Journal of Machine Learning Research 18 (2017) 1-52 Submitted 12/15; Revised 12/16; Published 3/17

Group Sparse Optimization via $\ell_{p,q}$ Regularization

Yaohua Hu MAYHHU@SZU.EDU.CN
College of Mathematics and Statistics
Shenzhen University
Shenzhen 518060, P. R. China

Chong Li CLI@ZJU.EDU.CN
School of Mathematical Sciences
Zhejiang University
Hangzhou 310027, P. R. China

Kaiwen Meng MKWFLY@126.COM
School of Economics and Management
Southwest Jiaotong University
Chengdu 610031, P. R. China

Jing Qin QINJING@CUHK.EDU.HK
School of Life Sciences
The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong
and Shenzhen Research Institute
The Chinese University of Hong Kong
Shenzhen 518057, P. R. China

Xiaoqi Yang* MAYANGXQ@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Kowloon, Hong Kong

Editor: Mark Schmidt

Abstract

In this paper, we investigate a group sparse optimization problem via $\ell_{p,q}$ regularization in three aspects: theory, algorithm and application. In the theoretical aspect, by introducing a notion of group restricted eigenvalue condition, we establish an oracle property and a global recovery bound of order $\mathcal{O}(\lambda^{\frac{2}{2-q}})$ for any point in a level set of the $\ell_{p,q}$ regularization problem, and by virtue of modern variational analysis techniques, we also provide a local analysis of recovery bound of order $\mathcal{O}(\lambda^2)$ for a path of local minima. In the algorithmic aspect, we apply the well-known proximal gradient method to solve the $\ell_{p,q}$ regularization problems, either by analytically solving some specific $\ell_{p,q}$ regularization subproblems, or by using the Newton method to solve general $\ell_{p,q}$ regularization subproblems. In particular, we establish a local linear convergence rate of the proximal gradient method for solving the $\ell_{1,q}$ regularization problem under some mild conditions and by first proving a second-order growth condition. As a consequence, the local linear convergence rate of proximal gradient method for solving the usual ℓ_q regularization problem $(0 < q < 1)$ is obtained. Finally in

$*$. Corresponding author.

the aspect of application, we present some numerical results on both the simulated data and the real data in gene transcriptional regulation.

Keywords: group sparse optimization, lower-order regularization, nonconvex optimization, restricted eigenvalue condition, proximal gradient method, iterative thresholding algorithm, gene regulation network

1. Introduction

In recent years, a great amount of attention has been paid to sparse optimization, which is to find the sparse solutions of an underdetermined linear system. The sparse optimization problem arises in a wide range of fields, such as compressive sensing, machine learning, pattern analysis and graphical modeling; see Blumensath and Davies (2008); Candès et al. (2006b); Chen et al. (2001); Donoho (2006a); Fan and Li (2001); Tibshirani (1994) and references therein.

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

$$Ax = b + \varepsilon,$$

where $A \in \mathbb{R}^{m\times n}$ 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 system is seriously ill-conditioned and may have infinitely many 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 measured by the ℓ_p norm $\|x\|_p$ ($p = 0$, see Blumensath and Davies (2008); $p = 1$, see Beck and Teboulle (2009); Chen et al. (2001); Daubechies et al. (2004); Donoho (2006a); Tibshirani (1994); Wright et al. (2009); Yang and Zhang (2011); and $p = 1/2$, see Chartrand and Staneva (2008); Xu et al. (2012)). The ℓ_p norm $\|x\|_p$ for $p > 0$ is defined by

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

while the ℓ_0 norm $\|x\|_0$ is defined by the number of nonzero components of x. The sparse optimization problem can be modeled as

$$\begin{array}{ll} \min & \|Ax - b\|_2 \\ \text{s.t.} & \|x\|_0 \le s, \end{array}$$

where s 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 ℓ_0 regularization problem is

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

where $\lambda > 0$ is the regularization parameter, providing a tradeoff between accuracy and sparsity. However, the ℓ_0 regularization problem is nonconvex and non-Lipschitz, and thus it is generally intractable to solve it directly (indeed, it is NP-hard; see Natarajan, 1995).

To overcome this difficulty, two typical relaxations of the ℓ_0 regularization problem are introduced, which are the ℓ_1 regularization problem

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

and the ℓ_q regularization problem $(0<q<1)$

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 + \lambda\|x\|_q^q. \tag{2}$$

1.1 ℓ_q Regularization Problems

The ℓ_1 regularization problem (1), also called Lasso (Tibshirani, 1994) or Basis Pursuit (Chen et al., 2001), has attracted much attention and has been accepted as one of the most useful tools for sparse optimization. Since the ℓ_1 regularization problem is a convex optimization problem, many exclusive and efficient algorithms have been proposed and developed for solving problem (1); see Beck and Teboulle (2009); Combettes and Wajs (2005); Daubechies et al. (2004); Hu et al. (2016); Nesterov (2012, 2013); Xiao and Zhang (2013); Yang and Zhang (2011). However, the ℓ_1 regularization problem (1) suffers some frustrations in practical applications. It was revealed by extensive empirical studies that the solutions obtained from the ℓ_1 regularization problem are much less sparse than the true sparse solution, that it cannot recover a signal or image with the least measurements when applied to compressed sensing, and that it often leads to sub-optimal sparsity in reality; see Chartrand (2007); Xu et al. (2012); Zhang (2010).

Recently, to overcome these drawbacks of ℓ_1 regularization, the lower-order regularization technique (that is, the ℓ_q regularization with $0<q<1$) is proposed to improve the performance of sparsity recovery of the ℓ_1 regularization problem. Chartrand and Staneva (2008) claimed that a weaker restricted isometry property is sufficient to guarantee perfect recovery in the ℓ_q regularization, and that it can recover sparse signals from fewer linear measurements than that required by the ℓ_1 regularization. Xu et al. (2012) showed that the $\ell_{1/2}$ regularization admits a significantly stronger sparsity promoting capability than the ℓ_1 regularization in the sense that it allows to obtain a more sparse solution or predict a sparse signal from a smaller amount of samplings. Qin et al. (2014) exhibited that the $\ell_{1/2}$ regularization achieves a more reliable solution in biological sense than the ℓ_1 regularization when applied to infer gene regulatory network from gene expression data of mouse embryonic stem cell. However, the ℓ_q regularization problem is nonconvex, nonsmooth and non-Lipschitz, and thus it is difficult in general to design efficient algorithms for solving it. It was presented in Ge et al. (2011) that finding the global minimal value of the ℓ_q regularization problem (2) is strongly NP-hard; while fortunately, computing a local minimum could be done in polynomial time. Some effective and efficient algorithms have been proposed to find a local minimum of problem (2), such as interior-point potential reduction algorithm (Ge et al., 2011), smoothing methods (Chen, 2012; Chen et al., 2010), splitting methods (Li and Pong, 2015a,b) and iterative reweighted minimization methods (Lai and Wang, 2011; Lai et al., 2013; Lu, 2014).

The ℓ_q regularization problem (2) is a variant of lower-order penalty problems, investigated in Huang and Yang (2003); Luo et al. (1996); Yang and Huang (2001), for a constrained optimization problem. The main advantage of the lower-order penalty functions

over the classical ℓ_1 penalty function is that they require weaker conditions to guarantee an exact penalization property and that their least exact penalty parameter is smaller; see Huang and Yang (2003). It was reported in Yang and Huang (2001) that the first- and second-order necessary optimality conditions of lower-order penalty problems converge to that of the original constrained optimization problem under a linearly independent constraint qualification.

Besides the preceding numerical algorithms, one of the most widely studied methods for solving the sparse optimization problem is the class of the iterative thresholding algorithms, which is studied in a unified framework of proximal gradient methods; see Beck and Teboulle (2009); Blumensath and Davies (2008); Combettes and Wajs (2005); Daubechies et al. (2004); Gong et al. (2013); Nesterov (2013); Xu et al. (2012) and references therein. It is convergent fast and of very low computational complexity. Benefitting from its simple formulation and low storage requirement, it is very efficient and applicable for large-scale sparse optimization problems. In particular, the iterative hard (resp. soft, half) thresholding algorithm for the ℓ_0 (resp. ℓ_1, $\ell_{1/2}$) regularization problem was studied in Blumensath and Davies (2008) (resp. Daubechies et al., 2004; Xu et al., 2012).

1.2 Global Recovery Bound

To estimate how far is the solution of regularization problems from that of the linear system, the global recovery bound (also called the ℓ_2 consistency) of the ℓ_1 regularization problem has been investigated in the literature. More specifically, under some mild conditions on A, such as the restricted isometry property (RIP, Candès and Tao, 2005) or restricted eigenvalue condition (REC, Bickel et al., 2009), van de Geer and Bühlmann (2009) established a deterministic recovery bound for the (convex) ℓ_1 regularization problem:

$$\|x^*(\ell_1) - \bar{x}\|_2^2 = \mathcal{O}(\lambda^2 s), \tag{3}$$

where $x^*(\ell_1)$ is a solution of problem (1), $\bar{x}$ is a solution of the linear system $Ax = b$, and sparsity $s := \|\bar{x}\|_0$. In the statistics literature, Bickel et al. (2009); Bunea et al. (2007); Meinshausen and Yu (2009); Zhang (2009) provided the recovery bound in a high probability for the ℓ_1 regularization problem when the size of the variable tends to infinity, under REC/RIP or some relevant conditions. However, to the best of our knowledge, the recovery bound for the general (nonconvex) ℓ_p regularization problem is still undiscovered. We will establish such a deterministic property in section 2.

1.3 Group Sparse Optimization

In applications, a wide class of problems usually have certain special structures, and recently, enhancing the recoverability due to the special structures has become an active topic in sparse optimization. One of the most popular structures is the group sparsity structure, 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. In general, the grouping information can be an arbitrary partition of x, and it is usually pre-defined based on prior knowledge of specific problems. Let $x := (x_{\mathcal{G}_1}^\top, \cdots, x_{\mathcal{G}_r}^\top)^\top$ represent the group structure of x. The group

sparsity of x with such a group structure can be measured by an $\ell_{p,q}$ norm, defined by

$$\|x\|_{p,q} := \left(\sum_{i=1}^{r} \|x_{\mathcal{G}_i}\|_p^q \right)^{1/q}.$$

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 applied in birthweight prediction (Bach, 2008; Yuan and Lin, 2006), dynamic MRI (Usman et al., 2011) and gene finding (Meier et al., 2008; Yang et al., 2010) with the $\ell_{2,1}$ norm. More specifically, the $\ell_{2,1}$ regularization problem

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

was introduced by Yuan and Lin (2006) to study the grouped variable selection in statistics under the name of group Lasso. The $\ell_{2,1}$ regularization, an important extension of the ℓ_1 regularization, proposes an ℓ_2 regularization for each group and ultimately yields the sparsity in a group manner. Since the $\ell_{2,1}$ regularization problem is a convex optimization problem, some effective algorithms have been proposed, such as, the spectral projected gradient method (van den Berg et al., 2008), SpaRSA (Wright et al., 2009) and the alternating direction method (Deng et al., 2011).

1.4 The Aim of This Paper

In this paper, we will investigate the group sparse optimization via $\ell_{p,q}$ regularization ($p \geq 1, 0 \leq q \leq 1$), also called the $\ell_{p,q}$ regularization problem

$$\min_{x\in\mathbb{R}^n} F(x) := \|Ax - b\|_2^2 + \lambda\|x\|_{p,q}^q. \tag{4}$$

We will investigate the oracle property and recovery bound for the $\ell_{p,q}$ regularization problem, which extends the existing results in two ways: one is the lower-order regularization, including the ℓ_q regularization problem ($q < 1$); the other is the group sparse optimization, including the $\ell_{2,1}$ regularization problem (group Lasso) as a special case. To this end, we will introduce the weaker notions of REC: the lower-order REC and the group REC (GREC). We will further establish the relationships between the new notions with the classical one: the lower-order REC is weaker than the classical REC, but the reverse is not true (see Example 1); and the GREC is weaker than the REC. Under the lower-order GREC, we will provide the oracle property and the global recovery bound for the $\ell_{p,q}$ regularization problem (see Theorem 9). Furthermore, we will conduct a local analysis of recovery bound for the $\ell_{p,q}$ regularization problem by virtue of modern variational analysis techniques (Rockafellar and Wets, 1998). More precisely, we assume that any nonzero group of $\bar{x}$ is active and the columns of A corresponding to the active components of $\bar{x}$ (a solution of $Ax = b$) are linearly independent, which matches the nature of the group sparsity structure. This leads us to the application of implicit function theorem and thus guarantees the existence of a local path around $\bar{x}$ which satisfies a second-order growth condition. As such, in the local recovery bound, we will establish a uniform recovery bound $\mathcal{O}(\lambda^2 S)$ for all the $\ell_{p,q}$ regularization problems; see Theorem 2.2.

The proximal gradient method is one of the most popular and practical methods for the sparse optimization problems, either convex or nonconvex problems. We will apply the proximal gradient method to solve the $\ell_{p,q}$ regularization problem (4). The advantage of the proximal gradient method is that the proximal optimization subproblems of some specific regularization have the analytical solutions, and the resulting algorithm is thus practically attractive. In the general cases when the analytical solutions of the proximal optimization subproblems seem not available, we will employ the Newton method to solve them. Furthermore, we will investigate a local linear convergence rate of the proximal gradient method for solving the $\ell_{p,q}$ regularization problem when $p = 1$ and $0 < q < 1$ under the assumption that any nonzero group of a local minimum is active. Problem (4) of the case $p = 1$ and $0 < q < 1$ possesses the properties that the regularization term $\|\cdot\|_{p,q}^q$ is concave near a local minimum and that the objective function $F(\cdot)$ of (4) satisfies a second-order growth condition, which plays an important role in the establishment of the local linear convergence rate. To the best of our knowledge, this is the first attempt to study the local linear convergence rate of proximal gradient method for solving the lower-order optimization problems. As a consequence of this result, we will obtain the local linear convergence rate of proximal gradient method for solving ℓ_q regularization problem $(0 < q < 1)$, which includes the iterative half thresholding algorithm $(q = 1/2)$ proposed in Xu et al. (2012) as a special case. The result on local linear convergence rate of proximal gradient method for solving the ℓ_q regularization problem is still new, as far as we know.

In the aspect of application, we will conduct some numerical experiments on both simulated data and real data in gene transcriptional regulation to demonstrate the performance of the proposed proximal gradient method. From the numerical results, it is demonstrated that the $\ell_{p,1/2}$ regularization is the best one among the $\ell_{p,q}$ regularizations for $q \in [0, 1]$, and it outperforms the $\ell_{p,1}$ and $\ell_{p,0}$ regularizations on both accuracy and robustness. This observation is consistent with several previous numerical studies on the ℓ_p regularization problem; see Chartrand and Staneva (2008); Xu et al. (2012). It is also illustrated from the numerical results that the proximal gradient method ($\ell_{2,1/2}$) outperforms most solvers in group sparse learning, such as OMP (Cai and Wang, 2011), FoBa (Zhang, 2011), ℓ_1-Magic (Candès et al., 2006a), ISTA (Daubechies et al., 2004), YALL1 (Yang and Zhang, 2011) etc. The R package of the proximal gradient method for solving group sparse optimization, named GSparO in CRAN, is available at https://CRAN.R-project.org/package=GSparO

1.5 Main Contributions

This paper is to investigate the group sparse optimization under a unified framework of the $\ell_{p,q}$ regularization problem (4). In this paper, we establish the oracle property and recovery bound, design an efficient numerical method for problem (4), and apply the proposed method to solve the problem of gene transcriptional regulation. The main contributions are presented as follows.

(i) We establish the following global recovery bound for the $\ell_{p,q}$ regularization problem (4) under the (p, q)-GREC:

$$\|x^* - \bar{x}\|_2^2 \leq \begin{cases} \mathcal{O}(\lambda^{\frac{2}{2-q}} S), & 2^{K-1} q = 1, \\ \mathcal{O}(\lambda^{\frac{2}{2-q}} S^{\frac{3-q}{2-q}}), & 2^{K-1} q > 1, \end{cases} \tag{5}$$

where $\bar{x}$ is a solution of $Ax = b$, $S := \|\bar{x}\|_{p,0}$ is the group sparsity, $0 < q \leq 1 \leq p \leq 2$, x^* is any point in the level set $\mathrm{lev}_F(\bar{x})$ of problem (4), and K is the smallest integer such that $2^{K-1}q \geq 1$.

(ii) By virtue of the variational analysis technique, for all the $\ell_{p,q}$ regularization problems, we establish a uniform local recovery bound

$$\|x^*_{p,q}(\lambda) - \bar{x}\|_2^2 \leq \mathcal{O}(\lambda^2 S) \quad \text{for small } \lambda,$$

where $0 < q < 1 \leq p$ and $x^*_{p,q}(\lambda)$ is a local optimal solution of problem (4) (near $\bar{x}$).

(iii) We present the analytical formulae for the proximal optimization subproblems of specific $\ell_{p,q}$ regularizations when $p = 1, 2$ and $q = 0, 1/2, 2/3, 1$. Moreover, we prove that any sequence $\{x^k\}$, generated by proximal gradient method for solving the $\ell_{1,q}$ regularization problem, linearly converges to a local minimum x^* under some mild conditions, that is, there exist $N \in \mathbb{N}$, $C > 0$ and $\eta \in (0, 1)$ such that

$$F(x^k) - F(x^*) \leq C\eta^k \quad \text{and} \quad \|x^k - x^*\|_2 \leq C\eta^k \quad \text{for any } k \geq N.$$

(iv) Our numerical experiments show that, measured by the biological golden standards, the accuracy of the gene regulation networks forecasting can be improved by exploiting the group structure of TF complexes. The successful application of group sparse optimization to gene transcriptional regulation will facilitate biologists to study the gene regulation of higher model organisms in a genome-wide scale.

