Abstract In this paper, we propose a proximal gradient algorithm for solving a general nonconvex and nonsmooth optimization model of minimizing the summation of a $C^{1,1}$ function and a grouped separable lsc function. This model includes the group sparse optimization via $\ell_{p,q}$ regularization as a special case. Our algorithmic scheme presents a unified framework for several well-known iterative thresholding algorithms. We establish some convergence results for the algorithm. We apply the proposed algorithm to the group sparse optimization problem and obtain analytical formulae for some specific $\ell_{p,q}$ regularizations. Finally, we present some numerical results to demonstrate the performance of the proposed algorithm.

Key words Group sparse optimization, $\ell_p$ regularization, nonconvex optimization, proximal gradient algorithm, iterative thresholding algorithm.

1 Introduction

In recent years, a great amount of attention has been paid to the sparse optimization problem, which is to find the sparse solutions of an under-determined linear system. The sparse optimization problem arises in a wide range of applications, such as variable selection, pattern analysis, graphical modeling and compressive sensing; see [5, 7, 9, 13, 15, 24] and references therein.

1.1 Background

In many applications, the underlying data usually can be represented approximately by a linear system of the form

$$Ax = b + \varepsilon,$$

where the matrix $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are known, $\varepsilon \in \mathbb{R}^m$ is an unknown noise vector, and $x = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$ is the variable to be estimated. If $m \ll n$, the above linear inverse
problem is seriously ill-posed and may have multiple solutions. The sparse optimization problem is to recover $x$ from information $b$ such that $x$ is of a sparse structure. The sparsity of variable $x$ has been investigated by using $\ell_p$ norm $\|x\|_p$ ($p = 0$, see [5, 6]; $p = 1$, see [2, 9, 11, 13, 24, 27, 30]; and $p = 1/2$, see [8, 28]). The $\ell_p$ norm $\|x\|_p$ of $x$ for $p > 0$ is defined as

$$\|x\|_p = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p},$$

while the $\ell_0$ norm $\|x\|_0$ denotes the number of nonzero components of $x$. Thus, the sparse optimization problem can be modeled as

$$\min \|x\|_0 \quad \text{s.t.} \quad \|Ax - b\|_2 \leq \delta,$$

or,

$$\min \|Ax - b\|_2 \quad \text{s.t.} \quad \|x\|_0 \leq K,$$

where $\delta$ is the error allowance and $K$ is the given sparsity level.

For the sparse optimization problem, a popular and practical technique is the regularization method, which is to transform the sparse optimization problem into an unconstrained optimization problem, called the regularization problem. For example, the $\ell_0$ regularization problem is

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2_2 + \lambda \|x\|_0, \quad (1.1)$$

where $\lambda > 0$ is the regularization parameter, providing a tradeoff between accuracy and sparsity. However, the $\ell_0$ regularization problem is nonconvex and non-Lipschitz, and thus it is generally intractable to solve the model (indeed, it is NP-hard; see [20]).

To overcome this difficulty, two typical relaxations of the $\ell_0$ regularization problem are introduced, which are the $\ell_1$ regularization problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2_2 + \lambda \|x\|_1 \quad (1.2)$$

and the $\ell_{1/2}$ regularization problem

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2_2 + \lambda \|x\|_{1/2} \quad (1.3)$$

The $\ell_1$ regularization model, also called Lasso [24] or Basis Pursuit [9], has attracted much attention and has been accepted as a most useful tool for the sparse optimization problem. Since the $\ell_1$ regularization problem is a convex optimization problem, many exclusive and efficient algorithms have been proposed and developed for solving (1.2), for instance, the interior-point methods [7, 9], LARs [14], the gradient projection method [16] and the alternating direction method [30]. The relationship between the $\ell_1$ and $\ell_0$ regularization model ($\ell_1/\ell_0$ equivalence) has been established in [13]: the solution of $\ell_1$ regularization problem coincides with one of the solutions of the $\ell_0$ regularization problem under some mild condition.
However, in many practical applications, the solutions obtained from the $\ell_1$ regularization problem are less sparse than those of the $\ell_0$ regularization problem, and it often leads to sub-optimal sparsity in reality; see e.g. [8, 28, 32].

Recently, the $\ell_{1/2}$ regularization model is proposed to improve the performance of sparsity recovery of the $\ell_1$ regularization model. Extensive computational studies in [8, 28] reveal that the $\ell_{1/2}$ regularization model admits a significantly stronger sparsity-promoting property than the $\ell_1$ regularization model in the sense that it allows to achieve more sparse solutions and recover a sparse signal from smaller amount of samplings. However, the $\ell_{1/2}$ regularization problem is a nonconvex, nonsmooth and non-Lipschitz optimization problem, and thus it is very difficult in general to design efficient algorithms for its solutions. It is presented in [17] that finding the global minimal value of the $\ell_{1/2}$ regularization problem (1.3) is strongly NP-hard, while fortunately, computing a local minimizer could be done in polynomial time. Some fast and efficient algorithms have been proposed to find a local minimizer of (1.3), such as the hybrid OMP-SG algorithm [10] and the interior-point potential reduction algorithm [17].

Besides the preceding numerical algorithms, the most widely studied class of the first order methods for the sparse optimization problem is various variants of the iterative thresholding algorithms; see [2, 5, 6, 11, 18, 28] and references therein. It is convergent and of very low computational complexity. Benefitting from its simple formulation and low storage requirement, it is very efficient and applicable even for large-scale sparse optimization problem. In particular, the iterative hard (resp. soft, half) thresholding algorithm for the $\ell_0$ (resp. $\ell_1$, $\ell_{1/2}$) regularization problem is studied in [5, 6] (resp. [2, 11], [28]).

