JORC (2013) 1:163–186
DOI 10.1007/s40305-013-0015-x

On the Linear Convergence of a Proximal Gradient Method for a Class of Nonsmooth Convex Minimization Problems

Haibin Zhang · Jiaojiao Jiang · Zhi-Quan Luo

Received: 7 December 2012 / Revised: 4 March 2013 / Accepted: 23 April 2013 /
Published online: 16 May 2013

Abstract We consider a class of nonsmooth convex optimization problems where the objective function is the composition of a strongly convex differentiable function with a linear mapping, regularized by the sum of both ℓ_1-norm and ℓ_2-norm of the optimization variables. This class of problems arise naturally from applications in sparse group Lasso, which is a popular technique for variable selection. An effective approach to solve such problems is by the Proximal Gradient Method (PGM). In this paper we prove a local error bound around the optimal solution set for this problem and use it to establish the linear convergence of the PGM method without assuming strong convexity of the overall objective function.

Keywords Proximal gradient method · Error bound · Linear convergence · Sparse group Lasso

1 Introduction

Consider an unconstrained nonsmooth convex optimization problem of the form

$$\min_{x\in\mathbb{R}^n} F(x) = f_1(x) + f_2(x), \tag{1}$$

This work was partially supported by the National Natural Science Foundation of China (Nos. 61179033, DMS-1015346). Part of this work was performed during a research visit by the first author to the University of Minnesota, with support from the Education Commission of Beijing Municipal Government.

H. Zhang · J. Jiang
College of Applied Science, Beijing University of Technology, Beijing 100124, China

Z.-Q. Luo (✉)
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis,
MN 55455, USA
e-mail: luozq@ece.umn.edu

where f_1 is a nonsmooth convex function given by

$$f_1(x) = \sum_{J \in \mathcal{J}} w_J \|x_J\| + \lambda \|x\|_1, \tag{2}$$

with $\mathcal{J}$ a partition of $\{1, \cdots, n\}$ and λ, $\{w_J\}_{J \in \mathcal{J}}$ some given nonnegative constants; $f_2(x)$ is a composite convex function

$$f_2(x) = h(Ax), \tag{3}$$

where $h : \mathbb{R}^m \mapsto \mathbb{R}$ is a continuously differentiable strongly convex function and $A \in \mathbb{R}^{m \times n}$ is a given matrix. Notice that unless A has full column rank (i.e., $\operatorname{rank}(A) = n$), the composite function $f_2(x)$ is not strongly convex.

1.1 Motivating Applications

Nonsmooth convex optimization problems of the form (1) arise in many contemporary statistical and signal applications [4, 24] including signal denoising, compressive sensing, sparse linear regression and high dimensional multinomial classification. To motivate the nonsmooth convex optimization problem (1), we briefly outline some application examples below.

Example 1 Suppose that we have a noisy measurement vector $d \in \mathbb{R}^m$ about an unknown sparse vector $x \in \mathbb{R}^n$, where the signal model is linear and given by

$$d \approx Ax$$

for some given matrix $A \in \mathbb{R}^{m \times n}$. A popular technique to estimate the sparse vector x is called Lasso [17, 23] which performs simultaneous estimation and variable selection. Furthermore, a related technique called group Lasso [22] acts like Lasso at the group level. Since the group Lasso does not yield sparsity within a group, a generalized model that yields sparsity at both the group and individual feature levels was proposed in [5]. This sparse group Lasso criterion is formulated as

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - d\|^2 + \sum_{J \in \mathcal{J}} w_J \|x_J\| + \lambda \|x\|_1, \tag{4}$$

where the minimization of $\|Ax - d\|^2$ has a denoising effect, while the middle term promotes or leads to sparse groups, where $\mathcal{J}$ is a partition of $\{1, 2, \cdots, n\}$ into groups. The ℓ_1-norm minimization sparsifies the solution x, and effectively selects the most significant components of x.

Obviously, the sparse Lasso problem (4) is in the form of the nonsmooth convex optimization problem (1) with $h(\cdot) = \frac{1}{2}\| \cdot - d\|^2$. Moreover, if $\lambda = 0$, (4) is reduced to group Lasso; if $w_J = 0$ for all $J \in \mathcal{J}$, (4) is exactly Lasso problem. We refer readers to [1, 9] for recent applications of the group Lasso technique.

Example 2 In logistic regression, we are given a set of n-dimensional feature vectors a_i $(i = 1, 2, \cdots, m)$, and the corresponding class labels $d_i \in \{0, 1\}$. The probability distribution of the class label d given a feature vector a and a logistic regression coefficient vector $x \in \mathbb{R}^n$ can be described by

$$p(d = 1 \mid a; x) = \frac{\exp(a^T x)}{1 + \exp(a^T x)}.$$

The logistic group Lasso technique [10] corresponds to selecting x by

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \bigl(\log\bigl(1 + \exp\bigl(a_i^T x\bigr)\bigr) - d_i a_i^T x\bigr) + \sum_{J \in \mathcal{J}} w_J \|x_J\|.$$

Again, this is in the form of the nonsmooth convex optimization problem (1) with $\lambda = 0$ and

$$h(u) = \sum_{i=1}^{m} \bigl(\log\bigl(1 + \exp(u_i)\bigr) - d_i u_i\bigr),$$

which is strongly convex in u. We refer readers to [6, 7, 13, 16, 18, 20–22] for further studies on group Lasso type of statistical techniques.

Example 3 Consider a high dimensional multinomial classification problem with K classes, N samples, and p covariates. Denote the data set as $(t_1, y_1), \cdots, (t_N, y_N)$, where for all $i = 1, \cdots, N$, $t_i \in \mathbb{R}^p$ is the observed covariate vector and $y_i \in \{1, \cdots, K\}$ is the categorical response. The covariate vectors can be organized in the $N \times p$ design matrix

$$T = (t_1 \quad t_2 \quad \cdots \quad t_N)^T,$$

and the model parameters can be grouped in a $K \times p$ matrix

$$x = [x_1 \quad x_2 \quad \cdots \quad x_p],$$

where $x_i \in \mathbb{R}^K$ denotes the parameter vector associated with the ith covariate.

Let $x_0 \in \mathbb{R}^K$. For $i = 1, \cdots, N$, we define

$$\eta^{(i)} = x_0 + x t_i, \quad \text{and} \quad q\bigl(\ell, \eta^{(i)}\bigr) = \frac{\exp(\eta_\ell^{(i)})}{\sum_{k=1}^{K} \exp(\eta_k^{(i)})}.$$

The log-likelihood function is

$$\ell(x_0, x) = \sum_{i=1}^{N} \log q\bigl(y_i, \eta^{(i)}\bigr).$$

The so called multinomial sparse group Lasso classifier [19] is given by the following optimization problem:

$$\min_{x_0 \in \mathbb{R}^K, x \in \mathbb{R}^{K \times p}} F(x) = -\ell(x_0, x) + \lambda \|x\|_1 + \sum_{J=1}^{p} w_J \|x_J\|. \tag{5}$$

The above problem (5) is can be cast in the form of (1) by adding an extra (vacuous) term $0\|x_0\|_1 + 0\|x_0\|$.

1.2 Proximal Gradient Method

A popular approach to solve the nonsmooth convex optimization problem (1) is by the so called *proximal gradient method* (PGM). For any convex function $\varphi(x)$ (possibly nonsmooth), the Moreau–Yoshida proximity operator [15] $\text{prox}_\varphi : \mathbb{R}^n \mapsto \mathbb{R}^n$ is defined as

$$\text{prox}_\varphi(x) = \arg\min_{y \in \mathbb{R}^n} \varphi(y) + \frac{1}{2}\|y - x\|^2. \tag{6}$$

Since $\frac{1}{2}\|\cdot - x\|^2$ is strongly convex and $\varphi(\cdot)$ is convex, the minimizer of (6) exists and is unique, so the prox-operator prox_φ is well-defined. The prox-operator is known to be non-expansive,

$$\left\|\text{prox}_\varphi(x) - \text{prox}_\varphi(y)\right\| \leqslant \|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

and is therefore Lipschitz continuous.

Notice that if $\varphi(x)$ is the indicator function $i_C(x)$ of a closed convex set C, then the corresponding proximity operator prox_φ becomes the standard projection operator to the convex set C. Thus prox-operator is a natural extension of the projection operator onto a convex set. For problems of large dimension, the computation of the proximity operator can be difficult due to nonsmoothness of $\varphi(\cdot)$. However, if φ has a separable structure, then the computation of the proximity operator decomposes naturally, yielding substantial efficiency. For instance, for the nonsmooth convex function $f_1(x)$ defined by (2), the proximity operator prox_{f_1} can be computed efficiently (e.g., in closed form) via the so called (group) shrinkage operator [12].

Using the proximity operator, we can write down the optimality condition for (1) as a fixed point equation

$$x = \text{prox}_{\alpha f_1}\big(x - \alpha \nabla f_2(x)\big), \tag{7}$$

for some $\alpha > 0$. The proximal gradient method (PGM) is to solve this fixed point equation via the iteration

$$x^{k+1} = \text{prox}_{\alpha_k f_1}\big(x^k - \alpha_k \nabla f_2(x^k)\big), \quad k = 0, 1, 2, \cdots, \tag{8}$$

where $\alpha_k > 0$ is a stepsize. Since the nonsmooth function f_1 (cf. (2)) has a separable structure, the resulting proximal step $\text{prox}_{\alpha_k f_1}(\cdot)$ decomposes naturally across groups (and/or coordinates) and can be computed efficiently via (group) shrinkage (see Sect. 2).