1.6 The Organization of This Paper

This paper is organized as follows. In section 2, we introduce the notions of q-REC and GREC, and establish the oracle property and (global and local) recovery bounds for the $\ell_{p,q}$ regularization problem. In section 3, we apply the proximal gradient method to solve the group sparse optimization using different types of $\ell_{p,q}$ regularization, and investigate the local linear convergence rate of the resulting proximal gradient method. Finally, section 4 exhibits the numerical results on both simulated data and real data in gene transcriptional regulation.

2. Global and Local Recovery Bounds

This section is devoted to the study of the oracle property and (global and local) recovery bounds for the $\ell_{p,q}$ regularization problem (4). To this end, we first present some basic inequalities of ℓ_p norm and introduce the notions of RECs, as well as their relationships.

The notations adopted in this paper are described as follows. We let the lowercase letters x, y, z denote the vectors, calligraphic letters $\mathcal{I}$, $\mathcal{T}$, $\mathcal{S}$, $\mathcal{J}$, $\mathcal{N}$ denote the index sets, capital letters N, S denote the numbers of groups in the index sets. In particular, we use $\mathcal{G}_i$ to denote the index set corresponding to the i-th group and $\mathcal{G}_{\mathcal{S}}$ to denote the index set $\{\mathcal{G}_i : i \in \mathcal{S}\}$. For $x \in \mathbb{R}^n$ and $\mathcal{T} \subseteq \{1, \ldots, n\}$, we use $x_{\mathcal{T}}$ to denote the subvector of x corresponding to $\mathcal{T}$. We use $\mathrm{sign} : \mathbb{R} \to \mathbb{R}$ to denote the signum function, defined by

$$\mathrm{sign}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0. \end{cases}$$

Throughout this paper, we assume that the group sparse optimization problem is of the group structure described as follows. Let $x := (x_{\mathcal{G}_1}^\top, \cdots, x_{\mathcal{G}_r}^\top)^\top$ represent the group structure of x, where $\{x_{\mathcal{G}_i} \in \mathbb{R}^{n_i} : i = 1, \cdots, r\}$ is the grouping of x, $\sum_{i=1}^r n_i = n$ and $n_{\max} := \max\{n_i : i \in \{1, \ldots, r\}\}$. For a group $x_{\mathcal{G}_i}$, we use $x_{\mathcal{G}_i} = 0$ (reps. $x_{\mathcal{G}_i} \neq 0$, $x_{\mathcal{G}_i} \neq_{\mathbf{a}} 0$) to denote a zero (reps. nonzero, active) group, where $x_{\mathcal{G}_i} = 0$ means that $x_j = 0$ for all $j \in \mathcal{G}_i$; $x_{\mathcal{G}_i} \neq 0$ means that $x_j \neq 0$ for some $j \in \mathcal{G}_i$; and $x_{\mathcal{G}_i} \neq_{\mathbf{a}} 0$ means that $x_j \neq 0$ for all $j \in \mathcal{G}_i$. It is trivial to see that

$$x_{\mathcal{G}_i} \neq_{\mathbf{a}} 0 \quad \Rightarrow \quad x_{\mathcal{G}_i} \neq 0.$$

For this group structure and $p > 0$, the $\ell_{p,q}$ norm of x is defined by

$$\|x\|_{p,q} = \begin{cases} (\sum_{i=1}^r \|x_{\mathcal{G}_i}\|_p^q)^{1/q}, & q > 0, \\ \sum_{i=1}^r \|x_{\mathcal{G}_i}\|_p^0, & q = 0, \end{cases} \tag{6}$$

which proposes the ℓ_p norm for each group and then processes the ℓ_q norm for the resulting vector. When $p = q$, the $\ell_{p,q}$ norm coincides with the ℓ_p norm, that is, $\|x\|_{p,p} = \|x\|_p$. Furthermore, all $\ell_{p,0}$ norms share the same formula, that is, $\|x\|_{p,0} = \|x\|_{2,0}$ for all $p > 0$. In particular, when the grouping structure is degenerated to the individual feature level, that is, if $n_{\max} = 1$ or $n = r$, we have $\|x\|_{p,q} = \|x\|_q$ for all $p > 0$ and $q > 0$.

Moreover, we assume that A and b in (4) are related by a linear model (noiseless)

$$b = A\bar{x}.$$

Let $\mathcal{S} := \{i \in \{1, \ldots, r\} : \bar{x}_{\mathcal{G}_i} \neq 0\}$ be the index set of nonzero groups of $\bar{x}$, $\mathcal{S}^c := \{1, \ldots, r\} \setminus \mathcal{S}$ be the complement of $\mathcal{S}$, $S := |\mathcal{S}|$ be the group sparsity of $\bar{x}$, and $n_{\mathbf{a}} := \sum_{i \in \mathcal{S}} n_i$.

2.1 Inequalities of $\ell_{p,q}$ Norm

We begin with some basic inequalities of the ℓ_p and $\ell_{p,q}$ norms, which will be useful in the later discussion of RECs and recovery bounds. First, we recall the following well-known inequality

$$\left(\sum_{i=1}^n |x_i|^{\gamma_2}\right)^{1/\gamma_2} \leq \left(\sum_{i=1}^n |x_i|^{\gamma_1}\right)^{1/\gamma_1} \quad \text{if } 0 < \gamma_1 \leq \gamma_2, \tag{7}$$

or equivalently ($x = (x_1, x_2, \ldots, x_n)^\top$),

$$\|x\|_{\gamma_2} \leq \|x\|_{\gamma_1} \quad \text{if } 0 < \gamma_1 \leq \gamma_2.$$

The following lemma improves Huang and Yang (2003, lem. 4.1) and extends to the $\ell_{p,q}$ norm. It will be useful in providing a shaper global recovery bound (see Theorem 9 later).

Lemma 1 *Let $0 < q \leq p \leq 2$, $x \in \mathbb{R}^n$ and K be the smallest integer such that $2^{K-1}q \geq 1$. Then the following relations hold.*

(i) $\|x\|_q^q \leq n^{1-2^{-K}} \|x\|_2^q$.

(ii) $\|x\|_{p,q}^q \leq r^{1-2^{-K}} \|x\|_{p,2}^q$.

Proof (i) Repeatedly using the property that $\|x\|_1 \le \sqrt{n}\|x\|_2$, one has that

$$\|x\|_q^q \le \sqrt{n}\left(\sum_{i=1}^n |x_i|^{2q}\right)^{1/2} \le \cdots \le n^{\frac{1}{2}+\cdots+\frac{1}{2^K}}\left(\sum_{i=1}^n |x_i|^{2^K q}\right)^{2^{-K}}.$$

Since $2^{K-1}q \ge 1$, by (7), we obtain that

$$\left(\sum_{i=1}^n |x_i|^{2^K q}\right)^{2^{-K}} = \left(\sum_{i=1}^n (|x_i|^2)^{2^{K-1}q}\right)^{\frac{1}{2^{K-1}q}\frac{q}{2}} \le \left(\sum_{i=1}^n |x_i|^2\right)^{q/2} = \|x\|_2^q.$$

Therefore, we arrive at the conclusion that

$$\|x\|_q^q \le n^{1-2^{-K}}\|x\|_2^q.$$

(ii) By (6), it is a direct consequence of (i). ■

For example, if $q = 1$, then $K = 1$; if $q = \frac{1}{2}$ or $\frac{2}{3}$, then $K = 2$. The following lemma describes the triangle inequality of $\|\cdot\|_{p,q}^q$.

Lemma 2 *Let $0 < q \le 1 \le p$ and $x, y \in \mathbb{R}^n$. Then*

$$\|x\|_{p,q}^q - \|y\|_{p,q}^q \le \|x - y\|_{p,q}^q.$$

Proof By the subadditivity of the ℓ_p norm and (7), it is easy to see that

$$\|x_{\mathcal{G}_i}\|_p^q - \|y_{\mathcal{G}_i}\|_p^q \le \|x_{\mathcal{G}_i} - y_{\mathcal{G}_i}\|_p^q, \quad \text{for } i = 1, \ldots, r.$$

Consequently, the conclusion directly follows from (6). ■

The following lemma will be beneficial to studying properties of the lower-order REC in Proposition 5 later.

Lemma 3 *Let $\gamma \ge 1$, and two finite sequences $\{y_i : i \in \mathcal{I}\}$ and $\{x_j : j \in \mathcal{J}\}$ satisfy that $y_i \ge x_j \ge 0$ for all $(i, j) \in \mathcal{I} \times \mathcal{J}$. If $\sum_{i\in\mathcal{I}} y_i \ge \sum_{j\in\mathcal{J}} x_j$, then $\sum_{i\in\mathcal{I}} y_i^\gamma \ge \sum_{j\in\mathcal{J}} x_j^\gamma$.*

Proof Set $\bar{y} := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} y_i$ and $\alpha := \min_{i\in\mathcal{I}} y_i$. By Huang and Yang (2003, lem. 4.1(ii)), one has that

$$\sum_{i\in\mathcal{I}} y_i^\gamma \ge \frac{1}{|\mathcal{I}|^{\gamma-1}}\left(\sum_{i\in\mathcal{I}} y_i\right)^\gamma = |\mathcal{I}|\bar{y}^\gamma. \tag{8}$$

On the other hand, let $M \in \mathbb{N}$ and $\beta \in [0, \alpha)$ be such that $\sum_{j\in\mathcal{J}} x_j = M\alpha + \beta$. Observing that $\gamma \ge 1$ and $0 \le x_j \le \alpha$ for all $j \in \mathcal{J}$, we obtain that $x_j^\gamma \le x_j\alpha^{\gamma-1}$, and thus, $\sum_{j\in\mathcal{J}} x_j^\gamma \le M\alpha^\gamma + \alpha^{\gamma-1}\beta$. By (8), it remains to show that

$$|\mathcal{I}|\bar{y}^\gamma \ge M\alpha^\gamma + \alpha^{\gamma-1}\beta. \tag{9}$$

If $|\mathcal{I}| > M$, the relation (9) is trivial since $\bar{y} \ge \alpha > \beta$; otherwise, $|\mathcal{I}| \le M$, from the facts that $|\mathcal{I}|\bar{y} \ge M\alpha + \beta$ (that is, $\sum_{i\in\mathcal{I}} y_i \ge \sum_{j\in\mathcal{J}} x_j$) and that $\gamma \ge 1$, it follows that

$$|\mathcal{I}|\bar{y}^\gamma \ge M^{1-\gamma}(M\alpha + \beta)^\gamma \ge M^{1-\gamma}(M^\gamma\alpha^\gamma + \gamma M^{\gamma-1}\alpha^{\gamma-1}\beta) \ge M\alpha^\gamma + \alpha^{\gamma-1}\beta.$$

Therefore, we verify the relation (9), and the proof is complete. ■

2.2 Group Restricted Eigenvalue Conditions

This subsection aims at the development of the critical conditions on the matrix A to guarantee the oracle property and the global recovery bound of the $\ell_{p,q}$ regularization problem (4). In particular, we will focus on the restricted eigenvalue condition (REC), and extend it to the lower-order setting and equip it with the group structure.

In the scenario of sparse optimization, given the sparsity level s, it is always assumed that the $2s$-sparse minimal eigenvalue of $A^\top A$ is positive (see Bickel et al., 2009; Bunea et al., 2007; Meinshausen and Yu, 2009), that is,

$$\phi_{\min}(2s) := \min_{\|x\|_0 \le 2s} \frac{x^\top A^\top A x}{x^\top x} > 0, \tag{10}$$

which is the minimal eigenvalue of any $2s \times 2s$ dimensional submatrix. It is well-known that the solution at sparsity level s of the linear system $Ax = b$ is unique if the condition (10) is satisfied; otherwise, assume that there are two distinct vectors $\hat{x}$ and $\tilde{x}$ such that $A\hat{x} = A\tilde{x}$ and $\|\hat{x}\|_0 = \|\tilde{x}\|_0 = s$. Then $x := \hat{x} - \tilde{x}$ is a vector such that $Ax = 0$ and $\|x\|_0 \le 2s$, and thus $\phi_{\min}(2s) = 0$, which is contradict with (10). Therefore, if the $2s$-sparse minimal eigenvalue of $A^\top A$ is zero (that is, $\phi_{\min}(2s) = 0$), one has no hope of recovering the true sparse solution from noisy observations.

However, only condition (10) is not enough and some further condition is required to maintain the nice recovery of regularization problems; see Bickel et al. (2009); Bunea et al. (2007); Meinshausen and Yu (2009); van de Geer and Bühlmann (2009); Zhang (2009) and references therein. For example, the REC was introduced in Bickel et al. (2009) to investigate the ℓ_2 consistency of the ℓ_1 regularization problem (Lasso), where the minimum in (10) is replaced by a minimum over a restricted set of vectors measured by an ℓ_1 norm inequality and the denominator is replaced by the ℓ_2 norm of only a part of x.

We now introduce the notion of the lower-order REC. Note that the residual $\hat{x} := x^*(\ell_q) - \bar{x}$, where $x^*(\ell_q)$ is an optimal solution of the ℓ_q regularization problem and $\bar{x}$ is a sparse solution of $Ax = b$, of the ℓ_q regularization problem always satisfies

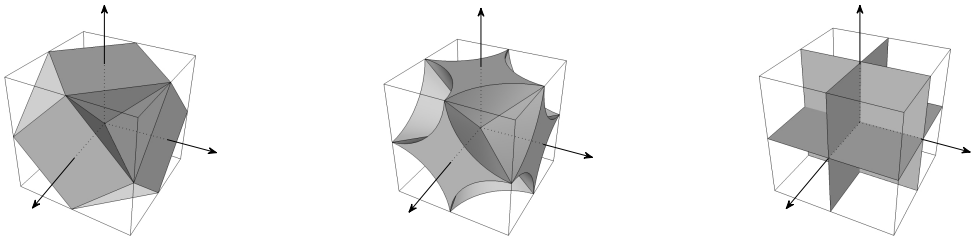
$$\|\hat{x}_{\mathcal{S}^c}\|_q \le \|\hat{x}_{\mathcal{S}}\|_q, \tag{11}$$

where $\mathcal{S}$ is the support of $\bar{x}$. Thus we introduce a lower-order REC, where the minimum is taken over a restricted set measured by an ℓ_q norm inequality such as (11), for establishing the global recovery bound of the ℓ_q regularization problem. Given $s \le t \ll n$, $x \in \mathbb{R}^n$ and $\mathcal{I} \subseteq \{1, \ldots, n\}$, we denote by $\mathcal{I}(x;t)$ the subset of $\{1, \ldots, n\}$ corresponding to the first t largest coordinates in absolute value of x in $\mathcal{I}^c$.

Definition 4 *Let $0 \le q \le 1$. The q-restricted eigenvalue condition relative to (s,t) (q-REC(s,t)) is said to be satisfied if*

$$\phi_q(s,t) := \min \left\{ \frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2} : |\mathcal{I}| \le s, \|x_{\mathcal{I}^c}\|_q \le \|x_{\mathcal{I}}\|_q, \mathcal{T} = \mathcal{I}(x;t) \cup \mathcal{I} \right\} > 0.$$

The q-REC describes a kind of restricted positive definiteness of $A^\top A$, which is valid only for the vectors satisfying the relation measured by an ℓ_q norm. The q-REC presents a unified framework of the REC-type conditions whenever $q \in [0,1]$. In particular, we note by



(a) REC (b) 1/2-REC (c) 0-REC

Figure 1: The geometric interpretation of the RECs: the gray regions show the feasible sets $C_q(s)$ ($q = 1, 1/2, 0$). The q-REC holds if and only if the null space of A does not intersect the gray region.

definition that 1-REC reduces to the classical REC (Bickel et al., 2009), and that $\phi_{\min}(2s) = \phi_0^2(s, s)$, and thus

$$(10) \quad \Leftrightarrow \quad 0\text{-REC}(s, s) \text{ is satisfied.}$$

It is well-known in the literature that the 1-REC is a stronger condition than the 0-REC (10). A natural question arises what are the relationships between the general q-RECs. To answer this question, associated with the q-REC, we consider the feasible set

$$C_q(s) := \{x \in \mathbb{R}^n : \|x_{\mathcal{I}^c}\|_q \le \|x_{\mathcal{I}}\|_q \text{ for some } |\mathcal{I}| \le s\},$$

which is a cone. Since the objective function associated with the q-REC is homogeneous, the q-REC(s, t) says that the null space of A does not cross over $C_q(s)$. Figure 1 presents the geometric interpretation of the q-RECs. It is shown in Figure 1 that $C_0(s) \subseteq C_{1/2}(s) \subseteq C_1(s)$, and thus

$$1\text{-REC} \quad \Rightarrow \quad 1/2\text{-REC} \quad \Rightarrow \quad 0\text{-REC}.$$

It is also observed from Figure 1 that the gap between the 1-REC and 1/2-REC and that between 1/2-REC and 0-REC are the matrices whose null spaces fall in the cones of $C_1(s) \setminus C_{1/2}(s)$ and $C_{1/2}(s) \setminus C_0(s)$, respectively.

We now provide a rigorous proof in the following proposition to identify the relationship between the feasible sets $C_q(s)$ and between the general q-RECs: the lower the q, the smaller the cone $C_q(s)$, and the weaker the q-REC.