In applications, a wide class of problems usually has certain special structures, and recently, enhancing the recoverability due to the special structures has become an active topic in the sparse optimization. One of the most popular sparse structures is the group sparsity, that is, the solution has a natural grouping of its components, and the components within each group are likely to be either all zeros or all nonzeros. Let $x = ([x]_1^T, \ldots, [x]_s^T)^T$ represent the group structure of $x$, where $\{[x]_i \in \mathbb{R}^{n_i} : i = 1, \ldots, s\}$ is the grouping of $x$, $[x]_i$ denotes the subvector of $x$ corresponding to the $i$-th group, and $\sum_{i=1}^s n_i = n$. The sparsity of $x = ([x]_1^T, \ldots, [x]_s^T)^T$ with a group structure can be measured by a $\ell_{p,q}$ norm. For $p \geq 1$ and $q > 0$, the $\ell_{p,q}$ norm of $x = ([x]_1^T, \ldots, [x]_s^T)^T$ is defined by

$$
\|x\|_{p,q} = \left( \sum_{i=1}^s \| [x]_i \|_p^q \right)^{1/q},
$$

while for $p \geq 1$, the $\ell_{p,0}$ norm is defined by

$$
\|x\|_{p,0} = \sum_{i=1}^s \| [x]_i \|_p^0.
$$

When $p = q \geq 1$, the $\ell_{p,q}$ norm coincides with the $\ell_p$ norm, i.e., $\|x\|_{p,p} = \|x\|_p$. Furthermore, all $\ell_{p,0}$ norms share the same formula, i.e., $\|x\|_{p,0} = \|x\|_{2,0}$ for all $p \geq 1$. In particular, when all $n_i = 1$, $\|x\|_{p,q} = \|x\|_q$ for all $p \geq 1$ and $q > 0$.
In general, the grouping can be any arbitrary partition of $x$, and it is usually pre-defined based on prior knowledge of specific problems. Exploiting the group sparsity structure can reduce the degrees of freedom in the solution, thereby leading to better recovery performance. Benefitting from these advantages, the group sparse optimization model has been extensively applied in birthweight prediction [1, 31], dynamic MRI [25] and gene finding [19, 29]. Based on the group structure, the $\ell_{2,1}$ regularization problem

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

is first introduced by Yuan and Lin [31] to study the grouped variable selection in statistics under the name of group Lasso. The $\ell_{2,1}$ regularization, an important extension of the $\ell_1$ regularization, proposes an $\ell_2$ regularization for each group and ultimately yields the sparsity in the group manner. Since the $\ell_{2,1}$ norm is a convex function and the $\ell_{2,1}$ regularization problem is a convex optimization problem, some effective algorithms have been proposed, for instance, the spectral projected gradient method [26], SpaRSA [27] and the alternating direction method [12].

1.2 The model description

Recall that the $\ell_0$ and $\ell_{p,0}$ norms can be respectively rewritten as

$$\|x\|_0 = \sum_{i=1}^{n} |x_i|^0 \quad \text{and} \quad \|x\|_{p,0} = \sum_{i=1}^{s} \| [x]_i \|_p^0,$$

adopting the convention $0^0 = 0$, and that the general $\ell_p$ and $\ell_{p,q}$ ($p > 0$) regularization terms can be respectively formulated as

$$\|x\|_p^p = \sum_{i=1}^{n} |x_i|^p \quad \text{and} \quad \|x\|_{p,q}^q = \sum_{i=1}^{s} \| [x]_i \|_p^q.$$

It is worth mentioning that $\|x\|_0, \|x\|_{p,0}, \|x\|_p^p$ and $\|x\|_{p,q}^q$ are all represented as a summation of some (grouped) separable lower semicontinuous (lsc) functions. Thus, throughout this paper, we consider the general nonconvex and nonsmooth optimization model

$$\min_{x \in \mathbb{R}^n} F(x) := f(x) + \phi(x). \quad (1.4)$$

The following assumptions are made for model (1.4) throughout the paper:

- $f : \mathbb{R}^n \to \mathbb{R}$ is a $C^{1,1}$ function, i.e., continuously differentiable with a Lipschitz continuous gradient satisfying

  $$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2, \forall x, y \in \mathbb{R}^n,$$

- $\phi(x)$ is a grouped separable function of the form $\phi(x) := \sum_{i=1}^{s} \varphi_i([x]_i)$ with each $\varphi_i : \mathbb{R}^{n_i} \to \mathbb{R}$ being a lsc function,
• $f$ and $\phi$ are both bounded from below on $\mathbb{R}^n$, i.e., $\inf_{x \in \mathbb{R}^n} \{ f(x), \phi(x) \} > -\infty$.

Evidently, the group sparse optimization via the $\ell_{p,q}$ ($p \geq 1, q \geq 0$) regularization

$$\min_{x \in \mathbb{R}^n} \| Ax - b \|_2^2 + \lambda \| x \|_{p,q}^q$$

is a special case of model (1.4) with $f(x) = \| Ax - b \|_2^2$, which is a $C^{1,1}$ function with $L = 2 \| A \|_2^2$, where $\| A \|_2 = \max_{x \neq 0} \frac{\| Ax \|_2}{\| x \|_2}$, and $\varphi_i(x_i) = \lambda \| x_i \|_{p,q}^q$.

1.3 Contributions

The main objective of this paper is to introduce the general model (1.4), propose a proximal gradient algorithm for this model and then apply the proximal gradient algorithm to the group sparse optimization.

The general model (1.4) includes a wide variety of interesting problems, especially the $\ell_p$ regularization problem. However, it is nonconvex, nonsmooth and non-Lipschitz, and thus it is very difficult in general to design efficient algorithms for solving (1.4). In this paper, we propose a proximal gradient algorithm to solve the model (1.4), which presents a unified framework of the well-known iterative thresholding algorithms, and establish some convergence properties under some mild condition on stepsize. An advantage of the proximal gradient algorithm is that it generates a non-increasing sequence of the objective function values.