Despite its popularity, the convergence analysis of the proximal gradient method is still rather limited. For instance, it is only known [3, Theorem 3.4; or 2, Proposition 2] that if the stepsize α_k satisfies

$$0 < \underline{\alpha} \leqslant \alpha_k \leqslant \bar{\alpha} < \frac{2}{L}, \quad k = 0, 1, 2, \cdots, \tag{9}$$

where L is the Lipschitz constant for the gradient $\nabla f_2(x)$:

$$\left\| \nabla f_2(x) - \nabla f_2(y) \right\| \leqslant L \|x - y\|, \quad \forall x, y \in \mathbb{R}^n,$$

then every sequence $\{x^k\}_{k \geqslant 0}$ generated by the proximal gradient algorithm (8) converges to a solution to (1). The rate of convergence is typically sublinear $O(1/k)$ [11]. The linear rate of convergence is still unknown except for some special cases. For instance, when $f_1(x) = i_C(x)$, the indicator function of the polyhedron C, the proximal gradient method (8) has been shown [8] to be globally linearly convergent to an optimal solution of (1), so long as the function f_2 has the composite structure (3). The significance of this convergence analysis lies in the fact it does not require strong convexity of f_2. More recently, Tseng [12] has proved that the PGM is linearly convergent for the case $f_1(x) = \sum_{J \in \mathcal{J}} w_J \|x_J\|$, again without assuming the strong convexity of f_2. The latter is particularly important for the applications described in Sect. 1.1. In particular, for either Lasso, Group Lasso or Sparse Group Lasso, the number of measurements is far less than the number of unknowns. Therefore, we have $m \ll n$, so the matrix A cannot have full column rank, implying that f_2 cannot be strongly convex.

In this paper, we extend Tseng's results of [12] to the case where f_1 is given by (2), namely, $f_1(x) = \sum_{J \in \mathcal{J}} w_J \|x_J\| + \lambda \|x\|_1$. In particular, we establish the linear convergence of the proximal gradient method (8) for the class of the nonsmooth convex minimization (1). Our result implies the linear convergence of PGM (8) for the sparse group Lasso problem (4) even if A does not have full column rank. This result significantly strengthens the sublinear convergence rate of PGM in the absence of strong convexity. Similar to the analysis of [8, 12], the key step in the linear convergence proof lies in the establishment of a local error bound that bounds the distance from an iterate to the optimal solution set in terms of the optimality residual $\|x - \text{prox}_{f_1}(x - \nabla f_2(x))\|$.

2 Preliminaries

We now develop some technical preliminaries needed for the subsequent convergence analysis in the next section.

For any vector $a \in \mathbb{R}^n$, we use $\text{sign}(a)$ to denote the vector whose ith component is

$$\text{sign}(a_i) := \begin{cases} 1, & \text{if } a_i > 0, \\ -1, & \text{if } a_i < 0, \\ [-1, 1], & \text{if } a_i = 0. \end{cases}$$

With this notation, the subdifferential of f_1 (cf. (2)) [14] can be written as $\partial f_1 = (\cdots, (\partial f_1)_J, \cdots)$ with

$$\begin{aligned}\big(\partial f_1(x)\big)_J &= w_J \partial \|x_J\| + \lambda \partial \|x_J\|_1 \\ &= \begin{cases} w_J \mathcal{B} + \lambda \mathcal{B}_\infty, & \text{if } x_J = 0, \\ w_J \frac{x_J}{\|x_J\|} + \lambda\, \mathrm{sign}(x_J), & \text{if } x_J \neq 0 \end{cases}\end{aligned} \tag{10}$$

for any $J \in \mathcal{J}$, where $\mathcal{B}$ and $\mathcal{B}_\infty$ are ℓ_2-norm and ℓ_∞-norm unit balls, respectively,

$$\mathcal{B} = \big\{s \in \mathbb{R}^{|J|} \mid \|s\| \leqslant 1\big\}, \qquad \mathcal{B}_\infty = \big\{t \in \mathbb{R}^{|J|} \mid \|t\|_\infty \leqslant 1\big\}.$$

Let us now consider a generic iteration of PGM (8). For convenience, we use x, x^+ and α to denote x^k, x^{k+1} and α_k, respectively. In light of the definition of prox-operator (6), we can equivalently express the PGM iteration (8) in terms of the optimality condition for the prox operator

$$x - \alpha \nabla f_2(x) \in \alpha \partial f_1\big(x^+\big) + x^+.$$

Using the separable structure of f_1 (cf. (2)), we can break up this optimality condition to the group level:

$$x_J - \alpha\big(\nabla f_2(x)\big)_J \in \alpha\big(\partial f_1\big(x^+\big)\big)_J + x_J^+, \quad \text{for all } J \in \mathcal{J}, \tag{11}$$

where $(\partial f_1(x^+))_J$ is given by (10). Notice that for any $J \in \mathcal{J}$, the component vector x_J^+ is uniquely defined by the above optimality condition (11).

Fix any x and any $J \in \mathcal{J}$. For each $j \in J$, let us denote

$$\beta_j(\alpha) = \begin{cases} 0, & \text{if } (x - \alpha\nabla f_2(x))_J \in \alpha(w_J \mathcal{B} + \lambda \mathcal{B}_\infty) \\ & \text{or } |(x - \alpha\nabla f_2(x))_j| \leqslant \alpha\lambda, \\ (x - \alpha\nabla f_2(x))_j - \alpha\lambda\, \mathrm{sign}((x - \alpha\nabla f_2(x))_j), & \text{else.} \end{cases} \tag{12}$$

Notice that, in the second case of (12), $\beta_j(\alpha)$ is simply equal to $\mathbf{Shrink}_{[-\alpha\lambda,\alpha\lambda]}((x - \alpha\nabla f_2(x))_j)$, where the shrinkage operator is the same as that in the compressive sensing algorithms. Namely, for any $\gamma > 0$, the shrinkage operator over the interval $[-\gamma, \gamma]$ is given by

$$\mathbf{Shrink}_{[-\gamma,\gamma]}(u) = \begin{cases} 0, & \text{if } |u| \leqslant \gamma, \\ u + \gamma, & \text{if } u \leqslant -\gamma, \\ u - \gamma, & \text{if } u \geqslant \gamma. \end{cases}$$

We now provide a complete characterization of the PGM iterate (8) by further simplifying the optimality condition (11).

Proposition 1 *The PGM iterate x^+ can be computed explicitly according to*

$$x_J^+ = \begin{cases} 0, & (x - \alpha\nabla f_2(x))_J \in \alpha(w_J \mathcal{B} + \lambda \mathcal{B}_\infty), \\ \beta_J(\alpha)(1 - w_J\alpha/\|\beta_J(\alpha)\|), & \textit{else.} \end{cases} \tag{13}$$

Proof Fix any x. If $(x-\alpha\nabla f_2(x))_J \in \alpha(w_J\mathcal{B}+\lambda\mathcal{B}_\infty)$, then it follows from (10) that the optimality condition (11) is satisfied at 0, implying $x_J^+=0$ (by the uniqueness of x_J^+). The converse is also true: if $(x-\alpha\nabla f_2(x))_J \notin \alpha(w_J\mathcal{B}+\lambda\mathcal{B}_\infty)$, then $x_J^+\neq 0$, because otherwise the optimality condition (11) would be violated.

Next we assume $(x-\alpha\nabla f_2(x))_J \notin \alpha(w_J\mathcal{B}+\lambda\mathcal{B}_\infty)$ so $x_J^+\neq 0$. If, in addition, $(x-\alpha\nabla f_2(x))_j\in[-\alpha\lambda,\alpha\lambda]$, then the optimality condition (11) implies $x_j^+=0$ (simply check that the optimality condition is satisfied at the point 0, and use the uniqueness of x_j^+).

The remaining case is both $(x-\alpha\nabla f_2(x))_J \notin \alpha(w_J\mathcal{B}+\lambda\mathcal{B}_\infty)$ and $|(x-\alpha\nabla f_2(x))_j|>\alpha\lambda$. In this case, $x_j^+\neq 0$ and the optimality condition (11) implies

$$\big(x-\alpha\nabla f_2(x)\big)_j = x_j^+ + \alpha w_J\frac{x_j^+}{\|x_J^+\|} + \alpha\lambda\,\mathrm{sign}\big(x_j^+\big).$$

Since the terms on the right hand side have the same sign, it follows that $\mathrm{sign}((x-\alpha\nabla f_2(x))_j)=\mathrm{sign}(x_j^+)$. Replacing the last term by $\mathrm{sign}((x-\alpha\nabla f_2(x))_j)$ and rearranging the terms, we obtain

$$\beta_j(\alpha) := \big(x-\alpha\nabla f_2(x)\big)_j - \alpha\lambda\,\mathrm{sign}\big(\big(x-\alpha\nabla f_2(x)\big)_j\big) = x_j^+ + \alpha w_J\frac{x_j^+}{\|x_J^+\|} \tag{14}$$

which further implies

$$\sqrt{\sum_{j\in J:x_j^+\neq 0}\beta_j^2(\alpha)} = \big\|\beta_J(\alpha)\big\| = \big\|x_J^+\big\|\left(1+\alpha w_J\frac{1}{\|x_J^+\|}\right),$$

where we have used the fact that $\beta_j(\alpha)=0$ whenever $x_j^+=0$ (see the definition of $\beta_j(\alpha)$ (12)). Hence, we have

$$\big\|x_J^+\big\| = \big\|\beta_J(\alpha)\big\| - \alpha w_J.$$

Substituting this relation into (14) yields

$$x_j^+ = \beta_j(\alpha)\big(1-w_J\alpha/\big\|\beta_J(\alpha)\big\|\big)$$

which establishes the proposition. □

Proposition 1 explicitly specifies how the PGM iterate x^+ can be computed. The only part that still requires further checking is to see whether the first condition in (12) holds. This can be accomplished easily by solving the following convex quadratic programming problem:

$$\min_{t\in\mathcal{B}_\infty}\sum_{j\in J}\big(x_j-\alpha\nabla f_2(x)_j-\alpha\lambda t_j\big)^2. \tag{15}$$

By Proposition 1, if the minimum value of (15) is less than or equal to $\alpha^2 w_J^2$, then we set $x_J^+ = 0$; else set

$$x_J^+ = \beta_J(\alpha)\bigl(1 - w_J\alpha / \|\beta_J(\alpha)\|\bigr), \tag{16}$$

where $\beta_J(\alpha)$ is defined by (12). In fact, due to the separable structure of the cost function, the minimum of (15) is attained at

$$t_j = \mathrm{proj}_{[-1,1]}\bigl(\bigl(x - \alpha\nabla f_2(x)\bigr)_j / \alpha\lambda\bigr), \quad j \in J$$

and the minimum value is simply

$$\sum_{j\in J} \beta_j^2(\alpha) = \|\beta_J(\alpha)\|^2,$$

where $\beta_j(\alpha)$ is defined by (12).