Proposition 5 *Let $0 \le q_1 \le q_2 \le 1$ and $1 \le s \le t \ll n$. Then the following statements are true:*

(i) *$C_{q_1}(s) \subseteq C_{q_2}(s)$, and*

(ii) *if the q_2-REC(s, t) holds, then the q_1-REC(s, t) holds.*

Proof (i) Fix $x \in C_{q_1}(s)$. We use $\mathcal{I}_*$ to denote the index set of the first s largest coordinates in absolute value of x. Since $x \in C_{q_1}(s)$, it follows that $\|x_{\mathcal{I}_*^c}\|_{q_1} \leq \|x_{\mathcal{I}_*}\|_{q_1}$ ($|\mathcal{I}_*| \leq s$ due to the construction of $\mathcal{I}_*$). By Lemma 3 (taking $\gamma = q_2/q_1$), one has that

$$\|x_{\mathcal{I}_*^c}\|_{q_2} \leq \|x_{\mathcal{I}_*}\|_{q_2},$$

that is, $x \in C_{q_2}(s)$. Hence it follows that $C_{q_1}(s) \subseteq C_{q_2}(s)$.
(ii) As proved by (i) that $C_{q_1}(s) \subseteq C_{q_2}(s)$, by the definition of q-REC, it follows that

$$\phi_{q_1}(s,t) \geq \phi_{q_2}(s,t) > 0.$$

The proof is complete. ■

To the best of our knowledge, this is the first work on introducing the lower-order REC and establishing the relationship of the lower-order RECs. In the following, we provide a counter example to show that the reverse of Proposition 5 is not true.

Example 1 (A matrix satisfying $1/2$-REC but not REC) *Consider the matrix*

$$A := \begin{pmatrix} a & a+c & a-c \\ \tilde{a} & \tilde{a}-\tilde{c} & \tilde{a}+\tilde{c} \end{pmatrix} \in \mathbb{R}^{2\times 3},$$

where $a > c > 0$ and $\tilde{a} > \tilde{c} > 0$. This matrix A does not satisfy the REC$(1,1)$. Indeed, by letting $\mathcal{J} = \{1\}$ and $x = (2,-1,-1)^\top$, we have $Ax = 0$ and thus $\phi(1,1) = 0$.

Below, we claim that A satisfies the $1/2$-REC$(1,1)$. It suffices to show that $\phi_{1/2}(1,1) > 0$. Let $x = (x_1,x_2,x_3)^\top$ satisfy the constraint associated with $1/2$-REC$(1,1)$. As $s = 1$, the deduction is divided into the following three cases.

(i) $\mathcal{J} = \{1\}$. *Then*

$$|x_1| \geq \|x_{\mathcal{J}^c}\|_{1/2} = |x_2| + |x_3| + 2|x_2|^{1/2}|x_3|^{1/2}. \tag{12}$$

Without loss of generality, we assume $|x_1| \geq |x_2| \geq |x_3|$. Hence, $\mathcal{T} = \{1,2\}$ and

$$\frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2} \geq \frac{\min\{a,\tilde{a}\}|x_1+x_2+x_3| + \min\{c,\tilde{c}\}|x_2-x_3|}{|x_1|+|x_2|}. \tag{13}$$

If $|x_2| \leq \frac{1}{3}|x_1|$, (13) reduces to

$$\frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2} \geq \frac{\frac{\min\{a,\tilde{a}\}}{3}|x_1|}{\frac{4}{3}|x_1|} = \frac{\min\{a,\tilde{a}\}}{4}. \tag{14}$$

If $|x_2| \geq \frac{1}{3}|x_1|$, substituting (12) into (13), one has that

$$\frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2} \geq \begin{cases} \frac{\min\{c,\tilde{c}\}}{8}, & |x_3| \leq \frac{1}{2}|x_2|, \\ \frac{\min\{a,\tilde{a}\}}{4}, & |x_3| \geq \frac{1}{2}|x_2|. \end{cases} \tag{15}$$

(ii) $\mathcal{J}=\{2\}$. *Since* $\mathcal{T}=\{2,1\}$ *or* $\{2,3\}$, *it follows from Huang and Yang (2003, lem. 4.1(i-i)) that*

$$|x_2|\geq\|x_{\mathcal{J}^c}\|_{1/2}\geq|x_1|+|x_3|. \tag{16}$$

Thus, it is easy to verify that $\|x_{\mathcal{T}}\|_2\leq 2|x_2|$ *and that*

$$\frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2}\geq\frac{|ax_1+(a+c)x_2+(a-c)x_3|}{2|x_2|}=\frac{|a(x_1+x_2+\frac{a-c}{a}x_3)+cx_2|}{2|x_2|}\geq\frac{c}{2}, \tag{17}$$

where the last inequality follows from (16) *and the fact that* $a>c$.

(iii) $\mathcal{J}=\{3\}$. *Similar to the deduction of* (ii), *one has that*

$$\frac{\|Ax\|_2}{\|x_{\mathcal{T}}\|_2}\geq\frac{|\tilde{a}x_1+(\tilde{a}-\tilde{c})x_2+(\tilde{a}+\tilde{c})x_3|}{2|x_3|}\geq\frac{\tilde{c}}{2}. \tag{18}$$

Therefore, by (14)-(15) *and* (17)-(18), *we conclude that* $\phi_{1/2}(1,1)\geq\frac{1}{8}\min\{c,\tilde{c}\}>0$, *and thus, the matrix* A *satisfies the* $1/2$-*REC*$(1,1)$.

In order to establish the oracle property and the global recovery bound for the $\ell_{p,q}$ regularization problem, we further introduce the notion of group restricted eigenvalue condition (GREC). Given $S\leq N\ll r$, $x\in\mathbb{R}^n$ and $\mathcal{J}\subseteq\{1,\ldots,r\}$, we use $\mathrm{rank}_i(x)$ to denote the rank of $\|x_{\mathcal{G}_i}\|_p$ among $\{\|x_{\mathcal{G}_j}\|_p:j\in\mathcal{J}^c\}$ (in a decreasing order), $\mathcal{J}(x;N)$ to denote the index set of the first N largest groups in the value of $\|x_{\mathcal{G}_i}\|_p$ among $\{\|x_{\mathcal{G}_j}\|_p:j\in\mathcal{J}^c\}$, that is,

$$\mathcal{J}(x;N):=\{i\in\mathcal{J}^c:\mathrm{rank}_i(x)\in\{1,\ldots,N\}\}.$$

Furthermore, by letting $R:=\lceil\frac{r-|\mathcal{J}|}{N}\rceil$, we denote

$$\mathcal{J}_k(x;N):=\begin{cases}\{i\in\mathcal{J}^c:\mathrm{rank}_i(x)\in\{kN+1,\ldots,(k+1)N\}\}, & k=1,\ldots,R-1,\\ \{i\in\mathcal{J}^c:\mathrm{rank}_i(x)\in\{RN+1,\ldots,r-|\mathcal{J}|\}\}, & k=R.\end{cases} \tag{19}$$

Note that the residual $\hat{x}:=x^*(\ell_{p,q})-\bar{x}$ of the $\ell_{p,q}$ regularization problem always satisfies $\|\hat{x}_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}\leq\|\hat{x}_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}$. Thus we introduce the notion of GREC, where the minimum is taken over a restricted set measured by an $\ell_{p,q}$ norm inequality, as follows.

Definition 6 *Let* $0<q\leq p\leq 2$. *The* (p,q)-*group restricted eigenvalue condition relative to* (S,N) ((p,q)-*GREC*(S,N)) *is said to be satisfied if*

$$\phi_{p,q}(S,N):=\min\left\{\frac{\|Ax\|_2}{\|x_{\mathcal{G}_{\mathcal{N}}}\|_{p,2}}:|\mathcal{J}|\leq S,\|x_{\mathcal{G}_{\mathcal{J}^c}}\|_{p,q}\leq\|x_{\mathcal{G}_{\mathcal{J}}}\|_{p,q},\mathcal{N}=\mathcal{J}(x;N)\cup\mathcal{J}\right\}>0.$$

The (p,q)-GREC extends the q-REC to the setting equipping with a pre-defined group structure. Handling the components in each group as one element, the (p,q)-GREC admits the fewer degree of freedom, which is S (about $s/n_{\max}$), on its associated constraint than that of the q-REC, and thus it characterizes a weaker condition than the q-REC. For example, the 0-REC(s,s) is to indicate the restricted positive definiteness of $A^\top A$, which is valid only for the vectors whose cardinality is less than $2s$; while the $(p,0)$-GREC(S,S) is to describe the restricted positive definiteness of $A^\top A$ on any $2S$-group support, whose degree

of freedom is much less than the $2s$-support. Thus the $(p,0)$-GREC(S,S) provides a broader condition than the 0-REC(s,s). Similar to the proof of Proposition 5, we can show that if $0 \le q_1 \le q_2 \le 1 \le p \le 2$ and the (p,q_2)-GREC(S,N) holds, then the (p,q_1)-GREC(S,N) also holds.

We end this subsection by providing the following lemma, which will be useful in establishing the global recovery bound for the $\ell_{p,q}$ regularization problem in Theorem 9.

Lemma 7 *Let $0 < q \le 1 \le p$, $\tau \ge 1$ and $x \in \mathbb{R}^n$, $\mathcal{N} := \mathcal{J}(x;N) \cup \mathcal{J}$ and $\mathcal{J}_k := \mathcal{J}_k(x;N)$ for $k = 1, \ldots, R$. Then the following inequalities hold*

$$\|x_{\mathcal{G}_{\mathcal{N}^c}}\|_{p,\tau} \le \sum_{k=1}^{R} \|x_{\mathcal{G}_{\mathcal{J}_k}}\|_{p,\tau} \le N^{\frac{1}{\tau}-\frac{1}{q}} \|x_{\mathcal{G}_{\mathcal{J}^c}}\|_{p,q}.$$

Proof By the definition of $\mathcal{J}_k$ (19), for each $j \in \mathcal{J}_k$, one has that

$$\|x_{\mathcal{G}_j}\|_p \le \|x_{\mathcal{G}_i}\|_p, \quad \text{for each } i \in \mathcal{J}_{k-1},$$

and thus

$$\|x_{\mathcal{G}_j}\|_p^q \le \frac{1}{N} \sum_{i \in \mathcal{J}_{k-1}} \|x_{\mathcal{G}_i}\|_p^q = \frac{1}{N} \|x_{\mathcal{G}_{\mathcal{J}_{k-1}}}\|_{p,q}^q.$$

Consequently, we obtain that

$$\|x_{\mathcal{G}_{\mathcal{J}_k}}\|_{p,\tau}^\tau = \sum_{i \in \mathcal{J}_k} \|x_{\mathcal{G}_i}\|_p^\tau \le N^{1-\tau/q} \|x_{\mathcal{G}_{\mathcal{J}_{k-1}}}\|_{p,q}^\tau.$$

Further by Huang and Yang (2003, lem. 4.1) ($\tau \ge 1$ and $q \le 1$), it follows that

$$\begin{aligned} \|x_{\mathcal{G}_{\mathcal{N}^c}}\|_{p,\tau} &= \left(\sum_{k=1}^{R} \sum_{i \in \mathcal{J}_k} \|x_{\mathcal{G}_i}\|_p^\tau\right)^{1/\tau} \\ &\le \sum_{k=1}^{R} \|x_{\mathcal{G}_{\mathcal{J}_k}}\|_{p,\tau} \\ &\le N^{\frac{1}{\tau}-\frac{1}{q}} \sum_{k=1}^{R} \|x_{\mathcal{G}_{\mathcal{J}_{k-1}}}\|_{p,q} \\ &\le N^{\frac{1}{\tau}-\frac{1}{q}} \|x_{\mathcal{G}_{\mathcal{J}^c}}\|_{p,q}. \end{aligned}$$

The proof is complete. ■

2.3 Global Recovery Bound

In recent years, many articles have been devoted to establishing the oracle property and the global recovery bound for the ℓ_1 regularization problem (1) under the RIP or REC; see Bickel et al. (2009); Meinshausen and Yu (2009); van de Geer and Bühlmann (2009); Zhang (2009). However, to the best of our knowledge, there is few paper devoted to investigating these properties for the lower-order regularization problem.

In the preceding subsections, we have introduced the general notion of (p,q)-GREC. Under the (p,q)-GREC(S,S), the solution of $Ax = b$ with group sparsity being S is unique. In this subsection, we will present the oracle property and the global recovery bound for

the $\ell_{p,q}$ regularization problem (4) under the (p,q)-GREC. The oracle property provides an upper bound on the squares error of the linear system and the violation of the true nonzero groups for each point in the level set of the objective function of problem (4)

$$\mathrm{lev}_F(\bar{x}) := \{x \in \mathbb{R}^n : \|Ax-b\|_2^2 + \lambda\|x\|_{p,q}^q \le \lambda\|\bar{x}\|_{p,q}^q\}.$$

Proposition 8 *Let $0 < q \le 1 \le p$, $S > 0$ and let the (p,q)-GREC(S,S) hold. Let $\bar{x}$ be the unique solution of $Ax = b$ at a group sparsity level S, and $\mathcal{S}$ be the index set of nonzero groups of $\bar{x}$. Let K be the smallest integer such that $2^{K-1}q \ge 1$. Then, for any $x^* \in \mathrm{lev}_F(\bar{x})$, the following oracle inequality holds*

$$\|Ax^* - A\bar{x}\|_2^2 + \lambda\|x^*_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}^q \le \lambda^{\frac{2}{2-q}} S^{\left(1-2^{-K}\right)\frac{2}{2-q}} / \phi_{p,q}^{\frac{2q}{2-q}}(S,S). \tag{20}$$

Moreover, letting $\mathcal{N}_* := \mathcal{S} \cup \mathcal{S}(x^*;S)$, we have

$$\|x^*_{\mathcal{G}_{\mathcal{N}_*}} - \bar{x}_{\mathcal{G}_{\mathcal{N}_*}}\|_{p,2}^2 \le \lambda^{\frac{2}{2-q}} S^{\left(1-2^{-K}\right)\frac{2}{2-q}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S).$$

Proof Let $x^* \in \mathrm{lev}_F(\bar{x})$. That is, $\|Ax^*-b\|_2^2 + \lambda\|x^*\|_{p,q}^q \le \lambda\|\bar{x}\|_{p,q}^q$. By Lemmas 1(ii) and 2, one has that

$$\begin{aligned} \|Ax^* - A\bar{x}\|_2^2 + \lambda\|x^*_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}^q &\le \lambda\|\bar{x}_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}^q - \lambda\|x^*_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}^q \\ &\le \lambda\|\bar{x}_{\mathcal{G}_{\mathcal{S}}} - x^*_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}^q \\ &\le \lambda S^{1-2^{-K}}\|\bar{x}_{\mathcal{G}_{\mathcal{S}}} - x^*_{\mathcal{G}_{\mathcal{S}}}\|_{p,2}^q. \end{aligned} \tag{21}$$

Noting that

$$\|x^*_{\mathcal{G}_{\mathcal{S}^c}} - \bar{x}_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}^q - \|x^*_{\mathcal{G}_{\mathcal{S}}} - \bar{x}_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}^q \le \|x^*_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}^q - (\|\bar{x}_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}^q - \|x^*_{\mathcal{G}_{\mathcal{S}}}\|_{p,q}^q) = \|x^*\|_{p,q}^q - \|\bar{x}\|_{p,q}^q \le 0.$$

Then the (p,q)-GREC(S,S) implies that

$$\|\bar{x}_{\mathcal{G}_{\mathcal{S}}} - x^*_{\mathcal{G}_{\mathcal{S}}}\|_{p,2} \le \|Ax^* - A\bar{x}\|_2 / \phi_{p,q}(S,S).$$

This, together with (21), yields that

$$\|Ax^* - A\bar{x}\|_2^2 + \lambda\|x^*_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}^q \le \lambda S^{1-2^{-K}}\|Ax^* - A\bar{x}\|_2^q / \phi_{p,q}^q(S,S), \tag{22}$$

and consequently,

$$\|Ax^* - A\bar{x}\|_2 \le \lambda^{\frac{1}{2-q}} S^{\left(1-2^{-K}\right)/(2-q)} / \phi_{p,q}^{\frac{q}{2-q}}(S,S). \tag{23}$$

Therefore, by (22) and (23), we arrive at the oracle inequality (20). Furthermore, by the definition of $\mathcal{N}_*$, the (p,q)-GREC(S,S) implies that

$$\|x^*_{\mathcal{G}_{\mathcal{N}_*}} - \bar{x}_{\mathcal{G}_{\mathcal{N}_*}}\|_{p,2}^2 \le \|Ax^* - A\bar{x}\|_2^2 / \phi_{p,q}^2(S,S) \le \lambda^{\frac{2}{2-q}} S^{\left(1-2^{-K}\right)\frac{2}{2-q}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S).$$

The proof is complete. ■

One of the main results of this section is presented as follows, where we establish the global recovery bound for the $\ell_{p,q}$ regularization problem under the (p,q)-GREC. We will apply oracle inequality (20) and Lemma 7 in our proof.