We also apply the proximal gradient algorithm to the group sparse optimization. More specifically, we apply proximal gradient algorithm to solve the group sparse optimization problem via $\ell_{p,q}$ regularization, $p = 1, 2$ and $q = 0, 1/2, 1$ respectively. For the specific regularization problem, we show that the proximal step can be given explicitly by an analytical formula, and the resulting algorithm is practically attractive. In particular, when the group size is 1, the analytical formulae for each group reduce to the ones established in [5, 11, 28] for each variable.

We further perform some numerical experiments to demonstrate the performance of the proposed algorithm, applying to the $\ell_{p,q}$ regularization problem. From the numerical results, it is demonstrated that the proposed algorithm for the $\ell_{p,1/2}$ regularization problem outperforms it for the $\ell_{p,1}$ and $\ell_{p,0}$ regularization problem on speed, accuracy and robustness. This observation is consistent with several previous numerical examples on the $\ell_p$ regularization problem; see [8, 28].

1.4 Paper organization

This paper is organized as follows. In Section 2, we propose the proximal gradient algorithm and demonstrate its convergence property for the general model (1.4). In Section 3, we apply the proximal gradient algorithm to the group sparse optimization using different types of $\ell_{p,q}$ regularization. Finally, Section 4 exhibits the numerical results.
2 Proximal gradient algorithm

We first recall some basic steps in the gradient algorithm and the proximal algorithm. Consider an unconstrained optimization problem
\[ \min_{x \in \mathbb{R}^n} h(x). \] (2.1)

If \( h \) is differentiable, then the iteration formula of the gradient algorithm for solving (2.1) is given by
\[ x_{k+1} = x_k - v \nabla h(x_k), \]
where \( v > 0 \) is the stepsize.

When \( h \) is a nonsmooth convex function, the iteration formula of the proximal algorithm [21] for solving (2.1) is presented by
\[ x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \{ h(x) + \frac{1}{2v} \| x - x_k \|_2^2 \}. \] (2.2)

To solve the general nonconvex and nonsmooth optimization problem (1.4), we propose a proximal gradient algorithm, which first applies the gradient step on the function \( f \) and then performs the proximal step on the function \( \phi \). Thus, the proximal gradient algorithm for problem (1.4) is described as follows.

Proximal gradient algorithm

Select a constant stepsize \( v \), start with an initial point \( x_0 \in \mathbb{R}^n \), and generate a sequence \( \{x_k\} \subseteq \mathbb{R}^n \) via the iteration
\[ z_k = x_k - v \nabla f(x_k), \] (2.3)
\[ x_{k+1} \in \arg \min_{x \in \mathbb{R}^n} \{ \phi(x) + \frac{1}{2v} \| x - z_k \|_2^2 \}, \] (2.4)

where (2.3) is the gradient step and (2.4) is the proximal step.

Remark 2.1. The existence of solutions of (2.4) follows from [22, Theorem 1.9] and thus the proximal gradient algorithm is well-posed.

Furthermore, for the general model (1.4), the main advantage of the proximal gradient algorithm is that (2.4) can be achieved in parallel computation. Indeed, by our assumption, the function \( \phi \) and the regularization term \( \| x - z_k \|_2^2 \) are both grouped separable, and hence the proximal step (2.4) can be viewed as a cycle of low dimensional proximal regularization problems
\[ \arg \min_{[x_i] \in \mathbb{R}^{n_i}} \{ \varphi_i([x_i]) + \frac{1}{2v} \| [x_i] - [z_k]_i \|_2^2 \}, \quad i = 1, \ldots, s. \]

Remark 2.2. When \( \phi(x) \) is a convex function, the proximal regularization term \( \| x - z_k \|_2^2 \) assures a unique solution in (2.4); see [21]. When \( \phi(x) \) is in general a nonconvex function, we will show that the proximal regularization term \( \| x - z_k \|_2^2 \) can guarantee a non-increasing
property of the objective values and some convergence properties; see Theorem 2.1. When \( \phi(x) \) is chosen as \( \|x\|_2, \|x\|_0, \|x\|_{1/2}, \) or \( \|x\|_{1/2} \), we will drive the analytical solution for (2.4) in Lemma 3.1. In particular, when \( \phi(x) \) is \( \|x\|_0, \|x\|_{1/2} \) or \( \|x\|_1 \), the analytical solution of (2.4) has been obtained; see [5, 11, 28]. It is worth noting that in the proximal gradient algorithm, the main computation will be the proximal step (2.4). Thus, when analytical formula for (2.4) is available, the resulting algorithm is practically attractive.

In the following, we give an equivalent representation of the proximal gradient algorithm and show the relationship with the well-known iterative thresholding algorithms. For convenience, we define an auxiliary function

\[
Q(x, y) := f(y) + \langle \nabla f(y), x - y \rangle + \phi(x) + \frac{1}{2v} \|x - y\|_2^2,
\]

and denote the gradient operator of \( f \) and the proximal operator of \( \phi \) on \( \mathbb{R}^n \) respectively by

\[
G_f(x) := x - v \nabla f(x),
\]

and

\[
P_\phi(y) := \operatorname{Arg min}_{x \in \mathbb{R}^n} \{ \phi(x) + \frac{1}{2v} \|x - y\|_2^2 \}.
\]

Thus, the proximal gradient algorithm is equivalent to (see e.g. [2, 4])

\[
x_{k+1} \in (P_\phi \circ G_f)(x_k) = \operatorname{Arg min}_{x \in \mathbb{R}^n} \{ Q(x, x_k) \},
\]

where \( P_\phi \circ G_f \) is called the proximal gradient operator.