In light of the preceding discussion, the updating formula (13) in Proposition 1 can be rewritten as

$$x_J^+ = \begin{cases} 0, & \text{if } \|\beta_J(\alpha)\| \leqslant \alpha w_J, \\ \beta_J(\alpha)(1 - w_J\alpha/\|\beta_J(\alpha)\|), & \text{otherwise.} \end{cases} \tag{17}$$

Hence, we can summarize the PGM iteration as follows:

Proximal Gradient Method (PGM)

Step 1 Given initial guess x^0 and a small positive number ε, set $k = 0$.

Step 2 Select a step size α_k by some rule (e.g., Armijo), for $J \in \mathcal{J}$, set

$$x_J^{k+1} = \begin{cases} 0, & \text{if } \|\beta_J(\alpha_k)\| \leqslant \alpha_k w_J; \\ \beta_J(\alpha_k)(1 - w_J\alpha_k/\|\beta_J(\alpha_k)\|), & \text{otherwise,} \end{cases}$$

where $\beta_j(\alpha_k)$ is defined by (12).

Step 3 If $\|x^{k+1} - x^k\| \leqslant \varepsilon$, then stop, else, set $k = k + 1$, go to **Step 2**.

Another useful property in the analysis of PGM is the fact that Ax is invariant over the optimal solution set of (1). Denote the optimal solution set of (1) by

$$\bar{X} = \Bigl\{x^* \in \mathbb{R}^n \mid F\bigl(x^*\bigr) = \min_x F(x)\Bigr\}. \tag{18}$$

Proposition 2 *Consider the nonsmooth convex minimization problem* $\min_x F(x) = f_1(x) + f_2(x)$, *where* f_1 *and* f_2 *are given by* (2) *and* (3), *respectively. Then* Ax^* *is invariant over* $\bar{X}$ *in the sense that there exists* $\bar{y} \in \operatorname{dom} h$ *such that*

$$Ax^* = \bar{y}, \quad \forall x^* \in \bar{X}. \tag{19}$$

Proof The argument is similar to Lemma 2.1 in [8]. Since $F(x) = f_1(x) + f_2(x) = f_1(x) + h(Ax)$ is continuous and convex, the optimal solution set $\bar{X}$ must be closed and convex. For any $x^*,\ \tilde{x} \in \bar{X}$, we have by the convexity of $\bar{X}$ that $(x^* + \tilde{x})/2 \in \bar{X}$. It follows that

$$F\big(x^*\big) = F(\tilde{x}) = \frac{1}{2}\big(F\big(x^*\big) + F(\tilde{x})\big) = F\left(\frac{x^* + \tilde{x}}{2}\right)$$

which further implies

$$\frac{1}{2}\big(f_1\big(x^*\big) + h\big(Ax^*\big) + f_1(\tilde{x}) + h(A\tilde{x})\big) = f_1\left(\frac{x^* + \tilde{x}}{2}\right) + h\left(A\left(\frac{x^* + \tilde{x}}{2}\right)\right).$$

By the convexity of $h(\cdot)$, we have

$$\frac{h(Ax^*) + h(A\tilde{x})}{2} \geqslant h\left(A\left(\frac{x^* + \tilde{x}}{2}\right)\right).$$

Combining the above two relations, we obtain

$$\frac{f_1(x^*) + f_1(\tilde{x})}{2} \leqslant f_1\left(\frac{x^* + \tilde{x}}{2}\right).$$

By the convexity of $f_1(x)$, it follows that

$$\frac{f_1(x^*) + f_1(\tilde{x})}{2} \geqslant f_1\left(\frac{x^* + \tilde{x}}{2}\right).$$

This implies that $\frac{f_1(x^*)+f_1(\tilde{x})}{2} = f_1(\frac{x^*+\tilde{x}}{2})$. Therefore, we obtain

$$\frac{h(Ax^*) + h(A\tilde{x})}{2} = h\left(A\left(\frac{x^* + \tilde{x}}{2}\right)\right).$$

By the strict convexity of $h(\cdot)$, we must have $Ax^* = A\tilde{x}$. □

3 Error Bound Condition

The global convergence of PGM is given by [3, Theorem 3.4]. In particular, under the following three assumptions:

Assumptions

(A1) f_1 is a lower semi-continuous convex functions from $\mathbb{R}^n$ to $(-\infty, +\infty)$ such that dom $f \neq \emptyset$;

(A2) $f_2 : \mathbb{R}^n \mapsto \mathbb{R}$ is convex with a L-Lipschitz continuous gradient ∇f_2, i.e.,

$$\big\|\nabla f_2(x) - \nabla f_2(y)\big\| \leqslant L\|x - y\|, \quad \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n$$

where $L > 0$;

(A3) $f_1(x) + f_2(x) \to +\infty$ as $\|x\| \to +\infty$.

and if the stepsize α_k satisfies

$$0 < \underline{\alpha} \leqslant \alpha_k \leqslant \bar{\alpha} < \frac{2}{L}, \quad k = 0, 1, 2, \cdots,$$

then every sequence $\{x^k\}_{k \geqslant 0}$ generated by proximal gradient algorithm converges to a solution to (1).

Let us now focus on the linear convergence of PGM. Traditionally, linear convergence of a first order optimization method is only possible under strong convexity and smoothness assumptions. Unfortunately in our case, the objective function $F(x)$ in (1) is neither smooth nor strongly convex. To establish linear convergence in the absence of strong convexity, we rely on the following error bound condition which estimates the distance from an iterate to the optimal solution set.

Error Bound Condition Let us define a distance function for the optimal solution set $\bar{X}$ (cf. (18)) as

$$\operatorname{dist}_{\bar{X}}(x) = \inf_{y \in \bar{X}} \|x - y\|$$

and define a residual function

$$r(x) = \operatorname{prox}_{f_1}\big(x - \nabla f_2(x)\big) - x. \tag{20}$$

Since the prox-operator is Lipschitz continuous (in fact non-expansive), it follows that the residual function $r(x)$ is continuous on dom f_2.

We say a *local error bound* holds around the optimal solution set $\bar{X}$ of (1) if for any $\xi \geqslant \min_x F$, there exist scalars $\kappa > 0$ and $\varepsilon > 0$, such that

$$\operatorname{dist}_{\bar{X}}(x) \leqslant \kappa \big\|r(x)\big\|, \quad \text{whenever } F(x) \leqslant \xi,\ \big\|r(x)\big\| \leqslant \varepsilon. \tag{21}$$

To simplify notations, we denote $g = \nabla f_2(x)$ and $\beta_J := \beta_J(1)$ (cf. (12)). In light of Proposition 1 and specializing the update formula (13) for the proximal step with $\alpha = 1$, we can write the residual function $r(x)$ as

$$r(x)_J = \begin{cases} -x_J, & \text{if } \|\beta_J\| \leqslant w_J, \\ \beta_J - x_J - w_J \beta_J / \|\beta_J\|, & \text{if } \|\beta_J\| > w_J \end{cases} \tag{22}$$

for all $J \in \mathcal{J}$.

Theorem 1 *Consider the nonsmooth convex minimization problem* (1) *with* $f_1(x)$ *and* $f_2(x)$ *defined by* (2) *and* (3). *Suppose* $f_1(x)$ *and* $f_2(x)$ *satisfy the assumptions* (A1)–(A3). *Then the error bound condition* (21) *holds.*

The proof of Theorem 1 is rather technical and extends the analysis of Tseng [12]. In particular, we need two intermediate lemmas described below. For simplicity, for any sequence $\{x^k\}_{k \geqslant 0}$ in $\mathbb{R}^n \setminus \bar{X}$, we adopt the following short notations:

$$r^k := r\big(x^k\big), \qquad \delta_k := \big\|x^k - \bar{x}^k\big\|, \qquad \bar{x}^k := \operatorname*{argmin}_{\bar{x} \in \bar{X}} \big\|x^k - \bar{x}\big\|$$

and

$$u^k := \frac{x^k - \bar{x}^k}{\delta_k} \to \bar{u} \neq 0. \tag{23}$$

Lemma 1 *Consider the nonsmooth convex minimization problem* (1) *with* f_1 *and* f_2 *defined by* (2) *and* (3), *respectively. Suppose* $f_1(x)$ *and* $f_2(x)$ *satisfy assumptions* (A1)–(A3). *Furthermore, suppose there exists a sequence* $x^1, x^2, \cdots \in \mathbb{R}^n \setminus \bar{X}$ *satisfying*

$$F\big(x^k\big) \leqslant \zeta, \quad \forall k \quad \text{and} \quad \big\{r^k\big\} \to 0, \qquad \left\{\frac{r^k}{\delta_k}\right\} \to 0 \tag{24}$$

and $A\bar{u} = 0$. *Let*

$$\hat{x}^k := \bar{x}^k + \delta_k^2 \bar{u}. \tag{25}$$

Then there exists a subsequence of $\{\hat{x}^k\}$ *along which the following*:

$$0 \in \bar{g}_J + w_J \partial \big\|\hat{x}_J^k\big\| + \lambda \partial \big\|\hat{x}_J^k\big\|_1 \tag{26}$$

is satisfied for all $J \in \mathcal{J}$.