Theorem 9 *Let $0 < q \le 1 \le p \le 2$, $S > 0$ and let the (p,q)-GREC(S,S) hold. Let $\bar{x}$ be the unique solution of $Ax = b$ at a group sparsity level S, and $\mathcal{S}$ be the index set of nonzero groups of $\bar{x}$. Let K be the smallest integer such that $2^{K-1}q \ge 1$. Then, for any $x^* \in \mathrm{lev}_F(\bar{x})$, the following global recovery bound for problem (4) holds*

$$\|x^* - \bar{x}\|_2^2 \le 2\lambda^{\frac{2}{2-q}} S^{\frac{q-2}{q} + \left(1-2^{-K}\right)\frac{4}{q(2-q)}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S). \tag{24}$$

More precisely,

$$\|x^* - \bar{x}\|_2^2 \le \begin{cases} \mathcal{O}(\lambda^{\frac{2}{2-q}} S), & 2^{K-1}q = 1, \\ \mathcal{O}(\lambda^{\frac{2}{2-q}} S^{\frac{3-q}{2-q}}), & 2^{K-1}q > 1. \end{cases} \tag{25}$$

Proof Let $\mathcal{N}_* := \mathcal{S} \cup \mathcal{S}(x^*; S)$ as in Proposition 8. Since $p \le 2$, it follows from Lemma 7 and Proposition 8 that

$$\|x^*_{\mathcal{G}_{\mathcal{N}_*^c}}\|_2^2 \le \|x^*_{\mathcal{G}_{\mathcal{N}_*^c}}\|_{p,2}^2 \le S^{1-2/q}\|x^*_{\mathcal{G}_{\mathcal{S}^c}}\|_{p,q}^2 \le \lambda^{\frac{2}{2-q}} S^{\frac{q-2}{q} + \left(1-2^{-K}\right)\frac{4}{q(2-q)}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S).$$

Then by Proposition 8, one has that

$$\begin{aligned} \|x^* - \bar{x}\|_2^2 &= \|x^*_{\mathcal{G}_{\mathcal{N}_*}} - \bar{x}_{\mathcal{G}_{\mathcal{N}_*}}\|_2^2 + \|x^*_{\mathcal{G}_{\mathcal{N}_*^c}}\|_2^2 \\ &\le \lambda^{\frac{2}{2-q}} S^{\left(1-2^{-K}\right)\frac{2}{2-q}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S) + \lambda^{\frac{2}{2-q}} S^{\frac{q-2}{q} + \left(1-2^{-K}\right)\frac{4}{q(2-q)}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S) \\ &\le 2\lambda^{\frac{2}{2-q}} S^{\frac{q-2}{q} + \left(1-2^{-K}\right)\frac{4}{q(2-q)}} / \phi_{p,q}^{\frac{4}{2-q}}(S,S), \end{aligned}$$

where the last inequality follows from the fact that $2^{K-1}q \ge 1$. This proves (24). In particular, if $2^{K-1}q = 1$, then $\frac{q-2}{q} + \left(1-2^{-K}\right)\frac{4}{q(2-q)} = 1$ and thus

$$\|x^* - \bar{x}\|_2^2 \le \mathcal{O}(\lambda^{\frac{2}{2-q}} S).$$

If $2^{K-1}q > 1$, then $2^{K-2}q < 1$. Hence, $\frac{q-2}{q} + \left(1-2^{-K}\right)\frac{4}{q(2-q)} < \frac{3-q}{2-q}$, and consequently

$$\|x^* - \bar{x}\|_2^2 \le \mathcal{O}(\lambda^{\frac{2}{2-q}} S^{\frac{3-q}{2-q}}).$$

Hence (25) is obtained. The proof is complete. ■

The global recovery bound (25) is deduced under general (p,q)-GREC(S,S), which is weaker than the REC as used in van de Geer and Bühlmann (2009). It shows that the sparse solution $\bar{x}$ may be recovered by any point x^* in the level set $\mathrm{lev}_F(\bar{x})$, in particular, when x^* is a global optimal solution of problem (4), as long as λ is sufficiently small. It is well-known that when $p = 2$ and $q = 1$, convex regularization problem (4) is numerically intractable for finding the sparse solution and that when $q < 1$ finding a point in the nonconvex level set $\mathrm{lev}_F(\bar{x})$ is equivalent to finding a global minimum of minimizing the indicator function of the nonconvex level set, which is NP-hard. Thus Theorem 9 is only a theoretical result and does not provide any insight for the numerical computation of a sparse solution by virtue of problem (4). We will design a proximal gradient method in section 3, test its numerical efficiency and provide some general guidance on which q is more attractive in practical applications in section 4.

To conclude this subsection, we illustrate by an example in which (24) does not hold when $q = 1$, but it does and is also tight when $q = \frac{1}{2}$. We will testify the recovery bound $\mathcal{O}(\lambda^{4/3}S)$ in (24) when $q = \frac{1}{2}$ by using a global optimization method.

Example 2 *By letting $a = \tilde{a} = 2$ and $c = \tilde{c} = 1$ in Example 1, we consider the following matrix:*

$$A := \begin{pmatrix} 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix}.$$

We set $b := (2, 2)^\top$ and then a true solution of $Ax = b$ is $\bar{x} := (1, 0, 0)^\top$. Denoting $x := (x_1, x_2, x_3)^\top$, the objective function associated with the ℓ_1 regularization problem (1) is

$$\begin{aligned} F(x) &:= \|Ax - b\|_2^2 + \lambda\|x\|_1 \\ &= (2x_1 + 3x_2 + x_3 - 2)^2 + (2x_1 + x_2 + 3x_3 - 2)^2 + \lambda(|x_1| + |x_2| + |x_3|). \end{aligned}$$

Let $x^*(\ell_1) := (x_1^*, x_2^*, x_3^*)^\top$ be an optimal solution of problem (1). Without loss of generality, we assume $\lambda \le 1$. The necessary condition of $x^*(\ell_1)$ being an optimal solution of problem (1) is $0 \in \partial F(x^*(\ell_1))$, that is,

$$0 \in 16x_1^* + 16x_2^* + 16x_3^* - 16 + \lambda\partial|x_1^*|, \tag{26a}$$

$$0 \in 16x_1^* + 20x_2^* + 12x_3^* - 16 + \lambda\partial|x_2^*|, \tag{26b}$$

$$0 \in 16x_1^* + 12x_2^* + 20x_3^* - 16 + \lambda\partial|x_3^*|, \tag{26c}$$

where $\partial|\mu| := \begin{cases} sign(\mu), & \mu \neq 0, \\ [-1, 1], & \mu = 0. \end{cases}$

We first show that $x_i^* \ge 0$ for $i = 1, 2, 3$ by contradiction. Indeed, if $x_1^* < 0$, (26a) *reduces to*

$$16x_1^* + 16x_2^* + 16x_3^* - 16 = \lambda.$$

Summing (26b) *and* (26c)*, we further have*

$$\lambda = 16x_1^* + 16x_2^* + 16x_3^* - 16 \in -\frac{\lambda}{2}(\partial|x_2^*| + \partial|x_3^*|),$$

which implies that $x_2^* \le 0$ and $x_3^* \le 0$. Hence, it follows that $F(x^*) > F(0)$, which indicates that x^* is not an optimal solution of problem (1)*, and thus, $x_1^* < 0$ is impossible. Similarly, we can show that $x_2^* \ge 0$ and $x_3^* \ge 0$.*

Next, we find the optimal solution $x^*(\ell_1)$ by only considering $x^*(\ell_1) \ge 0$. It is easy to obtain that the solution of (26) *and the corresponding objective value associated with problem* (1) *can be represented respectively by*

$$x_1^* = 1 - \frac{\lambda}{16} - 2x_3^*, \quad x_2^* = x_3^* \ \left(0 \le x_3^* \le \frac{1}{2} - \frac{\lambda}{32}\right), \quad \text{and} \quad F\left(x^*(\ell_1)\right) = \lambda - \frac{\lambda^2}{32}.$$

Hence, $x^*(\ell_1) := \left(0, \frac{1}{2} - \frac{\lambda}{32}, \frac{1}{2} - \frac{\lambda}{32}\right)^\top$ is an optimal solution of problem (1)*. The estimated error for this $x^*(\ell_1)$ is*

$$\|x^*(\ell_1) - \bar{x}\|_2^2 = 1 + \frac{1}{2}\left(1 - \frac{\lambda}{16}\right)^2 > 1,$$

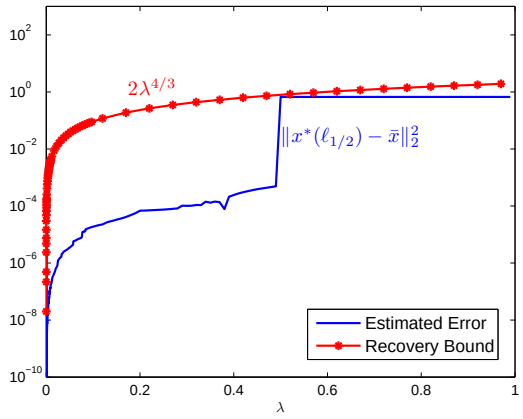


Figure 2: The illustration of the recovery bound (24) and estimated error.

which does not meet the recovery bound (25) *for any* $\lambda \leq 1$.

It is revealed from Example 1 that this matrix A *satisfies the* $1/2$*-REC*$(1,1)$. *Then the hypothesis of Theorem 9 is verified, and thus, Theorem 9 is applicable to establishing the recovery bound* (25) *for the* $\ell_{1/2}$ *regularization problem. Although we cannot obtain the closed-form solution of this nonconvex* $\ell_{1/2}$ *regularization problem, as it is of only 3-dimensions, we can apply a global optimization method, the filled function method (Ge (1990)), to find the global optimal solution* $x^*(\ell_{1/2})$ *and thus to testify the recovery bound* (25). *This is done by computing the* $\ell_{1/2}$ *regularization problem for many* λ *to plot the curve* $\|x^*(\ell_{1/2}) - \bar{x}\|_2^2$. *Figure 2 illustrates the variation of the estimated error* $\|x^*(\ell_{1/2}) - \bar{x}\|_2^2$ *and the bound* $2\lambda^{4/3}$ *(that is the right-hand side of* (24), *where* $S = 1$ *and* $\phi_{1/2}(1,1) \leq 1$ *(see Example 1)), when varying the regularization parameter* λ *from* 10^{-8} *to* 1. *It is illustrated from Figure 2 the recovery bound* (25) *is satisfied, and it is indeed tight, for this example.*

2.4 Local Recovery Bound

In the preceding subsection, we provided the global analysis of the recovery bound for the $\ell_{p,q}$ regularization problem under the (p,q)-GREC; see Theorem 9. One can also observe from Figure 2 that the global recovery bound (25) is tight for the $\ell_{1/2}$ regularization problem as the curves come together at $\lambda \simeq 0.5$, but there is still a big gap for the improvement when λ is small.

This subsection is devoted to providing a local analysis of the recovery bound for the $\ell_{p,q}$ regularization problem by virtue of the variational analysis technique (Rockafellar and Wets, 1998). For $x \in \mathbb{R}^n$ and $\delta \in \mathbb{R}_+$, we use $\mathbf{B}(x,\delta)$ to denote the open ball of radius δ centered at x. For a lower semi-continuous (lsc) function $f : \mathbb{R}^n \to \mathbb{R}$ and $x, w \in \mathbb{R}^n$, the subderivative of f at x along the direction w is defined by

$$df(\bar{x})(w) := \liminf_{\tau \downarrow 0,\ w' \to w} \frac{f(\bar{x} + \tau w') - f(\bar{x})}{\tau}.$$

To begin with, we show in the following lemma a significant advantage of lower-order regularization over the ℓ_1 regularization: the lower-order regularization term can easily induce the sparsity of the local minimum.

Lemma 10 *Let $0 < q < 1 \leq p$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a lsc function satisfying $df(0)(0) = 0$. Then the function $F := f + \lambda\|\cdot\|_{p,q}^q$ has a local minimum at 0 with the first-order growth condition being fulfilled, that is, there exist some $\epsilon > 0$ and $\delta > 0$ such that*

$$F(x) \geq F(0) + \epsilon\|x\|_2 \quad \textit{for any } x \in \mathbf{B}(0, \delta).$$

Proof Let $\varphi := \lambda\|\cdot\|_{p,q}^q$ and then $F = f + \varphi$. Since φ is grouped separable, by Rockafellar and Wets (1998, prop. 10.5), it follows from the definition that $d\varphi(0) = \delta_{\{0\}}$, where δ_X is the indicator function of X. Applying Rockafellar and Wets (1998, prop. 10.9), it follows that

$$dF(0) \geq df(0) + d\varphi(0) = df(0) + \delta_{\{0\}}. \tag{27}$$

By the assumption that f is finite and $df(0)(0) = 0$, its subderivative $df(0)$ is proper (see Rockafellar and Wets, 1998, ex. 3.19). Noting that $df(0)(0) = 0$, we obtain that $df(0) + \delta_{\{0\}} = \delta_{\{0\}}$. This, together with (27), yields that $dF(0) \geq \delta_{\{0\}}$. Therefore, by definition, there exist some $\epsilon > 0$ and $\delta > 0$ such that

$$F(x) \geq F(0) + \epsilon\|x\|_2 \quad \text{for any } x \in \mathbf{B}(0, \delta).$$

The proof is complete. ■

With the help of the above lemma, we can present in the following a local version of the recovery bound. This is done by constructing a path of local minima depending on the regularization parameter λ for the regularization problem, which starts from a sparse solution of the original problem and shares the same support as this sparse solution has, resulting in a sharper bound in terms of λ^2.

Theorem 11 *Let $\bar{x}$ be a solution of $Ax = b$, S be the group sparsity of $\bar{x}$, and B be a submatrix of A consisting of its columns corresponding to the active components of $\bar{x}$. Suppose that any nonzero group of $\bar{x}$ is active, and that the columns of A corresponding to the active components of $\bar{x}$ are linearly independent. Let $0 < q < 1 \leq p$. Then there exist $\kappa > 0$ and a path of local minima of problem (4), $x^*(\lambda)$, such that*

$$\|x^*(\lambda) - \bar{x}\|_2^2 \leq \lambda^2 q^2 S\|(B^\top B)^{-1}\|^2 \max_{\bar{x}_{\mathcal{G}_i} \neq 0} \left(\|\bar{x}_{\mathcal{G}_i}\|_p^{2(q-p)} \|\bar{x}_{\mathcal{G}_i}\|_{2p-2}^{2p-2}\right) \quad \textit{for any } \lambda < \kappa.$$

Proof Without loss of generality, we let $\bar{x}$ be of structure $\bar{x} = (\bar{z}^\top, 0)^\top$ with

$$\bar{z} = (\bar{x}_{\mathcal{G}_1}^\top, \ldots, \bar{x}_{\mathcal{G}_S}^\top)^\top \text{ and } \bar{x}_{\mathcal{G}_i} \neq_{\mathbf{a}} 0 \text{ for } i = 1, \ldots, S,$$

and let s be the sparsity of $\bar{x}$. Let $A = (B, D)$ with B being the submatrix involving the first s columns of A (corresponding to the active components of $\bar{x}$). By the assumption, we have that B is of full column rank and thus $B^\top B$ is invertible. In this setting, the linear relation $A\bar{x} = b$ reduces to $B\bar{z} = b$. The proof of this theorem is divided into the following three steps:

(a) construct a smooth path from $\bar{x}$ by the implicit function theorem;

(b) validate that every point of the constructed path is a local minimum of (4); and

(c) establish the recovery bound for the constructed path.

First, to show (a), we define $H : \mathbb{R}^{s+1} \to \mathbb{R}^s$ by

$$H(z,\lambda) = 2B^\top(Bz-b) + \lambda q \begin{pmatrix} \|z_{\mathcal{G}_1}\|_p^{q-p}\sigma(z_{\mathcal{G}_1}) \\ \vdots \\ \|z_{\mathcal{G}_S}\|_p^{q-p}\sigma(z_{\mathcal{G}_S}) \end{pmatrix},$$

where $\sigma(z_{\mathcal{G}_i}) = \text{vector}\left(|z_j|^{p-1}\text{sign}(z_j)\right)_{\mathcal{G}_i}$, denoting a vector consisting of $|z_j|^{p-1}\text{sign}(z_j)$ for all $j \in \mathcal{G}_i$. Let $\bar{\delta} > 0$ be sufficiently small such that $\text{sign}(z) = \text{sign}(\bar{z})$ for any $z \in \mathbf{B}(\bar{z}, \bar{\delta})$ and thus H is smooth on $\mathbf{B}(\bar{z}, \bar{\delta}) \times \mathbb{R}$. Note that $H(\bar{z}, 0) = 0$ and $\frac{\partial H}{\partial z}(\bar{z}, 0) = 2B^\top B$. By the implicit function theorem (Rudin, 1976), there exist some $\kappa > 0$, $\delta \in (0, \bar{\delta})$ and a unique smooth function $\xi : (-\kappa, \kappa) \to \mathbf{B}(\bar{z}, \delta)$ such that

$$\{(z,\lambda) \in \mathbf{B}(\bar{z},\bar{\delta}) \times (-\kappa,\kappa) : H(z,\lambda) = 0\} = \{(\xi(\lambda),\lambda) : \lambda \in (-\kappa,\kappa)\}, \tag{28}$$

and

$$\frac{d\xi}{d\lambda} = -q\left(2B^\top B + \lambda q \begin{pmatrix} M_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_{\mathcal{S}} \end{pmatrix}\right)^{-1} \begin{pmatrix} \|\xi(\lambda)_{\mathcal{G}_1}\|_p^{q-p}\sigma(\xi(\lambda)_{\mathcal{G}_1}) \\ \vdots \\ \|\xi(\lambda)_{\mathcal{G}_S}\|_p^{q-p}\sigma(\xi(\lambda)_{\mathcal{G}_S}) \end{pmatrix}, \tag{29}$$

where M_i for each $i = 1, \ldots, S$ is denoted by

$$M_i = (q-p)\|\xi(\lambda)_{\mathcal{G}_i}\|_p^{q-2p}(\sigma(\xi(\lambda)_{\mathcal{G}_i}))(\sigma(\xi(\lambda)_{\mathcal{G}_i}))^\top + (p-1)\|\xi(\lambda)_{\mathcal{G}_i}\|_p^{q-p}\text{diag}\left(|\xi(\lambda)_j|^{p-2}\right),$$

and $\text{diag}\left(|\xi(\lambda)_j|^{p-2}\right)$ denotes a diagonal matrix generated by vector $\left(|\xi(\lambda)_j|^{p-2}\right)$. Thus, by (28) and (29), we have constructed a smooth path $\xi(\lambda)$ near $\bar{z}$, $\lambda \in (-\kappa, \kappa)$, such that

$$2B^\top(B\xi(\lambda)-b) + \lambda q \begin{pmatrix} \|\xi(\lambda)_{\mathcal{G}_1}\|_p^{q-p}\sigma(\xi(\lambda)_{\mathcal{G}_1}) \\ \vdots \\ \|\xi(\lambda)_{\mathcal{G}_S}\|_p^{q-p}\sigma(\xi(\lambda)_{\mathcal{G}_S}) \end{pmatrix} = 0 \tag{30}$$

and

$$2B^\top B + \lambda q \begin{pmatrix} M_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_{\mathcal{S}} \end{pmatrix} \succ 0. \tag{31}$$

This shows that (a) is done as desired.