In particular, if \( f(x) = \|Ax - b\|_2^2, \phi(x) = \lambda \|x\|_p^p \) and \( v = 1/2 \), then the auxiliary function reduces to

\[
Q(x, y) = \|Ax - b\|_2^2 + \lambda \|x\|_p^p - \|Ax - Ay\|_2^2 + \|x - y\|_2^2,
\]

which is exactly the surrogate function of the \( \ell_p \) regularization, as in [5, 11, 28]. The proximal gradient algorithm generates the sequence \( \{x_k\} \) via the iteration of minimizing \( Q(x, x_k) \) (see (2.6)), while the iterative hard/half/soft thresholding algorithm generates the sequence \( \{x_k\} \) via the iteration of minimizing the corresponding surrogate function. Consequently, the proximal gradient algorithm presents a unified framework of the iterative thresholding algorithms, and reduces to the well-known iterative hard (resp. half, soft) thresholding algorithm for the \( \ell_0 \) (resp. \( \ell_{1/2}, \ell_1 \)) regularization problem respectively.

We first recall a well-known property of a \( C^{1,1} \) function, which is repeatedly used in this paper, so as to make the paper self-contained.

**Lemma 2.1** (see [4, Proposition A.24]). If \( h \) is a \( C^{1,1} \) function, then for each \( x, y \in \mathbb{R}^n \), we have

\[
|h(y) - h(x) - \langle \nabla h(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|_2^2.
\]
We state and prove the following lemma, which is the key to prove the non-increasing property of the objective function values and the convergence of the proximal gradient algorithm (2.3)-(2.4).

**Lemma 2.2.** Suppose $v < 1/L$. Then $Q(x, y) \geq F(x)$ for each $x, y \in \mathbb{R}^n$.

**Proof.** Since $v < 1/L$, for each $x, y \in \mathbb{R}^n$, we have

\[
Q(x, y) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2
\]

\[
\geq f(x) + \phi(x)
\]

\[
= F(x),
\]

where the second inequality follows from Lemma 2.1.

Now we establish some convergence properties of the proximal gradient algorithm.

**Theorem 2.1.** Suppose that the sequence $\{x_k\}$ is generated by the proximal gradient algorithm with the stepsize $v < 1/L$. Then the following statements hold:

(i) the sequences $\{Q(x_{k+1}, x_k)\}$ and $\{F(x_k)\}$ are non-increasing,

(ii) the sequence of iterates $\{x_k\}$ satisfies $\lim_{k \to \infty} \|x_{k+1} - x_k\|_2 = 0$, and

(iii) each cluster point $x^*$ of $\{x_k\}$ is the fixed point of the proximal gradient operator, i.e., $x^* \in (P_\phi \circ G_f)(x^*)$.

**Proof.** (i) For any $k \in \mathbb{N}$, by (2.5) and (2.6), we have

\[
F(x_k) = Q(x_k, x_k) \geq Q(x_{k+1}, x_k).
\]

Since $v < 1/L$, Lemma 2.2 is applicable to conclude that $Q(x_{k+1}, x_k) \geq F(x_{k+1})$. Thus, we obtain the non-increasing property of $\{Q(x_{k+1}, x_k)\}$ and $\{F(x_k)\}$.

(ii) By (2.7) and (2.5), we obtain

\[
F(x_k) \geq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \phi(x_{k+1}) + \frac{1}{2v} \|x_{k+1} - x_k\|^2.
\]

After rearranging, we arrive at

\[
F(x_k) - F(x_{k+1}) \geq f(x_k) - f(x_{k+1}) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2v} \|x_{k+1} - x_k\|^2
\]

\[
\geq \frac{1 - vL}{2v} \|x_{k+1} - x_k\|^2,
\]

where the second inequality follows from Lemma 2.1. Thus we obtain the relations

\[
\|x_{k+1} - x_k\|^2 \leq \frac{2v}{1 - vL} (F(x_k) - F(x_{k+1})),
\]

\[8\]
Lemma 3.1. The group can be given explicitly by an analytical formula, as shown in the following lemma.

When \( p \) is a group polarization problem (3.1) as follows:

\[
\ell \quad \text{proximal gradient algorithm (2.3)-(2.4)} \quad \text{to solve the}
\]

\[\ell \quad \text{group sparse optimization via the proximal gradient algorithm (2.3)-(2.4)} \quad \text{to solve the}
\]

\[\ell \quad \text{Applications to group sparse optimization}
\]

Let \( x^* \) be a cluster point of \( \{x_k\} \). Then there exists a subsequence \( \{x_{k_i}\} \) such that \( x_{k_i} \to x^* \) as \( i \to \infty \). By the statement (ii), we have that \( \{x_{k_i+1}\} \) also converges to \( x^* \). Next we show the continuity of the proximal gradient operator \( P_{\phi} \circ G_f \). By the assumption that \( f \) is a \( C^{1,1} \) function, we have \( \nabla f(\cdot) \) is continuous, and then \( G_f(\cdot) \) is also continuous. Since \( x_{k_i+1} \in P_{\phi}(G_f(x_{k_i})) \), \( G_f(x_{k_i}) \to G_f(x^*) \) and \( \phi \) is isc and bounded from below, the hypotheses in [22, Theorem 1.25] are verified and thus it is applicable to conclude that the cluster points of \( \{x_{k_i+1}\} \) lie in \( P_{\phi}(G_f(x^*)) \). Since further \( \{x_{k_i+1}\} \) converges to \( x^* \), it follows that \( x^* \in (P_{\phi} \circ G_f)(x^*) \).

\[\text{3 Applications to group sparse optimization}
\]

The group sparse optimization via the \( \ell_{p,q} \ (p \geq 1, q \geq 0) \) regularization is described as follows:

\[
\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda \|x\|_{p,q}^q.
\]  

(3.1)

In this section, we apply the proximal gradient algorithm (2.3)-(2.4) to solve the \( \ell_{p,q} \) regularization problem (3.1) as follows:

\[
z_k = G_A(x_k) := x_k - 2vA^\top(Ax_k - b),
\]

(3.2)

\[
x_{k+1} \in P_{p,q}(z_k) := \text{Arg min}_{x \in \mathbb{R}^n} \{\lambda \|x\|_{p,q}^q + \frac{1}{2v}\|x - z_k\|_2^2\}.
\]

(3.3)

Recall that the \( \ell_{p,q} \) norm and \( \ell_2 \) norm are both grouped separable functions. Then the proximal step (3.3) can be achieved parallel in each group, and equivalent to solve a cycle of low dimensional optimization problems

\[
[x_{k+1}]_i \in \text{Arg min}_{x \in \mathbb{R}^n} \{\lambda \|[x]_i\|_p^p + \frac{1}{2v}\|[x]_i - [z_k]_i\|_2^2\}, \text{ for } i = 1, \cdots, s.
\]

(3.4)

When \( p \) and \( q \) are given as some specific numbers, the solution of subproblem (3.4) of each group can be given explicitly by an analytical formula, as shown in the following lemma.