Lemma 2 *Suppose* $f_1(x)$ *and* $f_2(x)$ *satisfy assumptions* (A1)–(A3). *Moreover, suppose there exists a sequence* $x^1, x^2, \cdots \in \mathbb{R}^n \setminus \bar{X}$ *satisfying* (24). *Then there exists a* $\kappa > 0$ *such that*

$$\big\|x^k - \bar{x}^k\big\| \leqslant \kappa \big\|Ax^k - \bar{y}\big\| \quad \forall k. \tag{27}$$

The proof of Lemmas 1–2 is relegated to Appendix A and B. Assuming these lemmas hold, we can proceed to prove Theorem 1.

Proof of Theorem 1 We argue by contradiction. Suppose there exists a $\zeta \geqslant \min F$ such that (21) fails to hold for all $\kappa > 0$ and $\varepsilon > 0$. Then there exists a sequence $x^1, x^2, \cdots \in \mathbb{R}^n \setminus \bar{X}$ satisfying (24).

Let $\bar{y} = Ax^*$ for any $x^* \in \bar{X}$ (note that $\bar{y}$ is independent of x^*, cf. Proposition 2) and let

$$g^k := \nabla f_2\big(x^k\big) = A^T \nabla h\big(Ax^k\big), \qquad \bar{g} := A^T \nabla h(\bar{y}). \tag{28}$$

By Proposition 2, $\bar{g}^k = A^T \nabla h(A\bar{x}^k) = A^T \nabla h(\bar{y}) = \bar{g}$ for all k. Since

$$r^k \in \arg\min_d f_1\big(x^k + d\big) + \frac{1}{2}\big\|d + g^k\big\|^2,$$

it follows from the convexity that

$$0 \in \partial f_1\big(x^k + r^k\big) + r^k + g^k.$$

The latter is also the optimality condition for

$$r^k \in \arg\min_d \big\langle g^k + r^k, d\big\rangle + f_1\big(x^k + d\big). \tag{29}$$

We use an argument similar to that of [12]. In particular, by evaluating the right hand side of (29) at r^k and $\bar{x}^k - x^k$, respectively, we have

$$\langle g^k + r^k, r^k \rangle + f_1\big(x^k + r^k\big) \leqslant \langle g^k + r^k, \bar{x}^k - x^k \rangle + f_1\big(\bar{x}^k\big). \tag{30}$$

Similarly, since $\bar{x}^k \in \bar{X}$ and $\bar{g}^k = \bar{g}$, it follows that

$$0 \in \arg\min_d f_1\big(\bar{x}^k + d\big) + \frac{1}{2}\big\|d + \bar{g}^k\big\|^2,$$

which is further equivalent to

$$0 \in \arg\min_d \langle \bar{g}, d \rangle + f_1\big(\bar{x}^k + d\big). \tag{31}$$

By evaluating the right hand side of (31) at 0 and $x^k + r^k - \bar{x}^k$, respectively, we obtain

$$\langle \bar{g} + 0, 0 \rangle + f_1\big(\bar{x}^k + 0\big) \leqslant \langle \bar{g}, x^k + r^k - \bar{x}^k \rangle + f_1\big(\bar{x}^k + x^k + r^k - \bar{x}^k\big),$$

i.e.

$$f_1\big(\bar{x}^k\big) \leqslant \langle \bar{g}, x^k + r^k - \bar{x}^k \rangle + f_1\big(x^k + r^k\big). \tag{32}$$

Adding (30) and (32) yields

$$\langle g^k - \bar{g}, x^k - \bar{x}^k \rangle + \big\|r^k\big\|^2 \leqslant \langle \bar{g} - g^k, r^k \rangle + \langle r^k, \bar{x}^k - x^k \rangle.$$

By (28), the strong convexity of h and Lemma 2 (cf. (27)), we obtain

$$\langle g^k - \bar{g}, x^k - \bar{x}^k \rangle = \langle \nabla h\big(Ax^k\big) - \nabla h(\bar{y}), Ax^k - \bar{y} \rangle \geqslant \sigma \big\|Ax^k - \bar{y}\big\|^2 \geqslant \frac{\sigma}{\kappa^2}\big\|x^k - \bar{x}^k\big\|^2.$$

Moreover, since $\|A\| := \max_{\|d\|=1} \|Ad\|$, it follows that

$$\langle \bar{g} - g^k, r^k \rangle = \langle \nabla h(\bar{y}) - \nabla h\big(Ax^k\big), Ar^k \rangle \leqslant L\|A\|^2 \big\|x^k - \bar{x}^k\big\| \big\|r^k\big\|.$$

Combining the above three inequalities gives

$$\frac{\sigma}{\kappa^2}\big\|x^k - \bar{x}^k\big\|^2 + \big\|r^k\big\|^2 \leqslant L\|A\|^2 \big\|x^k - \bar{x}^k\big\| \big\|r^k\big\| + \big\|r^k\big\| \big\|x^k - \bar{x}^k\big\|,$$

which further implies

$$\frac{\sigma}{\kappa^2}\big\|x^k - \bar{x}^k\big\|^2 \leqslant \big(L\|A\|^2 + 1\big)\big\|x^k - \bar{x}^k\big\| \big\|r^k\big\|, \quad \forall k.$$

Canceling out a factor of $\|x^k - \bar{x}^k\|$ yields

$$\frac{\sigma}{\kappa^2}\big\|x^k - \bar{x}^k\big\| \leqslant \big(L\|A\|^2 + 1\big)\big\|r^k\big\|, \quad \forall k,$$

which contradicts (24). This completes the proof of Theorem 1. □

4 Linear Convergence

We now establish the linear convergence of the PGM (8) under the local error bound condition (21). Let $F(x) = f_1(x) + f_2(x)$ where f_1 and f_2 are defined by (2) and (3), respectively. Suppose that ∇f_2 is Lipschitz continuous with modulus L. Let $\{x^k\}_{k\geqslant 0}$ be a sequence generated by the PGM (8). There are three key steps in the linear convergence proof which we outline below. The framework was first established by Luo and Tseng in 1992 [8].

Step 1 **Sufficient decrease.** Suppose the step size α_k is chosen according to (9), but with $\bar{\alpha} < \frac{1}{L}$. Then for all $k \geqslant 0$, we have

$$F\big(x^k\big) - F\big(x^{k+1}\big) \geqslant c_1 \big\|x^{k+1} - x^k\big\|^2, \quad \text{for some } c_1 > 0. \tag{33}$$

Step 2 **Local error bound.** Let $\bar{X}$ denote the set of optimal solutions satisfying (7) and let $\operatorname{dist}_{\bar{X}}(x) := \min_{x^*\in\bar{X}} \|x - x^*\|$. Then for any $\xi \geqslant \min F(x)$, there exist some $\kappa,\ \varepsilon > 0$ such that

$$\operatorname{dist}_{\bar{X}}(x) \leqslant \kappa \big\|x - \operatorname{prox}_{f_1}\big[x - \nabla f_2(x)\big]\big\|, \tag{34}$$

for all x such that $\|x - \operatorname{prox}_{f_1}[x - \nabla f_2(x)]\| \leqslant \varepsilon$.

Step 3 **Cost-to-go estimate.** There exists a constant $c_2 > 0$ s.t.

$$F\big(x^k\big) - F^* \leqslant c_2\big(\operatorname{dist}^2_{\bar{X}}\big(x^k\big) + \big\|x^{k+1} - x^k\big\|^2\big), \quad \forall k. \tag{35}$$

We first establish the sufficient decrease property (33). Notice that the PGM iteration (8) can be equivalently written as

$$x^{k+1} = \operatorname*{argmin}_x \left\{ f_1(x) + \big\langle \nabla f_2\big(x^k\big), x - x^k \big\rangle + \frac{1}{2\alpha_k}\big\|x - x^k\big\|^2 \right\}, \quad k = 0, 1, 2, \cdots.$$

Plugging the values of $x = x^{k+1}$ and x^k, respectively, into the right hand side yields

$$f_1\big(x^{k+1}\big) + \big\langle \nabla f_2\big(x^k\big), x^{k+1} - x^k \big\rangle + \frac{1}{2\alpha_k}\big\|x^{k+1} - x^k\big\|^2 \leqslant f_1\big(x^k\big).$$

Since L is the Lipschitz constant of ∇f_2, it follows from the Taylor expansion of f_2 that

$$\begin{aligned}
F\big(x^{k+1}\big) - F\big(x^k\big) &\leqslant f_1\big(x^{k+1}\big) - f_1\big(x^k\big) + \big\langle \nabla f_2\big(x^k\big), x^{k+1} - x^k \big\rangle + \frac{L}{2}\big\|x^{k+1} - x^k\big\|^2 \\
&\leqslant -\frac{1}{2\alpha_k}\big\|x^{k+1} - x^k\big\|^2 + \frac{L}{2}\big\|x^{k+1} - x^k\big\|^2 \\
&= -\left(\frac{1}{2\alpha_k} - \frac{L}{2}\right)\big\|x^{k+1} - x^k\big\|^2 \leqslant -\frac{1 - \bar{\alpha}L}{2\bar{\alpha}}\big\|x^{k+1} - x^k\big\|^2
\end{aligned}$$

where the last step is due to (9). Since $\bar{\alpha} < 1/L$, it follows that the sufficient decrease condition (33) holds for all $k \geqslant 0$ with

$$c_1 = \frac{1 - \bar{\alpha}L}{2\bar{\alpha}} > 0.$$

The local error bound condition holds due to Theorem 1. So we need to establish the cost-to-go estimate (35). Let $\bar{x}^k \in \bar{X}$ be s.t. $\operatorname{dist}_{\bar{X}}(x^k) = \|x^k - \bar{x}^k\|$. The optimality of x^{k+1} implies

$$\begin{aligned}
&f_1\big(\bar{x}^k\big) + \big\langle \nabla f_2\big(x^k\big), \bar{x}^k - x^k \big\rangle + \frac{1}{2\alpha_k}\big\|\bar{x}^k - x^k\big\|^2 \\
&\quad \geqslant f_1\big(x^{k+1}\big) + \big\langle \nabla f_2\big(x^k\big), x^{k+1} - x^k \big\rangle + \frac{1}{2\alpha_k}\big\|x^{k+1} - x^k\big\|^2
\end{aligned}$$

implying

$$\big\langle \nabla f_2\big(x^k\big), x^{k+1} - \bar{x}^k \big\rangle + f_1\big(x^{k+1}\big) - f_1\big(\bar{x}^k\big) \leqslant \frac{1}{2\alpha_k}\operatorname{dist}^2_{\bar{X}}\big(x^k\big) \leqslant \frac{1}{2\underline{\alpha}}\operatorname{dist}^2_{\bar{X}}\big(x^k\big).$$