For fixed $\lambda \in (-\kappa, \kappa)$, let $x^*(\lambda) := (\xi(\lambda)^\top, 0)^\top$. To verify (b), we prove that $x^*(\lambda)$, with $\xi(\lambda)$ satisfying (30) and (31), is a local minimum of problem (4). Let $h : \mathbb{R}^s \to \mathbb{R}$ be a function with $h(z) := \|Bz-b\|_2^2 + \lambda\|z\|_{p,q}^q$ for any $z \in \mathbb{R}^s$. Note that $h(\xi(\lambda)) = \|Ax^*(\lambda)-b\|_2^2 + \lambda\|x^*(\lambda)\|_{p,q}^q$ and that h is smooth around $\xi(\lambda)$. By noting that $\xi(\lambda)$ satisfies

(30) and (31) (the first- and second- derivative of h at $\xi(\lambda)$), one has that h satisfies the second-order growth condition at $\xi(\lambda)$, that is, there exist $\epsilon_\lambda > 0$ and $\delta_\lambda > 0$ such that

$$h(z) \geq h(\xi(\lambda)) + 2\epsilon_\lambda \|z - \xi(\lambda)\|_2^2 \quad \text{for any } z \in \mathbf{B}(\xi(\lambda), \delta_\lambda). \tag{32}$$

In what follows, let $\epsilon_\lambda > 0$ and $\delta_\lambda > 0$ be given as above, and select $\epsilon_0 > 0$ such that

$$\sqrt{\epsilon_\lambda \epsilon_0} - \|B\|\|D\| > 0. \tag{33}$$

According to Lemma 10 (with $\|D\cdot\|_2^2 + 2\langle B\xi(\lambda) - b, D\cdot\rangle - 2\epsilon_0\|\cdot\|_2^2$ in place of f), there exists $\delta_0 > 0$ such that

$$\|Dy\|_2^2 + 2\langle B\xi(\lambda) - b, Dy\rangle - 2\epsilon_0\|y\|_2^2 + \lambda\|y\|_{p,q}^q \geq 0 \quad \text{for any } y \in \mathbf{B}(0, \delta_0). \tag{34}$$

Thus, for each $x := (z, y) \in \mathbf{B}(\xi(\lambda), \delta_\lambda) \times \mathbf{B}(0, \delta_0)$, it follows that

$$\begin{aligned}
&\|Ax - b\|_2^2 + \lambda\|x\|_{p,q}^q \\
&= \|Bz - b + Dy\|_2^2 + \lambda\|z\|_{p,q}^q + \lambda\|y\|_{p,q}^q \\
&= \|Bz - b\|_2^2 + \lambda\|z\|_{p,q}^q + \|Dy\|_2^2 + 2\langle Bz - b, Dy\rangle + \lambda\|y\|_{p,q}^q \\
&= h(z) + \|Dy\|_2^2 + 2\langle B\xi(\lambda) - b, Dy\rangle + \lambda\|y\|_{p,q}^q + 2\langle B(z - \xi(\lambda)), Dy\rangle.
\end{aligned}$$

By (32) and (34), it yields that

$$\begin{aligned}
&\|Ax - b\|_2^2 + \lambda\|x\|_{p,q}^q \\
&\geq h(\xi(\lambda)) + 2\epsilon_\lambda\|z - \xi(\lambda)\|_2^2 + 2\epsilon_0\|y\|_2^2 + 2\langle B(z - \xi(\lambda)), Dy\rangle \\
&\geq h(\xi(\lambda)) + 4\sqrt{\epsilon_\lambda\epsilon_0}\|z - \xi(\lambda)\|_2\|y\|_2 - 2\|B\|\|D\|\|z - \xi(\lambda)\|_2\|y\|_2 \\
&= \|Ax^*(\lambda) - b\|_2^2 + \lambda\|x^*(\lambda)\|_{p,q}^q + 2(2\sqrt{\epsilon_\lambda\epsilon_0} - \|B\|\|D\|)\|z - \xi(\lambda)\|_2\|y\|_2 \\
&\geq \|Ax^*(\lambda) - b\|_2^2 + \lambda\|x^*(\lambda)\|_{p,q}^q,
\end{aligned}$$

where the last inequality follows from (33). Hence $x^*(\lambda)$ is a local minimum of problem (4), and (b) is verified.

Finally, we check (c) by providing an upper bound on the distance from $\xi(\lambda)$ to $\bar{z}$. By (30), one has that

$$\xi(\lambda) - \bar{z} = -\frac{\lambda q}{2}((B^\top B)^{-1}) \begin{pmatrix} \|\xi(\lambda)_{\mathcal{G}_1}\|_p^{q-p}\sigma(\xi(\lambda)_{\mathcal{G}_1}) \\ \vdots \\ \|\xi(\lambda)_{\mathcal{G}_S}\|_p^{q-p}\sigma(\xi(\lambda)_{\mathcal{G}_S}) \end{pmatrix}. \tag{35}$$

Noting that $\{\xi(\lambda) : \lambda \in (-\kappa, \kappa)\} \subseteq \mathbf{B}(\bar{z}, \bar{\delta})$, without loss of generality, we assume for any $\lambda < \kappa$ that

$$\|\xi(\lambda)_{\mathcal{G}_i}\|_p^{2(q-p)} \leq 2\|\bar{z}_{\mathcal{G}_i}\|_p^{2(q-p)} \quad \text{and} \quad \|\xi(\lambda)_{\mathcal{G}_i}\|_{2p-2}^{2p-2} \leq 2\|\bar{z}_{\mathcal{G}_i}\|_{2p-2}^{2p-2} \quad \text{for } i = 1, \ldots, S$$

(otherwise, we choose a smaller $\bar{\delta}$). Recall that $\sigma(\xi(\lambda)_{\mathcal{G}_i}) = \text{vector}\left(|\xi(\lambda)_j|^{p-1}\text{sign}(\xi(\lambda)_j)\right)_{\mathcal{G}_i}$. We obtain from (35) that

$$\begin{aligned}
\|\xi(\lambda) - \bar{z}\|_2^2 &\leq \tfrac{\lambda^2 q^2}{4}\|(B^\top B)^{-1}\|^2 \textstyle\sum_{i=1}^S \left(\|\xi(\lambda)_{\mathcal{G}_i}\|_p^{2(q-p)} \sum_{j\in\mathcal{G}_i} |\xi(\lambda)_j|^{2p-2}\right) \\
&= \tfrac{\lambda^2 q^2}{4}\|(B^\top B)^{-1}\|^2 \textstyle\sum_{i=1}^S \left(\|\xi(\lambda)_{\mathcal{G}_i}\|_p^{2(q-p)}\|\xi(\lambda)_{\mathcal{G}_i}\|_{2p-2}^{2p-2}\right) \\
&\leq \tfrac{\lambda^2 q^2}{4}\|(B^\top B)^{-1}\|^2 S \max_{i=1,\ldots,S}\left(\|\xi(\lambda)_{\mathcal{G}_i}\|_p^{2(q-p)}\|\xi(\lambda)_{\mathcal{G}_i}\|_{2p-2}^{2p-2}\right) \\
&\leq \lambda^2 q^2 S\|(B^\top B)^{-1}\|^2 \max_{i=1,\ldots,S}\left(\|\bar{z}_{\mathcal{G}_i}\|_p^{2(q-p)}\|\bar{z}_{\mathcal{G}_i}\|_{2p-2}^{2p-2}\right).
\end{aligned}$$

Hence we arrive at that

$$\|x^*(\lambda) - \bar{x}\|_2^2 = \|\xi(\lambda) - \bar{z}\|_2^2 \leq \lambda^2 q^2 S \|(B^\top B)^{-1}\|^2 \max_{\bar{x}_{\mathcal{G}_i} \neq 0} \left(\|\bar{x}_{\mathcal{G}_i}\|_p^{2(q-p)} \|\bar{x}_{\mathcal{G}_i}\|_{2p-2}^{2p-2} \right)$$

for any $\lambda < \kappa$, and the proof is complete. ■

Theorem 11 provides a uniform local recovery bound for all the $\ell_{p,q}$ regularization problems ($0 < q < 1 \leq p$), which is

$$\|x_{p,q}^*(\lambda) - \bar{x}\|_2^2 \leq \mathcal{O}(\lambda^2 S),$$

where $x_{p,q}^*(\lambda)$ is a local optimal solution of problem (4) (near $\bar{x}$). This bound improves the global recovery bound given in Theorem 9 (of order $\mathcal{O}(\lambda^{\frac{2}{2-q}})$) and shares the same one with the $\ell_{p,1}$ regularization problem (group Lasso); see Blumensath and Davies (2008); van de Geer and Bühlmann (2009). It is worth noting that our proof technique is not working when $q = 1$ as Lemma 10 fails in this case.

3. Proximal Gradient Method for Group Sparse Optimization

Many efficient algorithms have been proposed to solve sparse optimization problems, and one of the most popular optimization algorithms is the proximal gradient method (PGM); see Beck and Teboulle (2009); Combettes and Wajs (2005); Xiao and Zhang (2013) and references therein. It was reported in Combettes and Wajs (2005) that the PGM for solving the ℓ_1 regularization problem (1) reduces to the well-known iterative soft thresholding algorithm (ISTA), and that the ISTA has a local linear convergence rate under some assumptions; see Bredies and Lorenz (2008); Hale et al. (2008); Tao et al. (2016). Recently, the global convergence of the PGM for solving some types of nonconvex regularization problems have been studied under the framework of the Kurdyka-Łojasiewicz theory (Attouch et al., 2010; Bolte et al., 2013), the majorization-minimization scheme (Mairal, 2013), the coordinate gradient descent method (Tseng and Yun, 2009), the general iterative shrinkage and thresholding (Gong et al., 2013) and the successive upper-bound minimization approach (Razaviyayn et al., 2013).

In this section, we apply the PGM to solve the group sparse optimization problem (4) (PGM-GSO), which is stated as follows.

Algorithm 1 (PGM-GSO) *Select a 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*

$$z^k = x^k - 2vA^\top(Ax^k - b), \tag{36}$$

$$x^{k+1} \in \arg\min_{x \in \mathbb{R}^n} \left\{ \lambda \|x\|_{p,q}^q + \frac{1}{2v} \|x - z^k\|_2^2 \right\}. \tag{37}$$

Global convergence of the PGM-GSO falls in the framework of the Kurdyka-Łojasiewicz theory (see Attouch et al., 2010). In particular, following from Bolte et al. (2013, prop. 3), the sequence generated by the PGM-GSO converges to a critical point, especially a global minimum when $q \geq 1$ and a local minimum when $q = 0$ (inspired by the idea in Blumensath and Davies, 2008), as summarized as follows.

Theorem 12 *Let $p \geq 1$. Suppose that the sequence $\{x^k\}$ is generated by the PGM-GSO with $v < \frac{1}{2}\|A\|_2^{-2}$. Then the following statements hold:*

(i) *if $q \geq 1$, then $\{x^k\}$ converges to a global minimum of problem (4),*

(ii) *if $q = 0$, then $\{x^k\}$ converges to a local minimum of problem (4), and*

(iii) *if $0 < q < 1$, then $\{x^k\}$ converges to a critical point [1] of problem (4).*

Although the global convergence of the PGM-GSO has been provided in Theorem 12, some important issues of the PGM-GSO have not been discovered yet. The section is to continue the development of the PGM-GSO, concentrating on its efficiency and applicability. In particular, we will establish the local convergence rate of the PGM-GSO under some mild conditions, and derive the analytical solutions of subproblem (37) for some specific p and q.

3.1 Local Linear Convergence Rate

In this subsection, we establish the local linear convergence rate of the PGM-GSO for the case when $p = 1$ and $0 < q < 1$. For the reminder of this subsection, we always assume that $p = 1$ and $0 < q < 1$.

To begin with, by virtue of the second-order necessary condition of subproblem (37), the following lemma provides a lower bound for nonzero groups of sequence $\{x^k\}$ generated by the PGM-GSO and shows that the index set of nonzero groups of $\{x^k\}$ maintains constant for large k.

Lemma 13 *Let $K = (v\lambda q(1-q))^{\frac{1}{2-q}}$, and let $\{x^k\}$ be a sequence generated by the PGM-GSO with $v < \frac{1}{2}\|A\|_2^{-2}$. Then the following statements hold:*

(i) *For any i and k, if $x^k_{\mathcal{G}_i} \neq 0$, then $\|x^k_{\mathcal{G}_i}\|_1 \geq K$.*

(ii) *x^k shares the same index set of nonzero groups for large k, that is, there exist $N \in \mathbb{N}$ and $\mathcal{I} \subseteq \{1, \ldots, r\}$ such that*

$$\begin{cases} x^k_{\mathcal{G}_i} \neq 0, & i \in \mathcal{I}, \\ x^k_{\mathcal{G}_i} = 0, & i \notin \mathcal{I}, \end{cases} \quad \text{for all } k \geq N.$$

Proof (i) For each group $x^k_{\mathcal{G}_i}$, by (37), one has that

$$x^k_{\mathcal{G}_i} \in \underset{x \in \mathbb{R}^{n_i}}{\arg\min} \left\{ \lambda \|x\|_1^q + \frac{1}{2v}\|x - z^{k-1}_{\mathcal{G}_i}\|_2^2 \right\}. \tag{38}$$

If $x^k_{\mathcal{G}_i} \neq 0$, we define $\mathcal{A}^k_i := \{j \in \mathcal{G}_i : x^k_j \neq 0\}$ and $a^k_i := |\mathcal{A}^k_i|$. Without loss of generality, we assume that the first a^k_i components of $x^k_{\mathcal{G}_i}$ are nonzeros. Then (38) implies that

$$x^k_{\mathcal{G}_i} \in \underset{x \in \mathbb{R}^{a^k_i} \times \{0\}}{\arg\min} \left\{ \lambda \|x\|_1^q + \frac{1}{2v}\|x - z^{k-1}_{\mathcal{G}_i}\|_2^2 \right\}.$$

1. A point x is said to be a critical point of F if 0 belongs to its limiting subdifferential at x; see Mordukhovich (2006).

Its second-order necessary condition says that

$$\frac{1}{v}I_i^k + \lambda q(q-1)M_i^k \succeq 0,$$

where I_i^k is the identity matrix in $R^{a_i^k \times a_i^k}$ and $M_i^k = \|x^k_{\mathcal{A}_i^k}\|_1^{q-2}(\mathrm{sign}(x^k_{\mathcal{A}_i^k}))(\mathrm{sign}(x^k_{\mathcal{A}_i^k}))^\top$. Let e be the first column of I_i^k. Then, we obtain that

$$\frac{1}{v}e^\top I_i^k e + \lambda q(q-1)e^\top M_i^k e \geq 0,$$

that is,

$$\frac{1}{v} + \lambda q(q-1)\|x^k_{\mathcal{A}_i^k}\|_1^{q-2} \geq 0.$$

Consequently, it implies that

$$\|x^k_{\mathcal{G}_i}\|_1 = \|x^k_{\mathcal{A}_i^k}\|_1 \geq (v\lambda q(1-q))^{\frac{1}{2-q}} = K.$$

Hence, it completes the proof of (i).
(ii) Recall from Theorem 12 that $\{x^k\}$ converges to a critical point x^*. Then there exists $N \in \mathbb{N}$ such that $\|x^k - x^*\|_2 < \frac{K}{2\sqrt{n}}$, and thus,

$$\|x^{k+1} - x^k\|_2 \leq \|x^{k+1} - x^*\|_2 + \|x^k - x^*\|_2 < \frac{K}{\sqrt{n}}, \tag{39}$$

for any $k \geq N$. Proving by contradiction, without loss of generality, we assume that there exist $k \geq N$ and $i \in \{1, \ldots, r\}$ such that $x^{k+1}_{\mathcal{G}_i} \neq 0$ and $x^k_{\mathcal{G}_i} = 0$. Then it follows from (i) that

$$\|x^{k+1} - x^k\|_2 \geq \frac{1}{\sqrt{n}}\|x^{k+1} - x^k\|_1 \geq \frac{1}{\sqrt{n}}\|x^{k+1}_{\mathcal{G}_i} - x^k_{\mathcal{G}_i}\|_1 \geq \frac{K}{\sqrt{n}},$$

which yields a contradiction with (39). The proof is complete. ■

Let $x^* \in \mathbb{R}^n$, and let

$$\mathcal{S} := \big\{i \in \{1, \ldots, r\} : x^*_{\mathcal{G}_i} \neq 0\big\} \quad \text{and} \quad B := (A_{\cdot j})_{j \in \mathcal{G}_\mathcal{S}}.$$

Consider the following restricted problem

$$\min_{y \in \mathbb{R}^{n_\mathbf{a}}} \quad f(y) + \varphi(y), \tag{40}$$

where $n_\mathbf{a} := \sum_{i \in \mathcal{S}} n_i$, and

$$f : \mathbb{R}^{n_\mathbf{a}} \to \mathbb{R} \quad \text{with} \quad f(y) := \|By - b\|_2^2 \quad \text{for any } y \in \mathbb{R}^{n_\mathbf{a}},$$

$$\varphi : \mathbb{R}^{n_\mathbf{a}} \to \mathbb{R} \quad \text{with} \quad \varphi(y) := \lambda\|y\|_{1,q}^q \quad \text{for any } y \in \mathbb{R}^{n_\mathbf{a}}.$$

The following lemma provides the first- and second-order conditions for a local minimum of the $\ell_{1,q}$ regularization problem, and shows a second-order growth condition for the restricted problem (40), which is useful for establishing the local linear convergence rate of the PGM-GSO.