**Lemma 3.1.** Let \( z \in \mathbb{R}^l \), \( v > 0 \) and the proximal regularization \( R_{p,q}(x) := \lambda \|x\|_p^p + \frac{1}{2v}\|x-z\|_2^2 \).

Then the proximal operator

\[
P_{p,q}(z) \in \text{Arg min}_{x \in \mathbb{R}^l} \{R_{p,q}(x)\}
\]

has the following analytical formula:
(i) if \( p = 2 \) and \( q = 1 \), then

\[
P_{2,1}(z) = \begin{cases} 
  z - \frac{v\lambda}{\|z\|_2} z, & \|z\|_2 > v\lambda, \\
  0, & \text{otherwise,}
\end{cases}
\]  

(3.5)

(ii) if \( p \geq 1 \) and \( q = 0 \), then

\[
P_{p,0}(z) = \begin{cases} 
  z, & \|z\|_2 > \sqrt{2v\lambda}, \\
  0 \text{ or } z, & \|z\|_2 = \sqrt{2v\lambda}, \\
  0, & \|z\|_2 < \sqrt{2v\lambda},
\end{cases}
\]  

(3.6)

(iii) if \( p = 2 \) and \( q = 1/2 \), then

\[
P_{2,1/2}(z) = \begin{cases} 
  \frac{16\|z\|_2^{3/2}\cos^3(\frac{\pi}{4} - \frac{\psi(z)}{4})}{3\sqrt{3}v\lambda + 16\|z\|_2^{3/2}\cos^3(\frac{\pi}{4} - \frac{\psi(z)}{4})} z, & \|z\|_2 > \frac{3}{2}(v\lambda)^{2/3}, \\
  0 \text{ or } \frac{16\|z\|_2^{3/2}\cos^3(\frac{\pi}{4} - \frac{\psi(z)}{4})}{3\sqrt{3}v\lambda + 16\|z\|_2^{3/2}\cos^3(\frac{\pi}{4} - \frac{\psi(z)}{4})} z, & \|z\|_2 = \frac{3}{2}(v\lambda)^{2/3}, \\
  0, & \|z\|_2 < \frac{3}{2}(v\lambda)^{2/3},
\end{cases}
\]  

(3.7)

with

\[
\psi(z) = \arccos \left( \frac{3}{2} \left( \frac{\|z\|_2}{3^{1/3}\|z\|_1 \cos(\frac{\pi}{4} - \frac{\xi(z)}{4})} \right)^{3/2} \right).
\]  

(3.8)

(iv) if \( p = 1 \) and \( q = 1/2 \), then

\[
P_{1,1/2}(z) = \begin{cases} 
  \bar{z}, & R_{1,1/2}(\bar{z}) < R_{1,1/2}(0), \\
  0 \text{ or } \bar{z}, & R_{1,1/2}(\bar{z}) = R_{1,1/2}(0), \\
  0, & R_{1,1/2}(\bar{z}) > R_{1,1/2}(0),
\end{cases}
\]  

(3.9)

where

\[
\bar{z} = z - \frac{\sqrt{3}v\lambda \text{sgn}(z)}{4\sqrt{\|z\|_1 \cos(\frac{\pi}{4} - \frac{\xi(z)}{4})}},
\]

and

\[
\xi(z) = \arccos \left( \frac{3\lambda}{4} \left( \frac{3}{\|z\|_1^{3/2}} \right) \right).
\]  

(3.10)

Proof. Since the proximal regularization \( R_{p,q}(x) \) is non-differentiable only at 0, \( P_{p,q}(z) \) must be 0 or a point \( \bar{x}(\neq 0) \) satisfying the first order condition

\[
\lambda q \|\bar{x}\|_p^{q-1} [||\bar{x}_i||_1^{p-1} \text{sgn}(\bar{x}_i)] + \frac{1}{q} (\bar{x} - z) = 0,
\]  

(3.11)

where \([a_i]\) denotes a vector whose \( i \)-th component is \( a_i \), and \( \text{sgn}(\cdot) \) denotes the sign function. Thus, we just need to calculate such \( \bar{x} \), and then compare the objective function values \( R_{p,q}(\bar{x}) \) and \( R_{p,q}(0) \) to obtain the solution \( P_{p,q}(z) \).
(i) When $p = 2$ and $q = 1$, (3.11) reduces to
\[ \lambda \frac{\bar{x}}{\|\bar{x}\|_2} + \frac{1}{v}(\bar{x} - z) = 0, \]
which implies that $\|\bar{x}\|_2 = \|z\|_2 - v\lambda$ and $\bar{x} = \frac{\|z\|_2 - v\lambda}{\|z\|_2}z$. Comparing the objective function values $R_{2,1}(\frac{\|z\|_2 - v\lambda}{\|z\|_2}z) = \lambda\|z\|_2 - v\lambda + \frac{1}{2}v\lambda^2$ and $R_{2,1}(0) = \frac{1}{2v}\|z\|_2^2$, we obtain that
\[ \|z\|_2 > v\lambda \Rightarrow R_{2,1}(0) - R_{2,1}(\frac{\|z\|_2 - v\lambda}{\|z\|_2}z) = \frac{1}{2v}(\|z\|_2 - v\lambda)^2 > 0 \Rightarrow P_{2,1}(z) = 0, \]
and that
\[ \|z\|_2 \leq v\lambda \Rightarrow R_{2,1}(0) - R_{2,1}(\frac{\|z\|_2 - v\lambda}{\|z\|_2}z) = \frac{1}{2v}(\|z\|_2 - v\lambda)(\|z\|_2 + 3v\lambda) \leq 0 \Rightarrow P_{2,1}(z) = 0. \]
Therefore, we arrive at the formula (3.5).