Also, the mean value theorem shows

$$f_2\big(x^{k+1}\big) - f_2\big(\bar{x}^k\big) = \big\langle \nabla f_2\big(\eta^k\big), x^{k+1} - \bar{x}^k \big\rangle$$

for some η^k in the line segment joining x^{k+1} and $\bar{x}^k$. Combining the above two relations and using the triangular inequality

$$\big\|\eta^k - x^k\big\| \leqslant \big\|x^{k+1} - x^k\big\| + \big\|\bar{x}^k - x^k\big\| = \big\|x^{k+1} - x^k\big\| + \operatorname{dist}_{\bar{X}}\big(x^k\big)$$

yields

$$\begin{aligned}
F\big(x^{k+1}\big) - F\big(\bar{x}^k\big) &= f_1\big(x^{k+1}\big) + f_2\big(x^{k+1}\big) - f_1\big(\bar{x}^k\big) - f_2\big(\bar{x}^k\big) \\
&= \big\langle \nabla f_2\big(\eta^k\big), x^{k+1} - \bar{x}^k \big\rangle + f_1\big(x^{k+1}\big) - f_1\big(\bar{x}^k\big) \\
&= \big\langle \nabla f_2\big(x^k\big), x^{k+1} - \bar{x}^k \big\rangle + \big\langle \nabla f_2\big(\eta^k\big) - \nabla f_2\big(x^k\big), x^{k+1} - \bar{x}^k \big\rangle \\
&\quad + f_1\big(x^{k+1}\big) - f_1\big(\bar{x}^k\big) \\
&\leqslant \big\langle \nabla f_2\big(x^k\big), x^{k+1} - \bar{x}^k \big\rangle + L\big\|\eta^k - x^k\big\|\big\|x^{k+1} - \bar{x}^k\big\| \\
&\quad + f_1\big(x^{k+1}\big) - f_1\big(\bar{x}^k\big) \\
&\leqslant \frac{1}{2\underline{\alpha}}\operatorname{dist}^2_{\bar{X}}\big(x^k\big) + L\big(\big\|x^{k+1} - x^k\big\| + \big\|x^k - \bar{x}^k\big\|\big)\big(\big\|x^{k+1} - x^k\big\| \\
&\quad + \operatorname{dist}_{\bar{X}}\big(x^k\big)\big) \\
&= \frac{1}{2\underline{\alpha}}\operatorname{dist}^2_{\bar{X}}\big(x^k\big) + L\big(\big\|x^{k+1} - x^k\big\| + \operatorname{dist}_{\bar{X}}\big(x^k\big)\big)^2 \\
&= O\big(\big\|x^{k+1} - x^k\big\|^2 + \operatorname{dist}^2_{\bar{X}}\big(x^k\big)\big),
\end{aligned}$$

where the first inequality is due to the Lipschitz continuity of ∇f_2 and the second inequality follows from the triangular inequality. This establishes the cost-to-go estimate (35). We are now ready to combine the three steps outlined above to establish the following main linear convergence result.

Theorem 2 *Assume that f_1 is convex and f_2 is convex differentiable with a Lipschitz continuous ∇f_2. Moreover, suppose $\bar{X}$ is nonempty and a local error bound (34) holds around the solution set $\bar{X}$, and that the step size α_k is chosen according to*

$$0 < \underline{\alpha} \leqslant \alpha_k \leqslant \bar{\alpha} < 1/L, \quad k = 0, 1, 2, \cdots.$$

Then the PGM algorithm (8) generates a sequence of iterates $x^0, x^1, \cdots, x^k, \cdots$ that converges linearly to a solution in $\bar{X}$.

Proof First, the sufficient decrease condition (33) implies $\|x^{k+1} - x^k\|^2 \to 0$. Since

$$\begin{aligned}\big\|x^k - \operatorname{prox}_{f_1}\big[x^k - \nabla f_2\big(x^k\big)\big]\big\| &\leqslant \frac{1}{\underline{\alpha}}\big\|x^k - \operatorname{prox}_{\alpha_k f_1}\big[x^k - \alpha_k \nabla f_2\big(x^k\big)\big]\big\| \\ &= \frac{1}{\underline{\alpha}}\big\|x^k - x^{k+1}\big\|,\end{aligned}$$

where we used the fact $\liminf_k \alpha_k \geqslant \underline{\alpha}$, it follows that

$$\big\|x^k - \operatorname{prox}_{f_1}\big[x^k - \nabla f_2\big(x^k\big)\big]\big\| \to 0.$$

Since the function values $F(x^k)$ is monotonically decreasing (33), it follows that the local error bound (34) holds for some κ and ϵ. In particular, for sufficiently large k, we have

$$\operatorname{dist}_{\bar{X}}\big(x^k\big) \leqslant \kappa \big\|x^k - \operatorname{prox}_{f_1}\big[x^k - \nabla f_2\big(x^k\big)\big]\big\|$$

implying $\operatorname{dist}_{\bar{X}}(x^k) \to 0$. Consequently, by the cost-to-go estimate (35) we have

$$F\big(x^k\big) \to F^*.$$

Now we use the local error bound (34) and the cost-to-go estimate (35) to obtain

$$\begin{aligned}F\big(x^{k+1}\big) - F^* &\leqslant c_2\big(\operatorname{dist}^2_{\bar{X}}\big(x^k\big) + \big\|x^{k+1} - x^k\big\|^2\big) \\ &\leqslant c_2\big(\kappa^2\big\|x^k - \operatorname{prox}_{f_1}\big[x^k - \nabla f_2\big(x^k\big)\big]\big\|^2 + \big\|x^{k+1} - x^k\big\|^2\big) \\ &\leqslant \frac{\kappa^2 c_2}{\min\{1, \alpha_k^2\}}\big\|x^k - \operatorname{prox}_{\alpha_k f_1}\big[x^k - \alpha_k \nabla f_2\big(x^k\big)\big]\big\|^2 + c_2\big\|x^{k+1} - x^k\big\|^2 \\ &\leqslant \frac{\kappa^2 c_2 + c_2}{\min\{1, \underline{\alpha}^2\}}\big\|x^k - x^{k+1}\big\|^2 \\ &\leqslant \frac{\kappa^2 c_2 + c_2}{c_1 \min\{1, \underline{\alpha}^2\}}\big(F\big(x^k\big) - F\big(x^{k+1}\big)\big).\end{aligned}$$

Hence, we have

$$F\big(x^{k+1}\big) - F^* \leqslant \frac{c_3}{1 + c_3}\big(F\big(x^k\big) - F^*\big),$$

where

$$c_3 = \frac{\kappa^2 c_2 + c_2}{c_1 \min\{1, \underline{\alpha}^2\}}.$$

This implies the Q-linear convergence of $F(x^k) \to F^*$. In light of (33), this further implies the R-linear convergence of $\|x^{k+1} - x^k\|^2$. Thus, $\{x^k\}$ converges linearly to an optimal solution in $\bar{X}$. □

5 Closing Remarks

Motivated by the recent applications in sparse group Lasso, we have considered in this paper a class of nonsmooth convex minimization problems whose objective function is the sum of a smooth convex function, a nonsmooth ℓ_1-norm regularization term and an ℓ_2-norm regularization term. We have derived a proximal gradient method for this problem whose subproblem can be solved efficiently (in closed form). Moreover, we have established linear convergence of this method when the smooth part of the objective function consists of a strongly convex function composed with a linear mapping, even though the overall objective function is not strongly convex and the problem may have multiple solutions. The key step in the analysis is a local error bound condition which provides an estimate of the distance to the optimal solution set in terms of the size of the proximal gradient vector.

Appendix A: Proof of Lemma 1

For each $J \in \mathcal{J}$, if $\bar{u}_J = 0$, then $\hat{x}_J^k = \bar{x}_J^k$ and (26) holds automatically (since $\bar{x}^k \in \bar{X}$). In the remainder of the proof, we assume $\bar{u}_J \neq 0$.

Since f_2 is given by (3), it follows from (1), (18), and (19) that

$$\bar{X} = \left\{ x \;\middle|\; \sum_{J \in \mathcal{J}} w_J \|x_J\| + \lambda \|x\|_1 = \min F - h(\bar{y}),\ Ax = \bar{y} \right\},$$

and by the positivity of λ or $w_J > 0$ for $J \in \mathcal{J}$, $\bar{X}$ must be compact. In fact, the level sets of $F(x)$ must also be compact.

Let

$$g^k := \nabla f_2\bigl(x^k\bigr) = A^T \nabla h\bigl(Ax^k\bigr), \qquad \bar{g} := A^T \nabla h(\bar{y}). \tag{36}$$

By (19) and (36), $A\bar{x}^k = \bar{y}$ and $\nabla f_2(\bar{x}^k) = \bar{g}$ for all k, where $\bar{x}^k = \operatorname{argmin}_{\bar{x} \in \bar{X}} \|\bar{x} - x^k\|$.