Lemma 14 *Assume that x^* is a local minimum of problem (4). Suppose that any nonzero group of x^* is active, and the columns of B are linearly independent* [2]*. Then the following statements are true:*

(i) *The following first- and second-order conditions hold*

$$2B^\top(By^* - b) + \lambda q \begin{pmatrix} \|y^*_{\mathcal{G}_1}\|_1^{q-1}\mathrm{sign}(y^*_{\mathcal{G}_1}) \\ \vdots \\ \|y^*_{\mathcal{G}_S}\|_1^{q-1}\mathrm{sign}(y^*_{\mathcal{G}_S}) \end{pmatrix} = 0, \tag{41}$$

and

$$2B^\top B + \lambda q(q-1) \begin{pmatrix} M_1^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_S^* \end{pmatrix} \succ 0, \tag{42}$$

where

$$M_i^* = \|y^*_{\mathcal{G}_i}\|_1^{q-2} \left(\mathrm{sign}(y^*_{\mathcal{G}_i})\right)\left(\mathrm{sign}(y^*_{\mathcal{G}_i})\right)^\top.$$

(ii) *The second-order growth condition holds at y^* for problem (40), that is, there exist $\varepsilon > 0$ and $\delta > 0$ such that*

$$(f+\varphi)(y) \geq (f+\varphi)(y^*) + \varepsilon\|y - y^*\|_2^2 \quad \textit{for any } y \in \mathbf{B}(y^*, \delta). \tag{43}$$

Proof Without loss of generality, we assume that $\mathcal{S} := \{1, \ldots, S\}$. By assumption, x^* is of structure $x^* := (y^{*\top}, 0)^\top$ with

$$y^* := (x^{*\top}_{\mathcal{G}_1}, \ldots, x^{*\top}_{\mathcal{G}_S})^\top \text{ and } x^*_{\mathcal{G}_i} \neq_{\mathbf{a}} 0 \text{ for } i = 1, \ldots, S. \tag{44}$$

(i) By (44), one has that $\varphi(\cdot)$ is smooth around y^* with its first- and second-derivatives being

$$\varphi'(y^*) = \lambda q \begin{pmatrix} \|y^*_{\mathcal{G}_1}\|_1^{q-1}\mathrm{sign}(y^*_{\mathcal{G}_1}) \\ \vdots \\ \|y^*_{\mathcal{G}_S}\|_1^{q-1}\mathrm{sign}(y^*_{\mathcal{G}_S}) \end{pmatrix},$$

and

$$\varphi''(y^*) = \lambda q(q-1) \begin{pmatrix} M_1^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_S^* \end{pmatrix};$$

hence $(f+\varphi)(\cdot)$ is also smooth around y^*. Therefore, we obtain the following first- and second-order necessary conditions of problem (40)

$$f'(y^*) + \varphi'(y^*) = 0 \quad \text{and} \quad f''(y^*) + \varphi''(y^*) \succeq 0,$$

2. This assumption is mild, and it holds automatically for the case when $n_{\max} = 1$ (see Chen et al., 2010, thm. 2.1).

which are (41) and

$$2B^\top B + \lambda q(q-1)\begin{pmatrix} M_1^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_S^* \end{pmatrix} \succeq 0,$$

respectively. Proving by contradiction, we assume that (42) does not hold, that is, there exists some $w \neq 0$ such that

$$2w^\top B^\top Bw + \lambda q(q-1)\sum_{i=1}^{S}\left(\|y^*_{\mathcal{G}_i}\|_1^{q-2}\cdot\left(\sum_{j\in\mathcal{G}_i} w_j \mathrm{sign}(y_j^*)\right)^2\right) = 0. \tag{45}$$

By assumption, one has that $B^\top B \succ 0$, and thus it follows from (45) that

$$\left(\sum_{j\in\mathcal{G}_i} w_j \mathrm{sign}(y_j^*)\right)^2 > 0 \quad \text{for some } i \in \{1,\ldots,S\}. \tag{46}$$

Let $h:\mathbb{R}\to\mathbb{R}$ with $h(t) := \|B(y^*+tw)-b\|_2^2 + \lambda\|y^*+tw\|_p^p$ for any $t\in\mathbb{R}$. Clearly, $h(\cdot)$ has a local minimum at 0, and $h(\cdot)$ is smooth around 0 with its derivatives being

$$h'(0) = 2w^\top B^\top(By^*-b) + \lambda q\sum_{i=1}^{S}\left(\|y^*_{\mathcal{G}_i}\|_1^{q-1}\cdot\sum_{j\in\mathcal{G}_i} w_j\mathrm{sign}(y_j^*)\right) = 0,$$

$$h''(0) = 2w^\top B^\top Bw + \lambda q(q-1)\sum_{i=1}^{S}\left(\|y^*_{\mathcal{G}_i}\|_1^{q-2}\cdot\left(\sum_{j\in\mathcal{G}_i} w_j\mathrm{sign}(y_j^*)\right)^2\right) = 0 \text{ (by (45))},$$

$$h^{(3)}(0) = \lambda q(q-1)(q-2)\sum_{i=1}^{S}\left(\|y^*_{\mathcal{G}_i}\|_1^{q-3}\cdot\left(\sum_{j\in\mathcal{G}_i} w_j\mathrm{sign}(y_j^*)\right)^3\right) = 0,$$

$$h^{(4)}(0) = \lambda q(q-1)(q-2)(q-3)\sum_{i=1}^{S}\left(\|y^*_{\mathcal{G}_i}\|_1^{q-4}\cdot\left(\sum_{j\in\mathcal{G}_i} w_j\mathrm{sign}(y_j^*)\right)^4\right) < 0; \tag{47}$$

due to (46). However, by elementary of calculus, it is clear that $h^{(4)}(0)$ must be nonnegative (since $h(\cdot)$ obtains a local minimum at 0), which yields a contradiction to (47). Therefore, we proved (42).

(ii) By the structure of y^* (44), $\varphi(\cdot)$ is smooth around y^*, and thus, $(f+\varphi)(\cdot)$ is also smooth around y^* with its derivatives being

$$f'(y^*) + \varphi'(y^*) = 0 \quad \text{and} \quad f''(y^*) + \varphi''(y^*) \succ 0;$$

due to (41) and (42). Hence (43) follows from Rockafellar and Wets (1998, thm. 13.24). This completes the proof. ■

The key for the study of local convergence rate of the PGM-GSO is the descent property of the function $f+\varphi$ in each iteration step. The following lemma states some basic properties of active groups of sequence $\{x^k\}$ generated by the PGM-GSO.

Lemma 15 *Let $\{x^k\}$ be a sequence generated by the PGM-GSO with $v < \frac{1}{2}\|A\|_2^{-2}$, which converges to x^* (by Theorem 12). Assume that the assumptions in Lemma 14 are satisfied. We define*

$$\alpha := \|B\|_2^2, \quad L := 2\|A\|_2^2 \quad \text{and} \quad D_k := \varphi(y^k) - \varphi(y^{k+1}) + \langle f'(y^k), y^k - y^{k+1}\rangle.$$

Then there exist $\delta > 0$ and $N \in \mathbb{N}$ such that the following inequalities hold for any $k \geq N$:

$$D_k \geq \left(\frac{1}{v} - \alpha\right) \|y^k - y^{k+1}\|_2^2, \tag{48}$$

and

$$(f+\varphi)(y^{k+1}) \leq (f+\varphi)(y^k) - \left(1 - \frac{Lv}{2(1-v\alpha)}\right) D_k. \tag{49}$$

Proof By Lemma 13(ii) and the fact that $\{x^k\}$ converges to x^*, one has that x^k shares the same index set of nonzero groups with that of x^* for large k; further by the structure of y^* (44), we obtain that all components in nonzero groups of y^k are nonzero for large k. In another word, we have

$$\text{there exists } N \in \mathbb{N} \text{ such that } y^k \neq_{\mathbf{a}} 0 \text{ and } x^k_{\mathcal{G}_{S^c}} = 0 \text{ for any } k \geq N; \tag{50}$$

hence $\varphi(\cdot)$ is smooth around y^k for any $k \geq N$.

In view of PGM-GSO and the decomposition of $x = \left(y^\top, z^\top\right)^\top$, one has that

$$y^{k+1} \in \arg\min \left\{ \varphi(y) + \frac{1}{2v} \left\| y - \left(y^k - v f'(y^k)\right)\right\|_2^2 \right\}.$$

Its first-order necessary condition is

$$\varphi'(y^{k+1}) = \frac{1}{v}\left(y^k - v f'(y^k) - y^{k+1}\right). \tag{51}$$

Recall from (42) that $\varphi''(y^*) \succ -2B^\top B$. Since $\varphi(\cdot)$ is smooth around y^*, then there exists $\delta > 0$ such that $\varphi''(w) \succ -2B^\top B$ for any $w \in \mathbf{B}(y^*, \delta)$. Noting that $\{y^k\}$ converges to y^*, without loss of generality, we assume that $\|y^k - y^*\| < \delta$ for any $k \geq N$ (otherwise, we can choose a larger N). Therefore, one has that $\varphi''(y^k) \succ -2B^\top B$ for any $k \geq N$. Then by Taylor expansion, we can assume without loss of generality that the following inequality holds for any $k \geq N$ and any $w \in \mathbf{B}(y^*, \delta)$ (otherwise, we can choose a smaller δ):

$$\varphi(w) > \varphi(y^{k+1}) + \langle \varphi'(y^{k+1}), w - y^{k+1}\rangle - \alpha\|w - y^{k+1}\|_2^2.$$

Hence, by (51), it follows that

$$\varphi(w) - \varphi(y^{k+1}) > \frac{1}{v}\langle y^k - v f'(y^k) - y^{k+1}, w - y^{k+1}\rangle - \alpha\|w - y^{k+1}\|_2^2. \tag{52}$$

Then (48) follows by setting $w = y^k$. Furthermore, by the definition of $f(\cdot)$, it is of class $C_L^{1,1}$ and it follows from Bertsekas (1999, prop. A.24) that

$$\|f(y) - f(x) - f'(x)(y-x)\| \leq \frac{L}{2}\|y - x\|^2 \quad \text{for any } x, y.$$

Then, by the definition of D_k, it follows that

$$\begin{array}{ll}(f+\varphi)(y^{k+1})-(f+\varphi)(y^k)+D_k & = f(y^{k+1})-f(y^k)+\langle f'(y^k), y^k-y^{k+1}\rangle \\ & \le \frac{L}{2}\|y^k-y^{k+1}\|_2^2 \\ & \le \frac{Lv}{2(1-v\alpha)}D_k,\end{array}$$

where the last inequality follows from (48), and thus, (49) is proved. ■

The main result of this subsection is presented as follows, in which we establish the local linear convergence rate of the PGM-GSO to a local minimum for the case when $p = 1$ and $0 < q < 1$ under some mild assumptions.

Theorem 16 *Let $\{x^k\}$ be a sequence generated by the PGM-GSO with $v < \frac{1}{2}\|A\|_2^{-2}$. Then $\{x^k\}$ converges to a critical point x^* of problem (4). Assume that x^* is a local minimum of problem (4). Suppose that any nonzero group of x^* is active, and the columns of B are linearly independent. Then there exist $N \in \mathbb{N}$, $C > 0$ and $\eta \in (0, 1)$ such that*

$$F(x^k) - F(x^*) \le C\eta^k \quad \text{and} \quad \|x^k - x^*\|_2 \le C\eta^k \quad \text{for any } k \ge N. \tag{53}$$

Proof The convergence of $\{x^k\}$ to a critical point x^* of problem (4) directly follows from Theorem 12. Let D_k, N and δ be defined as in Lemma 15, and let

$$r_k := F(x^k) - F(x^*).$$

Note in (50) that $y^k \neq_{\mathbf{a}} 0$ and $x^k_{\mathcal{G}^c_S} = 0$ for any $k \ge N$. Thus

$$r_k = (f+\varphi)(y^k) - (f+\varphi)(y^*) \quad \text{for any } k \ge N.$$

It is trivial to see that $\varphi(\cdot)$ is smooth around y^* (as it is active) and that

$$\varphi''(y^*) = \lambda q(q-1)\begin{pmatrix} M_1^* & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & M_S^* \end{pmatrix} \prec 0, \quad f''(y^*) + \varphi''(y^*) \succ 0;$$

as shown in (42). This shows that $\varphi(\cdot)$ is concave around y^*, while $(f+\varphi)(\cdot)$ is convex around y^*. Without loss of generality, we assume that $\varphi(\cdot)$ is concave and $(f+\varphi)(\cdot)$ is convex in $\mathbf{B}(y^*, \delta)$ and that $y^k \in \mathbf{B}(y^*, \delta)$ for any $k \ge N$ (since $\{y^k\}$ converges to y^*).

By the convexity of $(f+\varphi)(\cdot)$ in $\mathbf{B}(y^*, \delta)$, it follows that for any $k \ge N$

$$\begin{array}{ll} r_k & = (f+\varphi)(y^k) - (f+\varphi)(y^*) \\ & \le \langle f'(y^k) + \varphi'(y^k), y^k - y^*\rangle \\ & = \langle f'(y^k) + \varphi'(y^k), y^k - y^{k+1}\rangle + \langle f'(y^k) + \varphi'(y^k), y^{k+1} - y^*\rangle \\ & = D_k - \varphi(y^k) + \varphi(y^{k+1}) + \langle \varphi'(y^k), y^k - y^{k+1}\rangle + \langle f'(y^k) + \varphi'(y^k), y^{k+1} - y^*\rangle. \end{array} \tag{54}$$

Noting that $\varphi(\cdot)$ is concave in $\mathbf{B}(y^*, \delta)$, it follows that

$$\varphi(y^k) - \varphi(y^{k+1}) \ge \langle \varphi'(y^k), y^k - y^{k+1}\rangle.$$

Consequently, (54) is reduced to

$$\begin{aligned} r_k &\le D_k + \langle f'(y^k) + \varphi'(y^k), y^{k+1} - y^* \rangle \\ &= D_k + \langle \varphi'(y^k) - \varphi'(y^{k+1}), y^{k+1} - y^* \rangle + \langle f'(y^k) + \varphi'(y^{k+1}), y^{k+1} - y^* \rangle \\ &\le D_k + \left(L_\varphi + \tfrac{1}{v}\right) \|y^k - y^{k+1}\|_2 \|y^{k+1} - y^*\|_2, \end{aligned} \tag{55}$$

where the last inequality follows from the smoothness of φ on $\mathbf{B}(y^*, \delta)$ and (51), and L_φ is the Lipschitz constant of $\varphi'(\cdot)$ on $\mathbf{B}(y^*, \delta)$. Let $\beta := 1 - \frac{Lv}{2(1-v\alpha)} \in (0,1)$ (due to the assumption $v < \frac{1}{L}$). Then, (49) is reduced to

$$r_k - r_{k+1} = (f+\varphi)(y^k) - (f+\varphi)(y^{k+1}) \ge \beta D_k > 0,$$

and thus, it follows from (55) and (48) that

$$\begin{aligned} \beta r_k &\le \beta D_k + \beta \left(L_\varphi + \tfrac{1}{v}\right) \|y^k - y^{k+1}\|_2 \|y^{k+1} - y^*\|_2 \\ &\le r_k - r_{k+1} + \beta \left(L_\varphi + \tfrac{1}{v}\right) \|y^{k+1} - y^*\|_2 \sqrt{\tfrac{v}{1-v\alpha} D_k} \\ &\le r_k - r_{k+1} + \left(L_\varphi + \tfrac{1}{v}\right) \sqrt{\tfrac{v\beta}{1-v\alpha}} \|y^{k+1} - y^*\|_2 \sqrt{r_k - r_{k+1}}. \end{aligned} \tag{56}$$

Recall from Lemma 14(ii) that there exists $c > 0$ such that

$$\|y - y^*\|_2^2 \le c\,((f+\varphi)(y) - (f+\varphi)(y^*)) \quad \text{for any } y \in \mathbf{B}(y^*, \delta).$$

Thus, it follows that

$$\|y^{k+1} - y^*\|_2^2 \le c r_{k+1} \le c r_k \quad \text{for any } k \ge N. \tag{57}$$

Let $\epsilon := \frac{c}{\beta}\left(L_\varphi + \frac{1}{v}\right)^2$. By Young's inequality, (56) yields that

$$\begin{aligned} \beta r_k &\le r_k - r_{k+1} + \tfrac{1}{2\epsilon} \|y^{k+1} - y^*\|_2^2 \left(L_\varphi + \tfrac{1}{v}\right)^2 + \tfrac{\epsilon v \beta}{2(1-v\alpha)} (r_k - r_{k+1}) \\ &\le r_k - r_{k+1} + \tfrac{\beta}{2} r_k + \tfrac{cv}{2(1-v\alpha)} \left(L_\varphi + \tfrac{1}{v}\right)^2 (r_k - r_{k+1}). \end{aligned} \tag{58}$$

Let $\gamma := \frac{cv}{2(1-v\alpha)}\left(L_\varphi + \frac{1}{v}\right)^2 > 0$. Then, (58) is reduced to

$$r_{k+1} \le \frac{1 + \gamma - \frac{\beta}{2}}{1+\gamma} r_k = \eta_1 r_k,$$

where $\eta_1 := \frac{1+\gamma-\frac{\beta}{2}}{1+\gamma} \in (0,1)$. Thus, by letting $C_1 := r_N \eta_1^{-N}$, it follows that

$$r_k \le \eta_1^{k-N} r_N = C_1 \eta_1^k \quad \text{for any } k \ge N.$$

By letting $\eta_2 = \sqrt{\eta_1}$ and $C_2 = \sqrt{cC_1}$, it follows from (57) that

$$\|x^k - x^*\|_2 = \|y^k - y^*\|_2 \le (cr_k)^{1/2} \le C_2 \eta_2^k \quad \text{for any } k \ge N.$$

Letting $C := \max\{C_1, C_2\}$ and $\eta := \max\{\eta_1, \eta_2\}$, we obtain (53). The proof is complete. ■

Theorem 16 establishes the linear convergence rate of the PGM for solving the $\ell_{1,q}$ regularization problem under two assumptions: (i) the critical point x^* of the sequence produced by the PGM is a local minimum of problem (4), and (ii) any nonzero group of the local minimum is an active one. The assumption (i) is important by which we are able to establish a second-order growth property, which plays a crucial role in our analysis. Note that the assumption (ii) is satisfied automatically for the sparse optimization problem ($n_{max} = 1$). Hence, when $n_{max} = 1$, we obtain the linear convergence rate of the PGM for solving ℓ_q regularization problem ($0 < q < 1$). This result is stated below as a corollary.