(ii) When $p \geq 1$ and $q = 0$, (3.11) implies $\bar{x} = z$. Comparing the objective function values $R_{2,0}(z) = \lambda$ and $R_{2,0}(0) = \frac{1}{2v}\|z\|_2^2$, we obtain the formula (3.6).

(iii) When $p = 2$ and $q = 1/2$, (3.11) reduces to
\[ \frac{\lambda \bar{x}}{2\|\bar{x}\|_2^{3/2}} + \frac{1}{v}(\bar{x} - z) = 0, \]
which implies
\[ \|\bar{x}\|_2^{3/2} - \|\bar{x}\|_2^{1/2}\|z\|_2 + \frac{v\lambda}{2} = 0. \]
Denote $\eta = \|\bar{x}\|_2^{1/2} \geq 0$. The equation (3.13) can be transformed into the following cubic algebraic equation
\[ \eta^3 - \|z\|_2\eta + \frac{v\lambda}{2} = 0. \]
Due to the hyperbolic solution of the cubic equation (see [23]), by denoting
\[ r = 2\sqrt{\frac{\|z\|_2}{3}}, \quad \alpha = \arccos\left(\frac{v\lambda}{4}(\frac{3}{\|z\|_2})^{3/2}\right) = \psi(z) \quad \text{and} \quad \beta = \arccosh\left(-\frac{v\lambda}{4}(\frac{3}{\|z\|_2})^{3/2}\right), \]
the solution of (3.14) can be expressed as the follows.

(a) If $0 \leq \|z\|_2 \leq 3\left(\frac{v\lambda}{4}\right)^{2/3}$, then the three roots of (3.14) are given by
\[ \eta_1 = r \cosh \frac{\beta}{3}, \quad \eta_2 = -\frac{r}{2} \cosh \frac{\beta}{3} + i\frac{\sqrt{3}r}{2} \sinh \frac{\beta}{3}, \quad \eta_3 = -\frac{r}{2} \cosh \frac{\beta}{3} - i\frac{\sqrt{3}r}{2} \sinh \frac{\beta}{3}. \]
However, this $\beta$ does not exist since the value of hyperbolic cosine must be positive. Thus, in this case, $P_{2,1/2}(z) = 0$. 11
(b) If \( \|z\|_2 > 3(\frac{v\lambda}{4})^{2/3} \), then the three roots of (3.14) are
\[
\eta_1 = r \cos(\frac{\pi}{3} - \frac{\alpha}{3}), \quad \eta_2 = -r \sin(\frac{\pi}{2} - \frac{\alpha}{3}), \quad \eta_3 = -r \cos(\frac{2\pi}{3} - \frac{\alpha}{3}).
\]
Only \( \eta_1 \) is the positive solution and thus the unique solution of (3.12) is given by
\[
\bar{x} = \frac{2\eta_1^2}{v\lambda} + 2\eta_1^3 z = \frac{16\|z\|_2^{3/2} \cos^3(\frac{\pi}{3} - \frac{\psi(z)}{3})}{3\sqrt{3} v\lambda + 16\|z\|_2^{3/2} \cos^3(\frac{\pi}{3} - \frac{\psi(z)}{3})} z.
\]
Finally, we compare the objective function values \( R_{2,1/2}(\bar{x}) \) and \( R_{2,1/2}(0) \). For this purpose, when \( \|z\|_2 > 3(\frac{v\lambda}{4})^{2/3} \), we define
\[
G(\|z\|_2) := \frac{\|z\|_2}{\|z\|_2} \left( R_{2,1/2}(0) - R_{2,1/2}(\bar{x}) \right)
= \frac{v}{\|z\|_2} \left( \frac{1}{2\|z\|_2^2 - \lambda \|\bar{x}\|_2^{1/2} - \frac{1}{2\eta_1} \|\bar{x} - z\|_2^{1/2}} \right)
= \|z\|_2 - \frac{\|\bar{x}\|_2^{1/2} + 2v\lambda \|z\|_2^{1/2}}{2\|z\|_2}
= \|z\|_2 - \frac{1}{2\|z\|_2} - \frac{3}{4} v\lambda \|\bar{x}\|_2^{-1/2},
\]
where the third equality holds since \( \bar{x} \) is proportional to \( z \), and fourth equality follows from (3.13). Since both \( \|z\|_2 \) and \( \|\bar{x}\|_2 \) are strictly increasing on \( \|z\|_2 \), \( G(\|z\|_2) \) is also strictly increasing when \( \|z\|_2 > 3(\frac{v\lambda}{4})^{2/3} \). Thus the unique solution of \( G(\|z\|_2) = 0 \) satisfies
\[
\|z\|_2 \|\bar{x}\|_2^{1/2} = \frac{3}{2} v\lambda,
\]
and further, (3.13) implies that the solution of \( G(\|z\|_2) = 0 \) is
\[
\|z\|_2 = \frac{3}{2}(v\lambda)^{2/3}.
\]
Therefore, we arrive at the formulae (3.7) and (3.8).

