this unique solution of EN (4.1). Set

$$\varphi_\gamma(x) := \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1 \tag{4.2}$$

and

$$\psi_\delta(x) := \frac{1}{2}\|Ax-b\|_2^2 + \delta\frac{1}{2}\|x\|_2^2, \tag{4.3}$$

which are the limits of $\varphi_{\gamma,\delta}(x)$ as $\delta \to 0$ and $\gamma \to 0$, respectively.

Proposition 4.1 *Assume that the least-squares problem*

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 \tag{4.4}$$

is consistent and let S be its nonempty set of solutions.

(i) *As $\delta \to 0$ (for each fixed $\gamma > 0$), $x_{\gamma,\delta} \to x_\gamma^\dagger$ and $x_\gamma^\dagger$ is the (ℓ_2) minimum-norm solution to the lasso* (1.2). *Moreover, as $\gamma \to 0$, every cluster point of $x_\gamma^\dagger$ is an (ℓ_1) minimum-norm solution of the least-squares problem* (4.4), *i.e., a point in the set $\arg\min\limits_{x\in S}\|x\|_1$.*

(ii) *As $\gamma \to 0$ (for each fixed $\delta > 0$), $x_{\gamma,\delta} \to \widehat{x}_\delta$ and $\widehat{x}_\delta$ is the unique solution to the ℓ_2 regularized problem*:

$$\min_x \psi_\delta(x) := \frac{1}{2}\|Ax-b\|_2^2 + \delta\frac{1}{2}\|x\|_2^2. \tag{4.5}$$

Moreover, as $\delta \to 0$, $\widehat{x}_\delta \to \widehat{x}$ which is the ℓ_2 minimal norm solution of (4.4), *that is, $\widehat{x} = \arg\min\limits_{x\in S}\|x\|_2$.*

Proof Since the subdifferential

$$\partial\varphi_{\gamma,\delta}(x) = A^t(Ax-b) + \delta x + \gamma\partial\|x\|_1,$$

it turns out that the optimality condition $0 \in \partial\varphi_{\gamma,\delta}(x_{\gamma,\delta})$ is reduced to

$$-\frac{1}{\gamma}(A^t(Ax_{\gamma,\delta}-b) + \delta x_{\gamma,\delta}) \in \partial\|x_{\gamma,\delta}\|_1. \tag{4.6}$$

Then the subdifferential inequality implies that

$$\gamma\|x\|_1 \geq \gamma\|x_{\gamma,\delta}\|_1 - \langle A^t(Ax_{\gamma,\delta}-b) + \delta x_{\gamma,\delta}, x - x_{\gamma,\delta}\rangle \tag{4.7}$$

for $x \in \mathbb{R}^n$. Replacing x with $x_{\gamma',\delta'}$ for $\gamma' > 0$ and $\delta' > 0$ yields

$$\gamma\|x_{\gamma',\delta'}\|_1 \geq \gamma\|x_{\gamma,\delta}\|_1 - \langle A^t(Ax_{\gamma,\delta} - b) + \delta x_{\gamma,\delta}, x_{\gamma',\delta'} - x_{\gamma,\delta}\rangle. \tag{4.8}$$

Interchange γ and γ' and δ and δ' to get

$$\gamma'\|x_{\gamma,\delta}\|_1 \geq \gamma'\|x_{\gamma',\delta'}\|_1 - \langle A^t(Ax_{\gamma',\delta'} - b) + \delta' x_{\gamma',\delta'}, x_{\gamma,\delta} - x_{\gamma',\delta'}\rangle. \tag{4.9}$$

Adding up (4.8) and (4.9) results in

$$\begin{aligned}(\gamma' - \gamma)(\|x_{\gamma,\delta}\|_1 - \|x_{\gamma',\delta'}\|_1) &\geq \|Ax_{\gamma,\delta} - Ax_{\gamma',\delta'}\|_2^2 + \langle \delta x_{\gamma,\delta} - \delta' x_{\gamma',\delta'}, x_{\gamma,\delta} - x_{\gamma',\delta'}\rangle \\ &\geq \|Ax_{\gamma,\delta} - Ax_{\gamma',\delta'}\|_2^2 + (\delta - \delta')\langle x_{\gamma,\delta}, x_{\gamma,\delta} - x_{\gamma',\delta'}\rangle \\ &\quad + \delta'\|x_{\gamma,\delta} - x_{\gamma',\delta'}\|_2^2. \end{aligned} \tag{4.10}$$

Since the elastic net is the Tikhonov regularization of the lasso (1.2), we know that

$$\|x_{\gamma,\delta}\|_2 \leq \|x_\gamma\|_2 \leq c\|x_\gamma\|_1 \leq c\|x\|_1, \quad x_\gamma \in S_\gamma,\ x \in S.$$

Here $S_\gamma = \arg\min_{x\in\mathbb{R}^n} \varphi_\gamma(x)$, $S = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2$, and c is a constant. It follows that $\{x_{\gamma,\delta}\}$ is bounded. Hence, it follows from (4.10) that $(\gamma, \delta) \mapsto x_{\gamma,\delta}$ is a continuous curve for $\gamma, \delta > 0$.

Now by the properties of Tikhonov's regularization, we have that for each fixed $\gamma > 0$, $x_{\gamma,\delta}$ converges as $\delta \to 0$ to the ℓ_2 minimal norm solution of the lasso (1.2), i.e., the unique element $x_\gamma^\dagger := \arg\min_{x\in S_\gamma} \|x\|_2$. Moreover, by Proposition 2.2, we find that every cluster point (as $\gamma \to 0$) of the net $(x_\gamma^\dagger)$ belongs to the set $\arg\min_{x\in S} \|x\|_1$.