Corollary 17 *Let $0 < q < 1$, and let $\{x^k\}$ be a sequence generated by the PGM for solving the following ℓ_q regularization problem*

$$\min_{x\in\mathbb{R}^n} F(x) := \|Ax - b\|_2^2 + \lambda\|x\|_q^q \tag{59}$$

with $v < \frac{1}{2}\|A\|_2^{-2}$. Then $\{x^k\}$ converges to a critical point x^* of problem (59). Further assume that x^* is a local minimum of problem (59). Then there exist $N \in \mathbb{N}$, $C > 0$ and $\eta \in (0, 1)$ such that

$$F(x^k) - F(x^*) \le C\eta^k \quad \text{and} \quad \|x^k - x^*\|_2 \le C\eta^k \quad \text{for any } k \ge N.$$

While we are carrying out the revision of our manuscript, we have found that the local linear convergence rate of the PGM has been studied in the literature. On one hand, the local linear convergence rate of the PGM for solving the ℓ_1 regularization problem (ISTA) has been established under some assumptions in Bredies and Lorenz (2008); Hale et al. (2008); Tao et al. (2016), and, under the framework of the so-called KL theory, it is established that the sequence generated by the PGM linearly converges to a critical point of a KL function if its KL exponent is in $(0, \frac{1}{2}]$; see Frankel et al. (2015); Li and Pong (2016); Xu and Yin (2013). However, the KL exponent of the ℓ_q regularized function is still unknown, and thus the linear convergence result in these references cannot directly be applied to the ℓ_q regularization problem. On the other hand, Zeng et al. (2015) has obtained the linear convergence rate of the ℓ_q regularization problem under the framework of a restricted KL property. However, it seems that their result is restrictive as it is assumed that the stepsize v and the regularization component q satisfy

$$\frac{q}{2} < \frac{\lambda_{\min}(A_{\mathcal{S}}^T A_{\mathcal{S}})}{\|A\|_2^2} \quad \text{and} \quad \frac{q}{4\lambda_{\min}(A_{\mathcal{S}}^T A_{\mathcal{S}})} < v < \frac{1}{2\|A\|_2^2},$$

where $\mathcal{S}$ is the active index of the limiting point x^*, while our result in Corollary 17 holds for all the stepsize $v < \frac{1}{2}\|A\|_2^{-2}$ and the regularization component $0 < q < 1$.

3.2 Analytical Solutions of Proximal Subproblems

Since the main computation of the PGM is the proximal step (37), it is significant to investigate the solutions of subproblem (37) for the specific applications so as to spread the application of the PGM. Note that $\|x\|_{p,q}^q$ and $\|x - z^k\|_2^2$ are both grouped separable. Then

the proximal step (37) can be achieved parallelly in each group, and is equivalent to solve a cycle of low dimensional proximal optimization subproblems

$$x_{\mathcal{G}_i}^{k+1} \in \arg\min_{x\in\mathbb{R}^{n_i}} \left\{ \lambda\|x_{\mathcal{G}_i}\|_p^q + \frac{1}{2v}\|x_{\mathcal{G}_i} - z_{\mathcal{G}_i}^k\|_2^2 \right\} \quad \text{for } i = 1, \cdots, r. \tag{60}$$

When p and q are given as some specific numbers, such as $p = 1, 2$ and $q = 0, 1/2, 2/3, 1$, the solution of subproblem (60) of each group can be given explicitly by an analytical formula, as shown in the following proposition.

Proposition 18 *Let $z \in \mathbb{R}^l$, $v > 0$ and the proximal regularization be*

$$Q_{p,q}(x) := \lambda\|x\|_p^q + \frac{1}{2v}\|x - z\|_2^2 \quad \textit{for any } x \in \mathbb{R}^l.$$

Then the proximal operator

$$P_{p,q}(z) \in \arg\min_{x\in\mathbb{R}^l} \{Q_{p,q}(x)\}$$

has the following analytical formula:

(i) *if $p = 2$ and $q = 1$, then*

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

(ii) *if $p = 2$ and $q = 0$, then*

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

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

$$P_{2,1/2}(z) = \begin{cases} \frac{16\|z\|_2^{3/2}\cos^3\left(\frac{\pi}{3} - \frac{\psi(z)}{3}\right)}{3\sqrt{3}v\lambda + 16\|z\|_2^{3/2}\cos^3\left(\frac{\pi}{3} - \frac{\psi(z)}{3}\right)} z, & \|z\|_2 > \frac{3}{2}(v\lambda)^{2/3}, \\ 0 \text{ or } \frac{16\|z\|_2^{3/2}\cos^3\left(\frac{\pi}{3} - \frac{\psi(z)}{3}\right)}{3\sqrt{3}v\lambda + 16\|z\|_2^{3/2}\cos^3\left(\frac{\pi}{3} - \frac{\psi(z)}{3}\right)} z, & \|z\|_2 = \frac{3}{2}(v\lambda)^{2/3}, \\ 0, & \|z\|_2 < \frac{3}{2}(v\lambda)^{2/3}, \end{cases} \tag{61}$$

with

$$\psi(z) = \arccos\left(\frac{v\lambda}{4}\left(\frac{3}{\|z\|_2}\right)^{3/2}\right), \tag{62}$$

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

$$P_{1,1/2}(z) = \begin{cases} \tilde{z}, & Q_{1,1/2}(\tilde{z}) < Q_{1,1/2}(0), \\ 0 \text{ or } \tilde{z}, & Q_{1,1/2}(\tilde{z}) = Q_{1,1/2}(0), \\ 0, & Q_{1,1/2}(\tilde{z}) > Q_{1,1/2}(0), \end{cases}$$

with

$$\tilde{z} = z - \frac{\sqrt{3}v\lambda}{4\sqrt{\|z\|_1}\cos\left(\frac{\pi}{3} - \frac{\xi(z)}{3}\right)}\text{sign}(z), \quad \xi(z) = \arccos\left(\frac{v\lambda l}{4}\left(\frac{3}{\|z\|_1}\right)^{3/2}\right),$$

(v) *if* $p = 2$ *and* $q = 2/3$, *then*

$$P_{2,2/3}(z) = \begin{cases} \frac{3\left(a^{3/2}+\sqrt{2\|z\|_2 - a^3}\right)}{32v\lambda a^2 + 3\left(a^{3/2}+\sqrt{2\|z\|_2-a^3}\right)}z, & \|z\|_2 > 2\left(\frac{2}{3}v\lambda\right)^{3/4}, \\ 0 \text{ or } \frac{3\left(a^{3/2}+\sqrt{2\|z\|_2 - a^3}\right)}{32v\lambda a^2 + 3\left(a^{3/2}+\sqrt{2\|z\|_2-a^3}\right)}z, & \|z\|_2 = 2\left(\frac{2}{3}v\lambda\right)^{3/4}, \\ 0, & \|z\|_2 < 2\left(\frac{2}{3}v\lambda\right)^{3/4}, \end{cases} \tag{63}$$

with

$$a = \frac{2}{\sqrt{3}}(2v\lambda)^{1/4}\left(\cosh\left(\frac{\varphi(z)}{3}\right)\right)^{1/2}, \quad \varphi(z) = \text{arccosh}\left(\frac{27\|z\|_2^2}{16(2v\lambda)^{3/2}}\right), \tag{64}$$

(vi) *if* $p = 1$ *and* $q = 2/3$, *then*

$$P_{1,2/3}(z) = \begin{cases} \bar{z}, & Q_{1,2/3}(\bar{z}) < Q_{1,2/3}(0), \\ 0 \text{ or } \bar{z}, & Q_{1,2/3}(\bar{z}) = Q_{1,2/3}(0), \\ 0, & Q_{1,2/3}(\bar{z}) > Q_{1,2/3}(0), \end{cases}$$

with

$$\bar{z} = z - \frac{4v\lambda\bar{a}^{1/2}}{3\left(\bar{a}^{3/2} + \sqrt{2\|z\|_1 - \bar{a}^3}\right)}\text{sign}(z),$$

and

$$\bar{a} = \frac{2}{\sqrt{3}}(2v\lambda l)^{1/4}\left(\cosh\left(\frac{\zeta(z)}{3}\right)\right)^{1/2}, \quad \zeta(z) = \text{arccosh}\left(\frac{27\|z\|_1^2}{16(2v\lambda l)^{3/2}}\right).$$

Proof Since the proximal regularization $Q_{p,q}(\cdot) := \lambda\|\cdot\|_p^q + \frac{1}{2v}\|\cdot - z\|_2^2$ is non-differentiable only at 0, $P_{p,q}(z)$ must be 0 or some point $\tilde{x}(\neq 0)$ satisfying the first-order optimality condition

$$\lambda q\|\tilde{x}\|_p^{q-p}\begin{pmatrix} |\tilde{x}_1|^{p-1}\text{sign}(\tilde{x}_1) \\ \vdots \\ |\tilde{x}_l|^{p-1}\text{sign}(\tilde{x}_l) \end{pmatrix} + \frac{1}{v}(\tilde{x} - z) = 0. \tag{65}$$

Thus, to derive the analytical formula of the proximal operator $P_{p,q}(z)$, we just need to calculate such $\tilde{x}$ via (65), and then compare the objective function values $Q_{p,q}(\tilde{x})$ and $Q_{p,q}(0)$ to obtain the solution inducing a smaller value. The proofs of the six statements follow in the above routine, and we only provide the detailed proofs of (iii) and (v) as samples.

(iii) When $p = 2$ and $q = 1/2$, (65) reduces to

$$\frac{\lambda\tilde{x}}{2\|\tilde{x}\|_2^{3/2}} + \frac{1}{v}(\tilde{x} - z) = 0, \tag{66}$$

and consequently,

$$\|\tilde{x}\|_2^{3/2} - \|z\|_2\|\tilde{x}\|_2^{1/2} + \frac{1}{2}v\lambda = 0. \tag{67}$$

Denote $\eta = \|\tilde{x}\|_2^{1/2} > 0$. The equation (67) can be transformed into the following cubic algebraic equation

$$\eta^3 - \|z\|_2\eta + \frac{1}{2}v\lambda = 0. \tag{68}$$

Due to the hyperbolic solution of the cubic equation (see Short, 1937), by denoting

$$r = 2\sqrt{\frac{\|z\|_2}{3}},\ \alpha = \arccos\left(\frac{v\lambda}{4}\left(\frac{3}{\|z\|_2}\right)^{3/2}\right) \text{ and } \beta = \operatorname{arccosh}\left(-\frac{v\lambda}{4}\left(\frac{3}{\|z\|_2}\right)^{3/2}\right),$$

the solution of (68) can be expressed as the follows.

(1) If $0 \le \|z\|_2 \le 3\left(\frac{v\lambda}{4}\right)^{2/3}$, then the three roots of (68) are given by

$$\eta_1 = r\cosh\frac{\beta}{3},\ \eta_2 = -\frac{r}{2}\cosh\frac{\beta}{3} + i\frac{\sqrt{3}r}{2}\sinh\frac{\beta}{3},\ \eta_3 = -\frac{r}{2}\cosh\frac{\beta}{3} - i\frac{\sqrt{3}r}{2}\sinh\frac{\beta}{3},$$

where i denotes the imaginary unit. However, this β does not exist since the value of hyperbolic cosine must be positive. Thus, in this case, $P_{2,1/2}(z) = 0$.

(2) If $\|z\|_2 > 3\left(\frac{v\lambda}{4}\right)^{2/3}$, then the three roots of (68) are

$$\eta_1 = r\cos\left(\frac{\pi}{3} - \frac{\alpha}{3}\right),\ \eta_2 = -r\sin\left(\frac{\pi}{2} - \frac{\alpha}{3}\right),\ \eta_3 = -r\cos\left(\frac{2\pi}{3} - \frac{\alpha}{3}\right).$$

The unique positive solution of (68) is $\|\tilde{x}\|_2^{1/2} = \eta_1$, and thus, the unique solution of (66) is given by

$$\tilde{x} = \frac{2\eta_1^3}{v\lambda + 2\eta_1^3}z = \frac{16\|z\|_2^{3/2}\cos^3\left(\frac{\pi}{3} - \frac{\psi(z)}{3}\right)}{3\sqrt{3}v\lambda + 16\|z\|_2^{3/2}\cos^3\left(\frac{\pi}{3} - \frac{\psi(z)}{3}\right)}z.$$

Finally, we compare the objective function values $Q_{2,1/2}(\tilde{x})$ and $Q_{2,1/2}(0)$. For this purpose, when $\|z\|_2 > 3\left(\frac{v\lambda}{4}\right)^{2/3}$, we define

$$\begin{aligned} H(\|z\|_2) &:= \tfrac{v}{\|\tilde{x}\|_2}\left(Q_{2,1/2}(0) - Q_{2,1/2}(\tilde{x})\right) \\ &= \tfrac{v}{\|\tilde{x}\|_2}\left(\tfrac{1}{2v}\|z\|_2^2 - \lambda\|\tilde{x}\|_2^{1/2} - \tfrac{1}{2v}\|\tilde{x} - z\|_2^2\right) \\ &= \|z\|_2 - \tfrac{\|\tilde{x}\|_2^2 + 2v\lambda\|\tilde{x}\|_2^{1/2}}{2\|\tilde{x}\|_2} \\ &= \tfrac{1}{2}\|z\|_2 - \tfrac{3}{4}v\lambda\|\tilde{x}\|_2^{-1/2}, \end{aligned}$$

where the third equality holds since that $\tilde{x}$ is proportional to z, and fourth equality follows from (67). Since both $\|z\|_2$ and $\|\tilde{x}\|_2$ are strictly increasing on $\|z\|_2$, $H(\|z\|_2)$ is also strictly increasing when $\|z\|_2 > 3\left(\frac{v\lambda}{4}\right)^{2/3}$. Thus the unique solution of $H(\|z\|_2) = 0$ satisfies

$$\|z\|_2\|\tilde{x}\|_2^{1/2} = \frac{3}{2}v\lambda,$$

and further, (67) implies that the solution of $H(\|z\|_2) = 0$ is

$$\|z\|_2 = \frac{3}{2}(v\lambda)^{2/3}.$$

Therefore, we arrive at the formulae (61) and (62).
(v) When $p = 2$ and $q = 2/3$, (65) reduces to

$$\frac{2\lambda\tilde{x}}{3\|\tilde{x}\|_2^{4/3}} + \frac{1}{v}(\tilde{x} - z) = 0, \tag{69}$$

and consequently,

$$\|\tilde{x}\|_2^{4/3} - \|z\|_2\|\tilde{x}\|_2^{1/3} + \frac{2}{3}v\lambda = 0. \tag{70}$$

Denote $\eta = \|\tilde{x}\|_2^{1/3} > 0$ and $h(t) = t^4 - \|z\|_2 t + \frac{2}{3}v\lambda$ for any $t \in \mathbb{R}$. Thus, η is the positive solution of $h(t) = 0$. Next, we seek η by the method of undetermined coefficients. Assume that

$$h(t) = t^4 - \|z\|_2 t + \frac{2}{3}v\lambda = (t^2 + at + b)(t^2 + ct + d), \quad \text{where } a, b, c, d \in \mathbb{R}. \tag{71}$$

By expansion and comparison, we have that

$$a + c = 0, \quad b + d + ac = 0, \quad ad + bc = -\|z\|_2, \quad bd = \frac{2}{3}v\lambda,$$

and thus,

$$c = -a, b = \frac{1}{2}\left(a^2 + \frac{\|z\|_2}{a}\right), d = \frac{1}{2}\left(a^2 - \frac{\|z\|_2}{a}\right), bd = \frac{1}{4}\left(a^4 - \frac{\|z\|_2^2}{a^2}\right) = \frac{2}{3}v\lambda. \tag{72}$$

By letting $M = a^2$, the last one of the above equalities reduces to the following cubic algebraic equation

$$M^3 - \frac{8}{3}v\lambda M - \|z\|_2^2 = 0. \tag{73}$$

According to the Cardano formula for the cubic equation, the root of (73) can be represented by

$$a^2 = M = \left(\frac{\|z\|_2^2}{2} + \sqrt{\frac{\|z\|_2^4}{4} - \left(\frac{8}{9}v\lambda\right)^3}\right)^{1/3} + \left(\frac{\|z\|_2^2}{2} - \sqrt{\frac{\|z\|_2^4}{4} - \left(\frac{8}{9}v\lambda\right)^3}\right)^{1/3},$$

which can also be reformulated in the following hyperbolic form (see Short (1937))

$$a^2 = M = \frac{4}{3}\sqrt{2v\lambda}\cosh\left(\frac{\varphi(z)}{3}\right), \tag{74}$$

where $\varphi(z)$ is given by (64). By (71) and (72), we have that η, the positive root of $h(t) = 0$, satisfies

$$\eta^2 + a\eta + \frac{1}{2}\left(a^2 + \frac{\|z\|_2}{a}\right) = 0 \quad \text{or} \quad \eta^2 - a\eta + \frac{1}{2}\left(a^2 - \frac{\|z\|_2}{a}\right) = 0.$$