(iv) When \( p = 1 \) and \( q = 1/2 \), (3.11) reduces to
\[
\frac{\lambda sgn(\bar{x})}{2 \sqrt{\|\bar{x}\|_1}} + \frac{1}{v}(\bar{x} - z) = 0,
\]
which implies
\[
0 = l + \frac{2}{v\lambda} \sqrt{\|\bar{x}\|_1} (\|\bar{x}\|_1 - \|z\|_1), \quad (3.15)
\]
by \( sgn(\bar{x}) = sgn(z) \). Denote \( \eta = \|\bar{x}\|_1^{1/2} \geq 0 \). Then (3.15) reduces to the following cubic algebraic equation
\[
\eta^3 - \|z\|_1 \eta + \frac{v\lambda l}{2} = 0. \quad (3.16)
\]
Similar as the proof of statement (iii), the unique solution of (3.15) is given by
\[ \bar{x} = z - \frac{v\lambda \text{sgn}(z)}{2\eta_1} = z - \frac{\sqrt{3}v\lambda \text{sgn}(z)}{4\sqrt{\|z\|_1} \cos\left(\frac{\xi(z)}{3} - \frac{\psi}{3}\right)} (= \bar{z}), \]
where \( \eta_1 = 2\sqrt{\frac{3}{4}} \sqrt{\|z\|_1} \cos\left(\frac{\xi(z)}{3} - \frac{\psi}{3}\right) \) with \( \xi(z) = \arccos\left(\frac{v\lambda l}{2}\right) \). Therefore, comparing \( R_{1,1/2}(\bar{x}) \) and \( R_{1,1/2}(0) \), we obtain that \( P_{1,1/2}(z) \) can be formulated as (3.9) and (3.10).

The convergence theorem of the proximal gradient algorithm (3.2)-(3.3) is stated as follows.

**Theorem 3.1.** Let \( p \geq 1 \). Suppose that the sequence \( \{x_k\} \) is generated by the proximal gradient algorithm (3.2)-(3.3) with \( v < \frac{1}{2}\|A\|^{-2} \). Then the following statements hold:

(i) if \( q \geq 1 \), then \( \{x_k\} \) converges to the global minimizer of problem (3.1),

(ii) if \( q = 0 \), then \( \{x_k\} \) converges to the local minimizer of problem (3.1), and

(iii) if \( 0 < q < 1 \), each cluster point of \( \{x_k\} \) is the fixed point of the proximal gradient operator.

**Proof.** (i) If \( q \geq 1 \), the \( \ell_{p,q} \) norm is convex and the convergence of the proximal gradient algorithm has been established in [2, Theorem 3.1] and [3, Theorem 1.2].

(ii) By Theorem 2.1 (i) and the fact that the \( \ell_{p,0} \) is level bounded, we have that \( \{x_k\} \) is bounded and then \( \{x_k\} \) has some cluster point. Thus, by Theorem 2.1 (iii), we assume that a subsequence \( \{x_{k_i}\} \) converges to the fixed point \( x^* \). If \( \|x_{k_i}\|_2 \geq \sqrt{2\nu\lambda} \) and \( [x_{k+1}]_{i} = 0 \), then \( \|x_{k+1} - x_k\|_2 \geq \sqrt{2\nu\lambda} \), which by Theorem 2.1 (ii) is impossible for large \( k \). Thus, by Lemma 3.1 (ii), for large \( k \), the index set of zero and non-zero groups will not change, and then \( x_{k+1} = (P_{p,0} \circ G_A)(x_k) \) as well as \( x^* = (P_{p,0} \circ G_A)(x^*) \). Then, for large \( k \), we obtain
\[
\|x_{k+1} - x^*\|_2 = \|(P_{p,0} \circ G_A)(x_k) - (P_{p,0} \circ G_A)(x^*)\|_2 \\
\leq \|G_A(x_k) - G_A(x^*)\|_2 \\
\leq \|x_k - x^*\|_2,
\]
where the first inequality follows from Lemma 3.1 (ii) and the fact the index set of zero and non-zero groups does not change, and the second inequality follows from the nonexpansive property of \( G_A \) since \( v < \frac{1}{2}\|A\|^{-2} \). Hence, the whole sequence \( \{x_k\} \) converges to \( x^* \).

Next, we prove that each fixed point of \( P_{p,0} \circ G_A \) is the local minimizer of the \( \ell_{p,0} \) regularization problem. Denote
\[ F_{p,0}(x) := \|Ax - b\|_2^2 + \lambda\|x\|_{p,0}. \]
For the fixed point \( x^* \), we need to show that there exists some \( \delta > 0 \) such that \( F_{p,0}(x^* + \varepsilon) \leq F_{p,0}(x^*) \) for any small perturbation \( \varepsilon \) with \( \|\varepsilon\|_2 \leq \delta \). Denote the index sets \( \Gamma = \{i : [x^*]_i = 0\} \), \( \Lambda = \{i : [x^*]_i \neq 0\} \) and \( \delta = \sqrt{2\lambda/2} \). Then, for any small perturbation \( \varepsilon \) with \( \|\varepsilon\|_2 \leq \delta \), we obtain that \( [x^* + \varepsilon]_i \neq 0 \) for all \( i \in \Lambda \) and that

\[
F_{p,0}(x^* + \varepsilon) - F_{p,0}(x^*) \\
\geq 2\langle A^\top (Ax^* - b), \varepsilon \rangle + \lambda \|x^* + \varepsilon\|_{p,0} - \lambda \|x^*\|_{n,0} \\
= \sum_{i \in \Gamma} \left( \lambda \|\varepsilon_i\|_2^0 + 2\langle [A^\top]_i (Ax^* - b), [\varepsilon]_i \rangle \right) + \sum_{i \in \Lambda} 2\langle [A^\top]_i (Ax^* - b), [\varepsilon]_i \rangle.
\]

For \( i \in \Gamma \), it follows from (3.6) and (3.2) that

\[
\|2v[A^\top]_i(Ax^* - b)\|_2 \leq \sqrt{2v\lambda},
\]

and that

\[
\lambda \|\varepsilon_i\|_2^0 + 2\langle [A^\top]_i (Ax^* - b), [\varepsilon]_i \rangle \geq 0.
\]

On the other hand, for \( i \in \Lambda \), it follows from (3.2) and (3.6)

\[
[A^\top]_i(Ax^* - b) = 0.
\]

Therefore, combining the preceding inequalities, for any small perturbation \( \varepsilon \) with \( \|\varepsilon\|_2 \leq \delta \), we arrive at

\[
F_{p,0}(x^* + \varepsilon) - F_{p,0}(x^*) \geq 0.
\]

Consequently, the fixed point \( x^* \) is a local minimizer of the \( \ell_{p,0} \) regularization problem, and hence, \( \{x_k\} \) converges to the local minimizer of problem (3.1).