Since the level sets of F are compact, it follows from (24) that the sequence $\{x^k\}$ must be bounded. By further passing to a subsequence if necessary, we can assume that $x^k \to \bar{x}$ for some $\bar{x}$. By assumption (24), the sequence $\{r(x^k)\}$ converges to zero. Since the residual function $r(x)$ is continuous (cf. (20)), this implies $r(\bar{x}) = 0$, so $\bar{x} \in \bar{X}$. Hence $\delta_k = \|x^k - \bar{x}^k\| \leqslant \|x^k - \bar{x}\| \to 0$, so that $\bar{x}^k \to \bar{x}$. Also, by (36), $g^k = \nabla f_2(x^k) \to \nabla f_2(\bar{x}) = \bar{g}$. Since $f_1(x^k) \geqslant 0$, it follows that

$h(Ax^k) = F(x^k) - f_1(x^k) \leqslant F(x^k) \leqslant \zeta$ for all k. Since h is strongly convex, its level set must be compact. This implies that $\{Ax^k\}$ and $\bar{y}$ lie in some compact convex subset Y of the open convex set $\operatorname{dom} h$. By the strong convexity and the assumption that ∇h is Lipschitz continuous on Y, we have

$$\sigma \|y - \bar{y}\|^2 \leqslant \langle \nabla h(y) - \nabla h(\bar{y}), y - \bar{y} \rangle \quad \text{and} \quad \|\nabla h(y) - \nabla h(\bar{y})\| \leqslant L\|y - \bar{y}\|, \quad \forall y \in Y. \tag{37}$$

Since by assumption

$$\frac{(Ax^k - \bar{y})}{\delta_k} = \frac{A(x^k - \bar{x}^k)}{\|x^k - \bar{x}^k\|} \to A\bar{u} = 0,$$

it follows that $\|Ax^k - \bar{y}\| = o(\delta_k)$. Since Ax^k and $\bar{y}$ are in Y, the Lipschitz continuity of ∇h on Y (see (37)) and (36) yield

$$g^k = \bar{g} + O(\|Ax^k - A\bar{x}^k\|) = \bar{g} + o(\delta_k). \tag{38}$$

Consider a group $J \in \mathcal{J}$. We decompose $J = J_0^k \cup J_1^k$, where $\bar{x}_j^k = 0$, iff $j \in J_0^k$, and $\bar{x}_j^k \neq 0$, iff $j \in J_1^k$. In general, J_0^k, J_1^k vary with iteration index k. Since there are only finitely many choices for J_0^k and J_1^k, by passing onto a subsequence $\mathcal{K}_0$ if necessary, we can assume that J_0^k and J_1^k are fixed. Let us denote them simply as J_0 and J_1, respectively. Then we have for all $k \in \mathcal{K}_0$

$$J = J_0 \cup J_1, \quad \bar{x}_j^k = 0 \text{ for } j \in J_0 \text{ and } \bar{x}_j^k \neq 0 \text{ for } j \in J_1. \tag{39}$$

By further passing to a subsequence if necessary, we consider the following three cases:

(a) $\|\beta_J^k\| \leqslant w_J$, for all k;
(b) $\|\beta_J^k\| > w_J$, and $\bar{x}_J^k \neq 0$ for all k;
(c) $\|\beta_J^k\| > w_J$, and $\bar{x}_J^k = 0$ for all k.

Case (a). In this case, the formula (22) implies that $r_J^k := (r(x^k))_J = -x_J^k$ for all k. Since $r^k \to 0$ and $x^k \to \bar{x}$, it follows that $\bar{x}_J = 0$. Also, by (23) and (24),

$$u_J^k = \frac{-r_J^k - \bar{x}_J^k}{\delta_k} = \frac{o(\delta_k) - \bar{x}_J^k}{\delta_k}, \quad \text{implying } \bar{u}_J = -\lim_{k\to\infty} \frac{\bar{x}_J^k}{\delta_k}. \tag{40}$$

Since $\bar{u}_J \neq 0$, it follows that $\bar{x}_J^k \neq 0$ for sufficiently large $k \in \mathcal{K}_0$, so $J_1 \neq \emptyset$. By $\nabla f_2(\bar{x}^k) = \bar{g}$, $\bar{x}^k \in \bar{X}$ and the optimality condition, we have

$$0 \in \bar{g}_J + w_J \frac{\bar{x}_J^k}{\|\bar{x}_J^k\|} + \lambda \operatorname{sign}(\bar{x}_J^k). \tag{41}$$

We first consider the entries in $J_0 = J \setminus J_1$. Since $\bar{x}_{J_0}^k = 0$, it follows from (40) that $\bar{u}_{J_0} = 0$. By (25), we have

$$\hat{x}_{J_0}^k = \bar{x}_{J_0}^k + \delta_k^2 \bar{u}_{J_0} = 0. \tag{42}$$

Also, by (41), we have

$$0 = \bar{x}^k_{J_0} \in -\frac{\|\bar{x}^k_J\|}{w_J}\big(\bar{g}_{J_0} + \lambda\,\mathrm{sign}\big(\bar{x}^k_{J_0}\big)\big). \tag{43}$$

Since $\|\bar{x}^k_J\| \neq 0$, it follows from (42)

$$0 = \hat{x}^k_{J_0} \in -C\big(\bar{g}_{J_0} + \lambda\,\mathrm{sign}\big(\hat{x}^k_{J_0}\big)\big), \quad \forall C \in \mathbb{R}.$$

Letting $C = \frac{\|\hat{x}^k_J\|}{w_J}$, we have

$$0 = \hat{x}^k_{J_0} \in -\frac{\|\hat{x}^k_J\|}{w_J}\big(\bar{g}_{J_0} + \lambda\,\mathrm{sign}\big(\hat{x}^k_{J_0}\big)\big). \tag{44}$$

It remains to consider the entries in J_1. Since for $j \in J_1$ and $k \in \mathcal{K}_0$, $\bar{x}^k_j \neq 0$, there exist a subsequence $\mathcal{K}_1 \subseteq \mathcal{K}_0$ and a constant vector s_{J_1},

$$\mathrm{sign}\big(\bar{x}^k_{J_1}\big) = s_{J_1}, \quad \text{for } k \in \mathcal{K}_1 \subseteq \mathcal{K}_0, \tag{45}$$

and by (41) we have

$$\frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_J\|} = -\frac{1}{w_J}\big(\bar{g}_{J_1} + \lambda\,\mathrm{sign}\big(\bar{x}^k_{J_1}\big)\big) = -\frac{1}{w_J}(\bar{g}_{J_1} + \lambda s_{J_1}), \tag{46}$$

implying that $\bar{x}^k_{J_1}/\|\bar{x}^k_J\|$ is constant and parallel to $\bar{u}_{J_1}$ (cf. (40)). Hence, we have

$$-\frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_J\|} = -\lim_{k\to+\infty}\frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_J\|} = \frac{\bar{u}_{J_1}}{\|\bar{u}_J\|} = \frac{1}{w_J}\big(\bar{g}_{J_1} + \lambda\,\mathrm{sign}\big(\bar{x}^k_{J_1}\big)\big).$$

Together with the above equation and by (41), that is, $\bar{x}^k_{J_1} = -\frac{\|\bar{x}^k_J\|}{w_J}(\bar{g}_{J_1} + \lambda\,\mathrm{sign}(\bar{x}^k_{J_1}))$, we have

$$\hat{x}^k_{J_1} = \bar{x}^k_{J_1} + \delta_k^2\bar{u}_{J_1} = -\frac{\|\bar{x}^k_J\| - \delta_k^2\|\bar{u}_J\|}{w_J}\big(\bar{g}_{J_1} + \lambda\,\mathrm{sign}\big(\bar{x}^k_{J_1}\big)\big). \tag{47}$$

Because $\bar{u}_{J_1} = -\lim_{k\to\infty}\bar{x}^k_{J_1}/\delta_k$ and $\bar{u}_{J_1} \neq 0$ (from $\bar{u}_J \neq 0$ and $\bar{u}_{J_0} = 0$), for sufficiently large $k \in \mathcal{K}_1$, $\delta_k = O(\|\bar{x}^k_{J_1}\|)$, therefore, $\|\bar{x}^k_J\| - \delta_k^2\|\bar{u}_J\| > 0$ and $\mathrm{sign}(\hat{x}^k_{J_1}) = \mathrm{sign}(\bar{x}^k_{J_1}) = s_{J_1}$. Then for sufficiently large $k \in \mathcal{K}_1$, by (46), (47), the two vectors $\bar{x}^k_{J_1}$ and $\hat{x}^k_{J_1}$ are parallel so that

$$\frac{\hat{x}^k_{J_1}}{\|\hat{x}^k_{J_1}\|} = \frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_{J_1}\|}.$$

Since $\|\hat{x}^k_J\| = \|\hat{x}^k_{J_1}\|$ and $\|\bar{x}^k_{J_1}\| = \|\bar{x}^k_J\|$, we have

$$\frac{\hat{x}^k_{J_1}}{\|\hat{x}^k_J\|} = \frac{\hat{x}^k_{J_1}}{\|\hat{x}^k_{J_1}\|} = \frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_{J_1}\|} = \frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_J\|}.$$

Substituting this into (46) and using $\operatorname{sign}(\hat{x}^k_{J_1}) = s_{J_1}$ yields

$$\bar{g}_{J_1} + \lambda \operatorname{sign}\bigl(\hat{x}^k_{J_1}\bigr) + w_J \frac{\hat{x}^k_{J_1}}{\|\hat{x}^k_J\|} = 0. \tag{48}$$

By (44) and (48), we obtain (26).

Case (b). Similar to Case (a), we will show that the two vectors $\bar{x}^k_J$ and $\bar{u}_J$ are parallel to the direction $\bar{g}_J + \lambda \operatorname{sign}(\bar{x}^k_J)$, and that $\bar{x}^k_{J_1}$ can be written as (47), while $\hat{x}^k_{J_0} = \bar{x}^k_{J_0} = \bar{u}_{J_0} = 0$.