Next fix $\delta > 0$ and let $\widehat{x}_\delta$ be the unique solution to the minimization (4.3). This uniqueness and Proposition 2.2 imply that $x_{\gamma,\delta} \to \widehat{x}_\delta$ as $\gamma \to 0$. Now the standard property of Tikhonov's regularization ensures that $\widehat{x}_\delta \to \arg\min_{x\in S} \|x\|_2$ as $\delta \to 0$.

The elastic net (4.1) can be solved by the proximal algorithm (3.4). Take $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \frac{1}{2}\delta\|x\|_2^2$ and $g(x) = \gamma\|x\|_1$, then the proximal algorithm (3.4) is reduced to

$$x_{k+1} = \text{prox}_{\lambda_k\gamma\|\cdot\|_1}(x_k - \lambda_k[A^t(Ax_k - b) + \delta x_k]). \tag{4.11}$$

The convergence of this algorithm is given as follows.

Theorem 4.1 *Assume*

$$0 < \liminf_{k\to\infty} \lambda_k \leq \limsup_{k\to\infty} \lambda_n < \frac{2}{\|A\|_2^2 + \delta}.$$

Then the sequence (x_k) *generated by the algorithm* (4.11) *converges to the solution of the EN* (4.1).

We can also take $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) = \gamma\|x\|_1 + \frac{1}{2}\delta\|x\|_2^2$, then $\text{prox}_{\mu g}(x) = \text{prox}_{\nu\|\cdot\|_1}\left(\frac{1}{1+\mu\delta}x\right)$ with $\nu = \frac{\mu\gamma}{1+\mu\delta}$, and the proximal algorithm (3.4) is reduced to

$$x_{k+1} = \text{prox}_{\nu_k\|\cdot\|_1}\Big(\frac{1}{1 + \delta\gamma_k}(x_k - \lambda_k A^t(Ax_k - b))\Big), \tag{4.12}$$

where $\nu_k = \frac{\gamma\lambda_k}{1+\delta\gamma_k}$. Convergence of this algorithm is given below.

Theorem 4.2 *Assume*

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

Then the sequence (x_k) *generated by the algorithm* (4.12) *converges to the solution of the EN* (4.1).

4.2 The S-lasso

The smooth-lasso (S-lasso for short) of Hebiri and van der Geer [10] is formulated as the minimization problem

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1 + \delta\sum_{j=1}^{n-1}(x_{j+1}-x_j)^2. \tag{4.13}$$

This is also an adaption to the fused lasso of Tibshirani, et al [17],

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1 + \delta\sum_{j=1}^{n-1}|x_{j+1}-x_j|. \tag{4.14}$$

A more general version of the S-lasso is the minimization

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \gamma\|x\|_1 + \delta\|Jx\|_2^2, \tag{4.15}$$

where J is a $k \times n$ matrix. The S-lasso of (4.13) corresponds to the choice of J given by

$$J = \begin{bmatrix} 1 & -1 & \cdots & 0 & 0 \\ & 1 & -1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ & & & 1 & -1 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

We can apply the proximal algorithm (3.4) to the S-lasso (4.15) by taking

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

Since $\nabla f(x) = A^t(Ax-b) + 2\delta J^t J(x)$, we find that the proximal algorithm (3.4) is reduced to the algorithm

$$x_{k+1} = \operatorname{prox}_{\lambda_k\gamma\|\cdot\|_1}(x_k - \lambda_k[A^t(Ax_k-b) + 2\delta J^t J(x_k)]). \tag{4.16}$$

The convergence of this algorithm is given below.

Theorem 4.3 *Assume*

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

Then the sequence (x_k) *generated by the algorithm* (4.16) *converges to the solution of the EN* (4.15).

References

[1] Candes, E. J., Romberg, J. and Tao, T., Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Inform. Theory*, **52**(2), 2006, 489–509.

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

[3] Candes, E. J. and Tao, T., Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, **52**(12), 2006, 5406–5425.

[4] Candes, E. J. and Wakin, M. B., An introduction to compressive sampling, IEEE Signal Processing Magazine, 2008, 21–30.

[5] Cipra, B. A., ℓ_1-magic, SIAM News, **39**(9), 2006.

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

[7] Donoho, D., Compressed sensing, *IEEE Trans. Inform. Theory*, **52**(4), 2006, 1289–1306.

[8] Friedman, J., Hastie, T. and Tibshirani, R., A note on the group lasso and a sparse group lasso, arXiv:1001.0736V1.

[9] Geobel, K. and Kirk, W. A., Topics in Metric Fixed Point Theory, Cambridge Studies in Advanced Mathematics, Vol. 28, Cambridge University Press, 1990.

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

[11] Marino, G. and Xu, H. K., Convergence of generalized proximal point algorithms, *Comm. Pure Appl. Anal.*, **3**(3), 2004, 791–808.

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

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

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

[15] Raasch, T., On the L-curve criterion in ℓ_1 regularization of linear discrete ill-posed problems, International Conference on Inverse Problems and Related Topics, Nanjing, 2012.

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

[17] Tibshirani, R., Saunders, M., Rosset, R., et al., Sparsity and smoothness via the fused lasso, *J. Royal Statist. Soc., Ser. B*, **67**, 2005, 91–108.

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

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

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