(iii) It is a direct conclusion of Theorem 3.1.

\[\blacksquare\]

**Remark 3.1.** Note from (3.6), (3.7) and (3.9) that the solution of the proximal regularization problem (3.4) might not be unique when \( R_{p,q}(\bar{x}) = R_{p,q}(0) \). To avoid this obstacle in numerical computations, we choose the solution \( P_{p,q}(\bar{x}) = 0 \) whenever \( R_{p,q}(\bar{x}) = R_{p,q}(0) \), which achieves a more sparse solution, in the definition of the proximal operator to guarantee a unique update.

**Remark 3.2.** When \( n_i = 1 \) for all groups, the data do not form any group structure in the feature space, and the sparsity is achieved only on the individual feature level. In this case, the proximal operators \( P_{2,1}(z) \) in (3.5), \( P_{2,0}(z) \) in (3.6) and \( P_{2,1/2}(z) \) in (3.7) and \( P_{1,1/2}(z) \) in (3.9) reduce to the soft thresholding function in [11], the hard thresholding function in [5] and the half thresholding function in [28], respectively.

Hence, by (3.2)-(3.3), Lemma 3.1 and Remark 2.1, we outline the proximal gradient algorithm for the group sparse optimization (in short, PGA-GSO) via \( \ell_{p,q} \) regularization model (3.1) as follows.
PGA-GSO
Select a stepsizes $v$, start an initial point $x_0 \in \mathbb{R}^n$, and generate a sequence $\{x_k\} \subseteq \mathbb{R}^n$ via the iteration, consisting of a gradient step

$$z_k = x_k - 2vA^\top(Ax_k - b),$$

and a cycle of $s$ low dimensional proximal steps

$$[x_{k+1}]_i = P_{p,q}([z_k]_i), \quad i = 1, 2, \ldots, s,$$

where $P_{p,q}(\cdot)$ is a proximal operator.

4 Numerical Experiments

The purpose of this section is to show the performance of the proposed proximal gradient algorithm for the sparse optimization problem. We apply and compare PGA-GSO with four types of $\ell_{p,q}$ regularization, $(p, q) = (2, 1), (2, 0), (2, 1/2)$ and $(1, 1/2)$.

In the numerical experiments, the numerical data are generated as follows. We first randomly generate an i.i.d. Gaussian ensemble $A \in \mathbb{R}^{m \times n}$ satisfying $A^\top A = I$. Then we generate a group sparse solution $\bar{x} \in \mathbb{R}^n$ via randomly splitting its components into $s$ groups and randomly picking $k$ of them as active groups, whose entries are also randomly generated as i.i.d. Gaussian, while the remaining groups are all set as zeros. We generate the data $b$ by the MATLAB script

$$b = A \ast \bar{x} + \text{sigma} \ast \text{randn}(m, 1),$$

where $\text{sigma}$ is the standard deviation of additive Gaussian noise. The problem size is set to $n = 1024$ and $m = 256$, and we test on the noisy measurement data with $\text{sigma} = 0.5\%$.

For each given sparsity level, which is $k/s$, we randomly generate the data $A$, $\bar{x}$, $b$ (as above), run the algorithm 500 times, and average the 500 numerical results to illustrate the performance of the algorithm. We choose the stepsizes $v = 1/2$ in all the testing. The two key criteria to characterize the performance are the relative error $\|x - \bar{x}\|_2/\|\bar{x}\|_2$ and the successful recovery rate, where the recovery is defined as success when the relative error between the recovered data and the true data is smaller than $10^{-2}$, otherwise, it regarded as failure.

We carry out two experiments with the initial point $x_0 = 0$. In the first experiment, setting $s = 128$ (so group size $G = 1024/128 = 8$), we compare the convergence rate results and the successful recovery rates of the algorithm for different sparsity levels. In Figure 1, (a), (b), (c) respectively show the convergence rate results on sparsity level 1%, 5%, 10%, while (d) plots the successful recovery rates on different sparsity levels. When the solution is of high sparse level, as shown in Figure 1 (a), $\ell_{2,0}$, $\ell_{2,1/2}$ and $\ell_{1,1/2}$ perform perfect and achieve a fast convergence rate. As demonstrated in Figure 1 (b) and (c), when the sparsity level drops to 5% and 10%, $\ell_{2,1/2}$ and $\ell_{1,1/2}$ perform better and arrive at a more accurate level than $\ell_{2,1}$ and $\ell_{2,0}$, and it surprises us that $\ell_{1,1/2}$ performs a little better than $\ell_{2,1/2}$. From Figure 1 (d), it is illustrated that $\ell_{2,1/2}$ and $\ell_{1,1/2}$ achieve a better successful recovery.
rate than $\ell_{2,0}$ and $\ell_{2,1}$. Moreover, we surprisingly see that $\ell_{1,1/2}$ also outperforms $\ell_{2,1/2}$ on the successful recovery rate. In this experiment, we also note that the running times are at the same level for the algorithm, about 0.9 second per 500 iteration.

Figure 1: Convergence results and recovery rates for different sparsity levels.

The second experiment is performed to show the sensitivity analysis on the group size ($G = 4, 8, 16, 32$) of PGA-GSO with four types of $\ell_{p,q}$ regularization. As shown in Figure 2, the four types of $\ell_{p,q}$ reach a higher successful recovery rate for the larger group size. We also note that the larger the group size, the shorter the running time.

References


Figure 2: Sensitivity analysis on group size