First, since $\bar{x}^k \in \bar{X}$ and $\bar{x}^k_J \neq 0$, it follows that $J_1 \neq \emptyset$ for $k \in \mathcal{K}_0$. The optimality condition and $\nabla f_2(\bar{x}^k) = \bar{g}$ imply (41), (45) and (46) hold for $k \in \mathcal{K}_1$. According to (45) and (46), the sign of various quantities are fixed for all sufficiently large $k \in \mathcal{K}_1$

$$-\operatorname{sign}(\bar{g}_{J_1}) = \operatorname{sign}\bigl(\bar{x}^k_{J_1}\bigr) = \operatorname{sign}\bigl(\bar{x}^k_{J_1} - \bar{g}_{J_1}\bigr) = s_{J_1} = \operatorname{sign}\bigl(x^k_{J_1} - g^k_{J_1}\bigr). \tag{49}$$

In light of (41), we can see

$$\bar{x}^k_{J_1} = -\frac{\|\bar{x}^k_J\|}{w_J}\bigl(\bar{g}_{J_1} + \lambda \operatorname{sign}\bigl(\bar{x}^k_{J_1}\bigr)\bigr). \tag{50}$$

We next use a limiting argument to show that $\bar{u}_{J_1}$ is also parallel to the direction $(\bar{g}_{J_1} + \lambda \operatorname{sign}(\bar{x}^k_{J_1}))$. Denote

$$\bar{\beta}^k_J := \bar{x}^k_J - \bar{g}_J - \lambda \operatorname{sign}\bigl(\bar{x}^k_J - \bar{g}_J\bigr) \tag{51}$$

and note that $\beta^k_J = x^k_J - g^k_J - \lambda \operatorname{sign}(x^k_J - g^k_J)$ (cf. (12)). It follows from (23) and (38) that

$$\beta^k_J = \bar{\beta}^k_J + \delta_k u^k_J - g^k_J + \bar{g}_J = \bar{\beta}^k_J + \delta_k u^k_J - o(\delta_k). \tag{52}$$

Using the optimality condition (41) and the property (49), we can simplify (51) as

$$\bar{\beta}^k_{J_1} = \bar{x}^k_{J_1} + w_J \frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_J\|} + \lambda \operatorname{sign}\bigl(\bar{x}^k_{J_1}\bigr) - \lambda \operatorname{sign}\bigl(\bar{x}^k_{J_1} - \bar{g}_{J_1}\bigr) = \bar{x}^k_{J_1}\left(1 + \frac{w_J}{\|\bar{x}^k_J\|}\right), \tag{53}$$

which, by $\|\bar{x}^k_{J_1}\| = \|\bar{x}^k_J\|$, further implies

$$\|\bar{\beta}^k_J\| \geqslant \|\bar{\beta}^k_{J_1}\| = \|\bar{x}^k_{J_1}\| + w_J > w_J.$$

Hence, we have $1 - w_J/\|\bar{\beta}^k_J\| \neq 0$. By $\bar{x}^k \in \bar{X}$ and (22), we obtain

$$0 = \bar{r}^k_J = \bar{\beta}^k_J - \bar{x}^k_J - w_J \bar{\beta}^k_J/\|\bar{\beta}^k_J\| = -\bar{x}^k_J + \bar{\beta}^k_J\bigl(1 - w_J/\|\bar{\beta}^k_J\|\bigr),$$

and by $\bar{x}^k_{J_0} = 0$, we have from (51)

$$\bar{\beta}^k_{J_0} = 0, \quad \text{and} \quad -\bar{g}_{J_0} - \lambda \operatorname{sign}(-\bar{g}_{J_0}) = 0. \tag{54}$$

From (53), we have

$$\frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|} = \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_{J_1}\|} = \frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_{J_1}\|} = \frac{\bar{x}^k_{J_1}}{\|\bar{x}^k_J\|}.$$

Moreover, it follows from $\bar{x}^k_{J_0} = 0$ and (54),

$$\frac{\bar{\beta}^k_{J_0}}{\|\bar{\beta}^k_J\|} = \frac{\bar{x}^k_{J_0}}{\|\bar{x}^k_J\|} = 0.$$

Therefore, we have

$$\frac{\bar{x}^k_J}{\|\bar{x}^k_J\|} = \frac{\bar{\beta}^k_J}{\|\bar{\beta}^k_J\|},$$

so by (41) and (46),

$$\frac{\bar{\beta}^k_J}{\|\bar{\beta}^k_J\|} \in -\frac{1}{w_J}\big(\bar{g}_J + \lambda\,\mathrm{sign}\big(\bar{x}^k_J\big)\big), \qquad \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|} = -\frac{1}{w_J}\big(\bar{g}_{J_1} + \lambda\,\mathrm{sign}\big(\bar{x}^k_{J_1}\big)\big). \tag{55}$$

Recall that $\bar{r}^k_J := r(\bar{x}^k_J) = 0$. It follows from (22) and (12) with $\alpha = 1$ that for sufficiently large $k \in \mathcal{K}_1$,

$$\begin{aligned} -r^k_{J_1} &= \bar{r}^k_{J_1} - r^k_{J_1} = g^k_{J_1} - \bar{g}_{J_1} + \lambda\,\mathrm{sign}\big(x^k_{J_1} - g^k_{J_1}\big) - s_{J_1} + w_J\left(\frac{\beta^k_{J_1}}{\|\beta^k_J\|} - \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|}\right) \\ &= o(\delta_k) + w_J\left(\frac{\beta^k_{J_1}}{\|\beta^k_J\|} - \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|}\right). \end{aligned} \tag{56}$$

Now, using (51), Taylor expansion and (55), we obtain

$$\begin{aligned} \frac{\beta^k_{J_1}}{\|\beta^k_J\|} - \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|} &= \frac{\bar{\beta}^k_{J_1} + \delta_k u^k_{J_1} - o(\delta_k)}{\|\beta^k_J\|} - \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|} = \frac{\bar{\beta}^k_{J_1}}{\|\beta^k_J\|} - \frac{\bar{\beta}^k_{J_1}}{\|\bar{\beta}^k_J\|} + \frac{\delta_k u^k_{J_1} - o(\delta_k)}{\|\beta^k_J\|} \\ &= -\frac{\langle \bar{\beta}^k_J, \delta_k u^k_J\rangle}{\|\bar{\beta}^k_J\|^3}\bar{\beta}^k_{J_1} + o(\delta_k) + \frac{\delta_k u^k_{J_1}}{\|\beta^k_J\|} \\ &= -\left\langle \frac{\bar{g}_{J_1} + \lambda\,\mathrm{sign}(\bar{x}^k_{J_1})}{w_J}, \frac{\delta_k u^k_{J_1}}{\|\bar{\beta}^k_J\|}\right\rangle\left(\frac{\bar{g}_{J_1} + \lambda\,\mathrm{sign}(\bar{x}^k_{J_1})}{w_J}\right) \\ &\quad + o(\delta_k) + \frac{\delta_k u^k_{J_1}}{\|\beta^k_J\|}. \end{aligned}$$

Multiplying both sides by $\frac{\|\bar{\beta}^k_J\|}{\delta_k}$, using (24) and (56), yields in the limit:

$$0 = -\left\langle \frac{\bar{g}_{J_1} + \lambda\,\mathrm{sign}(\bar{x}^k_{J_1})}{w_J}, \bar{u}_{J_1}\right\rangle\left(\frac{\bar{g}_{J_1} + \lambda\,\mathrm{sign}(\bar{x}^k_{J_1})}{w_J}\right) + \bar{u}_{J_1}. \tag{57}$$

This shows that $\bar{u}_{J_1}$ is parallel to the vector $(\bar{g}_{J_1} + \lambda\, \mathrm{sign}(\bar{x}^k_{J_1}))$. Since $\|\bar{g}_{J_1} + \lambda\, \mathrm{sign}(\bar{x}^k_{J_1})\| = w_J$ (by (46)), we have

$$\bar{u}_{J_1} = \pm \frac{\|\bar{u}_{J_1}\|}{w_J}\big(\bar{g}_{J_1} + \lambda\, \mathrm{sign}\big(\bar{x}^k_{J_1}\big)\big).$$

Combining this with (50), we obtain

$$\hat{x}^k_{J_1} = \bar{x}^k_{J_1} + \delta_k^2 \bar{u}_{J_1} = -\frac{\|\bar{x}^k_J\| \pm \delta_k^2 \|\bar{u}_{J_1}\|}{w_J}\big(\bar{g}_{J_1} + \lambda\, \mathrm{sign}\big(\bar{x}^k_{J_1}\big)\big), \tag{58}$$

where $\|\bar{x}^k_J\| - \delta_k^2\|\bar{u}_{J_1}\| > 0$ for sufficiently large $k \in \mathcal{K}_1$ (in fact, if $\bar{x}_J \neq 0$, then the statement is obvious; else $\bar{x}_J = 0$, because $\bar{u}_J = -\lim_{k\to\infty} \bar{x}^k_J/\delta_k$ and $\bar{u}_J \neq 0$, $\delta_k = O(\|\bar{x}^k_J\|)$).

Next we consider the entries in J_0. Clearly, we have $\bar{x}^k_{J_0} = 0$ by definition. We now show by a limiting argument that $\bar{u}_{J_0}$ is also zero. By (22), substituting $\beta^k_{J_0}$ into $-r^k_{J_0}$ and merging some similar items, we have

$$\begin{aligned}
-r^k_{J_0} &= -\beta^k_{J_0} + x^k_{J_0} + w_J \beta^k_{J_0}/\|\beta^k_J\| \\
&= \frac{w_J}{\|\beta^k_J\|} x^k_{J_0} + \left(1 - \frac{w_J}{\|\beta^k_J\|}\right)\big(g^k_{J_0} + \lambda\, \mathrm{sign}\big(x^k_{J_0} - g^k_{J_0}\big)\big).
\end{aligned}$$

Moreover, by (38), we have

$$\begin{aligned}
-r^k_{J_0} &= \frac{w_J}{\|\beta^k_J\|} x^k_{J_0} + \left(1 - \frac{w_J}{\|\beta^k_J\|}\right)\big(\bar{g}_{J_0} + o(\delta_k) + \lambda\, \mathrm{sign}\big(x^k_{J_0} - g^k_{J_0}\big)\big), \\
&= \frac{w_J}{\|\beta^k_J\|} x^k_{J_0} + \left(1 - \frac{w_J}{\|\beta^k_J\|}\right)\big(-\lambda\, \mathrm{sign}(\bar{x}_{J_0} - \bar{g}_{J_0}) + o(\delta_k) + \lambda\, \mathrm{sign}\big(x^k_{J_0} - g^k_{J_0}\big)\big),
\end{aligned}$$

where the second step follows from (54). Since the signs are fixed for sufficiently large $k \in \mathcal{K}_1$ (cf. (49)), we have $-\lambda\, \mathrm{sign}(\bar{x}_{J_0} - \bar{g}_{J_0}) + \lambda\, \mathrm{sign}(x^k_{J_0} - g^k_{J_0}) = 0$, therefore,

$$-r^k_{J_0} = \frac{w_J}{\|\beta^k_J\|} x^k_{J_0} + \left(1 - \frac{w_J}{\|\beta^k_J\|}\right) o(\delta_k).$$

Multiplying both sides of above equation with $\frac{1}{\delta_k}$ and letting $k \to +\infty$, we obtain by (24) that $x^k_{J_0}/\delta_k \to 0$. This further implies

$$\bar{u}_{J_0} = \lim_{k\to+\infty} \frac{x^k_{J_0} - \bar{x}^k_{J_0}}{\delta_k} = 0.$$

So, $\hat{x}^k_{J_0} = \bar{x}^k_{J_0} + \delta_k^2 \bar{u}_{J_0} = 0$. Combining this with (58) yields

$$\hat{x}^k_J = \bar{x}^k_J + \delta_k^2 \bar{u}_J \in -\frac{\|\bar{x}^k_J\| \pm \delta_k^2 \|\bar{u}_J\|}{w_J}\big(\bar{g}_J + \lambda\, \mathrm{sign}\big(\bar{x}^k_J\big)\big).$$

Using an argument identical to that for the Case (a), we can show that $\hat{x}_J^k$ satisfies (48) for sufficiently large $k \in \mathcal{K}_1$, thus establishing the desired property (26).

Case (c). In this case $0 = \bar{x}_J^k \to \bar{x}_J$, it follows from that $\bar{x}_J = 0$. By $\|\beta_J^k\| > w_J$ (the assumption for Case (c)), it follows that $x_J^k + r_J^k \neq 0$. By the optimality condition of the prox-operator we have

$$g_J^k + \lambda \operatorname{sign}\bigl(x_J^k + r_J^k\bigr) \ni -r_J^k - w_J \frac{x_J^k + r_J^k}{\|x_J^k + r_J^k\|}, \quad \forall k. \tag{59}$$

Notice that

$$\lim_{k\to\infty} \frac{r_k}{\delta_k} = 0, \qquad \lim_{k\to\infty} \frac{x_J^k}{\delta_k} = \lim_{k\to\infty} \frac{x_J^k - \bar{x}_J^k}{\delta_k} = \bar{u}_J,$$

where we have used the fact that $\bar{x}_J^k = 0$ in this case. Hence, we have

$$\lim_{k\to\infty} \frac{x_J^k + r_J^k}{\|x_J^k + r_J^k\|} = \lim_{k\to\infty} \frac{\frac{x_J^k}{\delta_k} + \frac{r_J^k}{\delta_k}}{\|\frac{x_J^k}{\delta_k} + \frac{r_J^k}{\delta_k}\|} = \frac{\bar{u}_J}{\|\bar{u}_J\|}.$$

Moreover, notice that

$$\lim_{k\to\infty} \operatorname{sign}\bigl(x_J^k + r_J^k\bigr) = \lim_{k\to\infty} \operatorname{sign}\left(\frac{x_J^k + r_J^k}{\delta_k}\right) \subseteq \operatorname{sign}\left(\lim_{k\to\infty} \frac{x_J^k + r_J^k}{\delta_k}\right) = \operatorname{sign}(\bar{u}_J).$$

Thus, by taking limit $k \to \infty$ in (59), we obtain

$$\bar{g}_J + \lambda \operatorname{sign}(\bar{u}_J) \ni -w_J \frac{\bar{u}_J}{\|\bar{u}_J\|}$$

implying

$$\bar{g}_J + \lambda \operatorname{sign}\bigl(\delta_k^2 \bar{u}_J\bigr) \ni -w_J \frac{\delta_k^2 \bar{u}_J}{\|\delta_k^2 \bar{u}_J\|}.$$

This is precisely the optimality condition for $\hat{x}_J^k = \bar{x}_J^k + \delta_k^2 \bar{u}_J = \delta_k^2 \bar{u}_J$. In other words, we have (26) as desired.

Appendix B: Proof of Lemma 2

We argue by contradiction. Suppose this is false. Then, by passing to a subsequent if necessary, we can assume that

$$\frac{\|Ax^k - \bar{y}\|}{\|x^k - \bar{x}^k\|} \to 0.$$

Since $\bar{y} = A\bar{x}^k$, this is equivalent to $\{Au^k\} \to 0$, where u^k is defined by (23). Then $\|u^k\| = 1$ for all k. By further passing to subsequence if necessary, we will assume that $u^k \to \bar{u}$ for some $\bar{u}$ with $\|\bar{u}\| = 1$. Then $A\bar{u} = 0$ and $\|\bar{u}\| = 1$. Moreover,

$$Ax^k = A\big(\bar{x}^k + \delta_k u^k\big) = \bar{y} + \delta_k Au^k = \bar{y} + o(\delta_k).$$

Since $\{u^k\} \to \bar{u}$ and $\|\bar{u}\| = 1$, we have $\langle u^k, \bar{u}\rangle \geqslant \frac{1}{2}$ for all k sufficiently large. Fix any such k and consider $\hat{x}^k$ defined by (25), namely, $\hat{x}^k = \bar{x}^k + \delta_k^2 \bar{u}$.

Since $A\bar{u} = 0$, it follows $A\hat{x}^k = A\bar{x}^k$, which further implies $\nabla f_2(\hat{x}^k) = A^T \nabla h(A\hat{x}^k) = A^T \nabla h(A\bar{x}^k) = \nabla f_2(\bar{x}^k) = \bar{g}$. By Lemma 1, since $\hat{x}^k$ satisfies (25), it follows that

$$0 \in \nabla f_2\big(\hat{x}^k\big)_J + w_J \partial \big\|\hat{x}_J^k\big\| + \lambda \partial \big\|\hat{x}_J^k\big\|_1 \tag{60}$$

for all $J \in \mathcal{J}$. Hence $\hat{x}^k \in \bar{X}$. Since $\langle x^k - \bar{x}^k, \bar{u}\rangle = \delta_k \langle u^k, \bar{u}\rangle > \delta_k/2$ and $\|\bar{u}\| = 1$, it follows that

$$\begin{aligned}\big\|x^k - \hat{x}^k\big\|^2 &= \big\|x^k - \bar{x}^k - \delta_k^2 \bar{u}\big\|^2 = \big\|x^k - \bar{x}^k\big\|^2 - 2\delta_k^2 \big\langle x^k - \bar{x}^k, \bar{u}\big\rangle + \delta_k^4 \\ &< \big\|x^k - \bar{x}^k\big\|^2 - \delta_k^3(1 - \delta_k) < \big\|x^k - \bar{x}^k\big\|^2, \quad \text{for all } 0 < \delta_k < 1,\end{aligned}$$

which contradicts $\bar{x}^k$ being the point in $\bar{X}$ nearest to x^k. This proves (27).

References

[1] Bach, F.: Consistency of the group Lasso and multiple kernel learning. J. Mach. Learn. Res. **9**, 1179–1225 (2008)

[2] Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. arXiv:0912.3522v4 [math.OC], 18 May 2010

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

[4] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1359 (2001)

[5] Friedman, J., Hastie, T., Tibshirani, R.: A note on the group Lasso and a sparse group Lasso. arXiv:1001.0736v1 [math.ST], 5 Jan 2010

[6] Kim, D., Sra, S., Dhillon, I.: A scalable trust-region algorithm with application to mixednorm regression. In: International Conference on Machine Learning (ICML), vol. 1 (2010)

[7] Liu, J., Ji, S., Ye, J.: SLEP: sparse learning with efficient projections. Arizona State University (2009)

[8] Luo, Z.Q., Tseng, P.: On the linear convergence of descent methods for convex essentially smooth minimization. SIAM J. Control Optim. **30**(2), 408–425 (1992)

[9] Ma, S., Song, X., Huang, J.: Supervised group Lasso with applications to microarray data analysis. BMC Bioinform. **8**(1), 60 (2007)

[10] Meier, L., Van de Geer, S., Buhlmann, P.: The group Lasso for logistic regression. J. R. Stat. Soc., Ser. B, Stat. Methodol. **70**(1), 53–71 (2008)

[11] Nesterov, Y.: Introductory Lectures on Convex Optimization. Kluwer, Boston (2004)

[12] Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**(2), 263–295 (2010)

[13] Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. **117**, 387–423 (2009)

[14] Rockafellar, R.T.: Convex Analysis. Princeton Univ. Press, Princeton (1970)

[15] Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, New York (1998)

[16] Roth, V., Fischer, B.: The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In: Proceedings of the 25th International Conference on Machine Learning, pp. 848–855. ACM, New York (2008)

[17] Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B **58**, 267–288 (1996)

[18] van den Berg, E., Schmidt, M., Friedlander, M., Murphy, K.: Group sparsity via linear-time projection. Technical Report TR-2008-09, Department of Computer Science, University of British Columbia (2008)

[19] Vincent, M., Hansen, N.R.: Sparse group Lasso and high dimensional multinomial classification. J. Comput. Stat. Data Anal. arXiv:1205.1245v1 [stat.ML], 6 May 2012

[20] Wright, S., Nowak, R., Figueiredo, M.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**(7), 2479–2493 (2009)

[21] Yang, H., Xu, Z., King, I., Lyu, M.: Online learning for group Lasso. In: 27th International Conference on Machine Learning (ICML2010) (2010)

[22] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Ser. B, Stat. Methodol. **68**(1), 49–67 (2006)

[23] Zou, H.: The adaptive Lasso and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)

[24] Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. B **67**(2), 301–320 (2005)