Hence, the real roots of the above equations, that is, the real roots of $h(t) = 0$, are

$$\eta_1 = \frac{1}{2}\left(|a| + \sqrt{\frac{2\|z\|_2}{|a|} - a^2}\right) \quad \text{and} \quad \eta_2 = \frac{1}{2}\left(|a| - \sqrt{\frac{2\|z\|_2}{|a|} - a^2}\right). \tag{75}$$

It is easy to see that $\eta_1 > \eta_2$ and that η_2 should be discarded as it induces the saddle point rather than a minimum (since $h(t) > 0$ when $t < \eta_2$). Thus, by (69), (74) and (75), one has

$$\tilde{x} = \frac{3\eta_1^4}{2v\lambda + 3\eta_1^4} z = \frac{3\left(a^{3/2} + \sqrt{2\|z\|_2 - a^3}\right)}{32v\lambda a^2 + 3\left(a^{3/2} + \sqrt{2\|z\|_2 - a^3}\right)} z,$$

where a is given by (64). Finally, we compare the objective function values $Q_{2,2/3}(\tilde{x})$ and $Q_{2,2/3}(0)$. For this purpose, we define

$$\begin{aligned} H(\|z\|_2) &:= \tfrac{v}{\|\tilde{x}\|_2}\left(Q_{2,2/3}(0) - Q_{2,2/3}(\tilde{x})\right) \\ &= \tfrac{v}{\|\tilde{x}\|_2}\left(\tfrac{1}{2v}\|z\|_2^2 - \lambda\|\tilde{x}\|_2^{2/3} - \tfrac{1}{2v}\|\tilde{x} - z\|_2^2\right) \\ &= \|z\|_2 - \tfrac{\|\tilde{x}\|_2^2 + 2v\lambda\|\tilde{x}\|_2^{2/3}}{2\|\tilde{x}\|_2} \\ &= \tfrac{1}{2}\|z\|_2 - \tfrac{2}{3}v\lambda\|\tilde{x}\|_2^{-1/3}, \end{aligned}$$

where the third equality holds since that $\tilde{x}$ is proportional to z, and fourth equality follows from (70). Since both $\|z\|_2$ and $\|\tilde{x}\|_2$ are strictly increasing on $\|z\|_2$, $H(\|z\|_2)$ is also strictly increasing when $\|z\|_2 > 4(\frac{2}{9}v\lambda)^{3/4}$. Thus the unique solution of $H(\|z\|_2) = 0$ satisfies

$$\|z\|_2\|\tilde{x}\|_2^{1/3} = \frac{4}{3}v\lambda,$$

and further, (70) implies that the solution of $H(\|z\|_2) = 0$ is

$$\|z\|_2 = 2\left(\frac{2}{3}v\lambda\right)^{3/4}.$$

Therefore, we arrive at the formulae (63) and (64). The proof is complete. ■

Remark 19 *Note from Proposition 18 that the solutions of the proximal optimization subproblems might not be unique when $Q_{p,q}(\tilde{x}) = Q_{p,q}(0)$. To avoid this obstacle in numerical computations, we select the solution $P_{p,q}(z) = 0$ whenever $Q_{p,q}(\tilde{x}) = Q_{p,q}(0)$, which achieves a more sparse solution, in the definition of the proximal operator to guarantee a unique update.*

Remark 20 *By Proposition 18, one sees that the PGM-GSO meets the group sparsity structure, since the components of each iterate within each group are likely to be either all zeros or all nonzeros. When $n_{\max} = 1$, 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 $P_{2,1}(z)$, $P_{2,0}(z)$, and $P_{2,1/2}(z)$ and $P_{1,1/2}(z)$ reduce to the soft thresholding function in Daubechies et al. (2004), the hard thresholding function in Blumensath and Davies (2008) and the half thresholding function in Xu et al. (2012), respectively.*

Remark 21 *Proposition 18 presents the analytical solution of the proximal optimization subproblems* (60) *when* $q = 0, 1/2, 2/3, 1$. *However, in other cases, the analytical solution of problem* (60) *seems not available, since the algebraic equation* (65) *does not have an analytical solution (it is difficult to find an analytical solution for the algebraic equation whose order is larger than four). Thus, in the general cases of* $q \in (0, 1)$, *we alternatively use the Newton method to solve the nonlinear equation* (65), *which is the optimality condition of the proximal optimization subproblem. The numerical simulation in Figure 5 of section 4 shows that the Newton method works in solving the proximal optimization subproblems* (60) *for the general* q, *while the* $\ell_{p,1/2}$ *regularization is the best one among the* $\ell_{p,q}$ *regularizations for* $q \in [0, 1]$.

4. Numerical Experiments

The purpose of this section is to carry out the numerical experiments of the proposed PGM for the $\ell_{p,q}$ regularization problem. We illustrate the performance of the PGM-GSO among different types of $\ell_{p,q}$ regularization, in particular, when (p, q) =(2,1), (2,0), (2,1/2), (1,1/2), (2,2/3) and (1,2/3), by comparing them with several state-of-the-art algorithms for simulated data and applying them to infer gene regulatory network from gene expression data of mouse embryonic stem cell. All numerical experiments are implemented in Matlab R2013b and executed on a personal desktop (Intel Core Duo E8500, 3.16 GHz, 4.00 GB of RAM). The R package of the PGM for solving group sparse optimization, named GSparO in CRAN, is available at https://CRAN.R-project.org/package=GSparO

4.1 Simulated Data

In the numerical experiments on simulated data, 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 r 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 to be zeros. We generate the data b by the Matlab script

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

where *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 $sigma = 0.1\%$. Assuming the group sparsity level S is predefined, the regularization parameter λ is iteratively updated by obeying the rule: we set the iterative threshold to be the S-th largest value of $\|z^k_{\mathcal{G}_i}\|_2$ and solve the λ by virtue of Proposition 18.

For each given sparsity level, which is k/r, we randomly generate the data A, $\bar{x}$, b (as above) 500 times, run the algorithm, and average the 500 numerical results to illustrate the performance of the algorithm. We choose the stepsize $v = 1/2$ in the tests of the PGM-GSO. 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 0.5%; otherwise, it is regarded as *failure*.

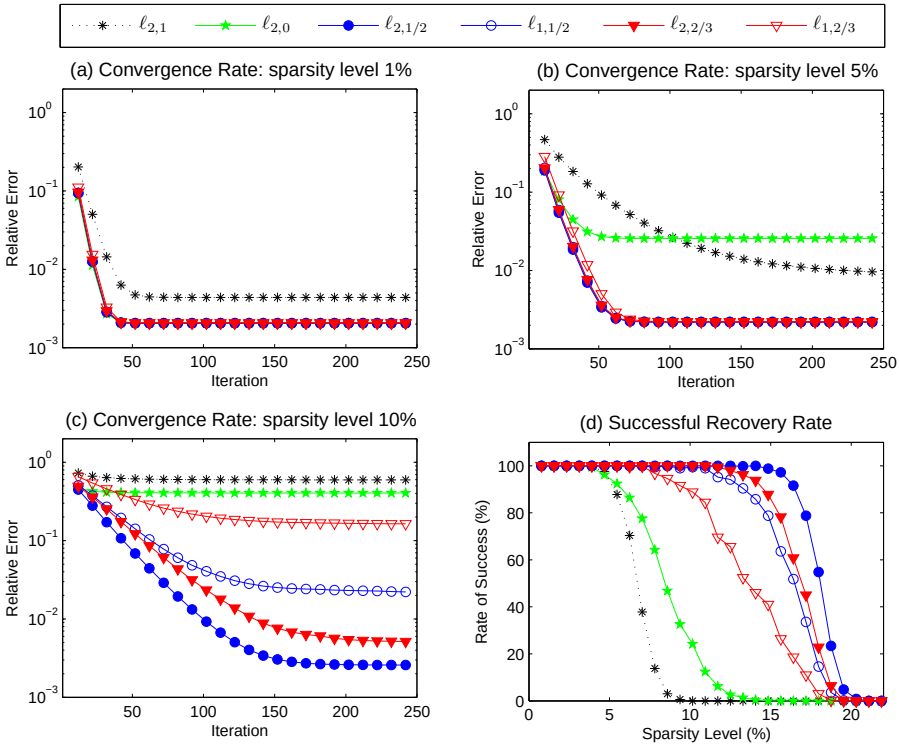


Figure 3: Convergence results and recovery rates for different sparsity levels.

We carry out six experiments with the initial point $x_0 = 0$ (unless otherwise specified). In the first experiment, setting $r = 128$ (so group size $G = 1024/128 = 8$), we compare the convergence rates and the successful recovery rates of the PGM-GSO with $(p, q) = (2, 1), (2, 0), (2, 1/2), (1, 1/2), (2, 2/3)$ and $(1, 2/3)$ for different sparsity levels. In Figure 3, (a), (b), and (c) illustrate the convergence rate results on sparsity level 1%, 5%, and 10%, respectively, while (d) plots the successful recovery rates on different sparsity levels. When the solution is of high sparse level, as shown in Figure 3(a), all $\ell_{p,q}$ regularization problems perform perfect and achieve a fast convergence rate. As demonstrated in Figure 3(b), when the sparsity level drops to 5%, $\ell_{p,1/2}$ and $\ell_{p,2/3}$ ($p = 1$ and 2) perform better and arrive at a more accurate level than $\ell_{2,1}$ and $\ell_{2,0}$. As illustrated in Figure 3(c), when the sparsity level is 10%, $\ell_{p,1/2}$ further outperforms $\ell_{p,2/3}$ ($p = 1$ or 2), and it surprises us that $\ell_{2,q}$ performs better than $\ell_{1,q}$ ($q = 1/2$ or $2/3$). From Figure 3(d), it is illustrated that $\ell_{p,1/2}$ achieves a better successful recovery rate than $\ell_{p,2/3}$ ($p = 1$ or 2), which outperforms $\ell_{2,0}$ and $\ell_{2,1}$. Moreover, we surprisingly see that $\ell_{2,q}$ also outperforms $\ell_{1,q}$ ($q = 1/2$ or $2/3$) on the successful recovery rate. In a word, $\ell_{2,1/2}$ performs as the best one of these six regularizations on both accuracy and robustness. In this experiment, we also note that the running times are at a same level, about 0.9 second per 500 iterations.

The second experiment is performed to show the sensitivity analysis on the group size ($G = 4, 8, 16, 32$) of the PGM-GSO with the six types of $\ell_{p,q}$ regularization. As shown in

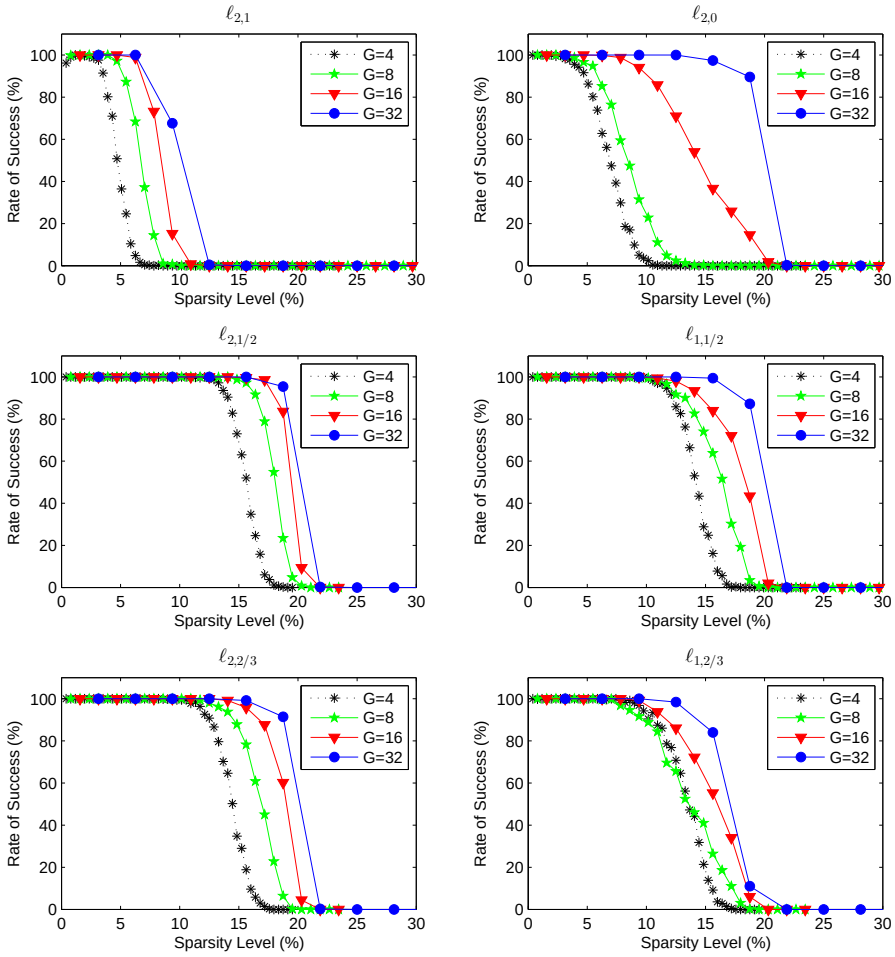


Figure 4: Sensitivity analysis on group size.

Figure 4, the six types of $\ell_{p,q}$ regularization 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.

The third experiment is implemented to study the variation of the PGM-GSO when varying the regularization order q (fix $p = 2$). Recall from Theorem 18 that the analytical solution of the proximal optimization subproblem (60) can be obtained when $q = 0, 1/2, 2/3, 1$. However, in other cases, the analytical solution of subproblem (60) seems not available, and thus we apply the Newton method to solve the nonlinear equation (65), which is the optimality condition of the proximal optimization subproblem. Figure 5 shows the variation of successful recovery rates by decreasing the regularization order q from 1 to 0. It is illustrated that the PGM-GSO achieves the best successful recovery rate when $q = 1/2$, which arrives at the same conclusion as the first experiment. The farther the distance of q from $1/2$, the lower the successful recovery rate.

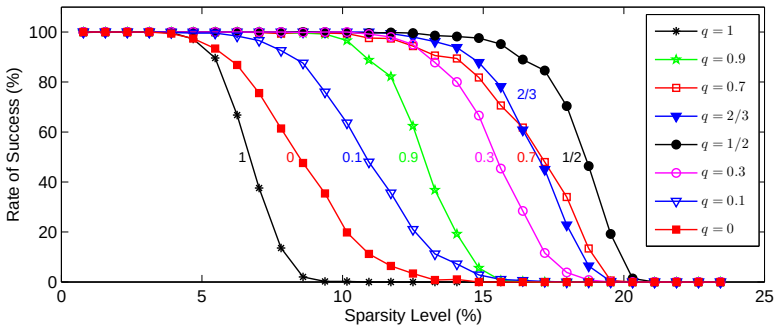


Figure 5: Variation of the PGM-GSO when varying the regularization order q.

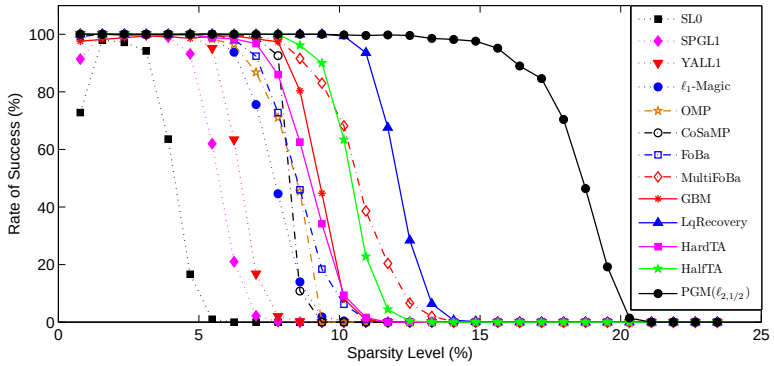


Figure 6: Comparison between the PGM-GSO and several state-of-the-art algorithms.

The fourth experiment is to compare the PGM-GSO with several state-of-the-art algorithms in the field of sparse optimization, either convex or nonconvex algorithms, as listed in Table 1. All these algorithms, including PGM-GSO, can successfully recover the signal when the solution is of high sparse level. However, some of these algorithms fails to obtain the group sparsity structure along with the group sparsity level decreases. Figure 7 plots the signals estimated by these algorithms in a random trial at a group sparsity level of 15%. It is illustrated that the solutions of the MultiFoBa and the PGM-GSO type solvers are of group sparsity structure, while other algorithms do not obtain the group sparse solutions. In these solvers, the MultiFoBa and the $\ell_{2,1}$ obtain the true active groups but inaccurate weights, the $\ell_{2,0}$ achieves the accurate weights but some incorrect active groups, while $\ell_{p,1/2}$ and $\ell_{p,2/3}$ recover perfect solutions in both true active groups and accurate weights. Figure 6 demonstrates the overall performance of these algorithms by plotting the successful recovery rates on different sparsity levels. It is indicated by Figure 6 that $\ell_{2,1/2}$ can achieve the higher successful recovery rate than other algorithms, by exploiting the group sparsity structure and lower-order regularization. From this experiment, it is demonstrated that the PGM-GSO ($\ell_{2,1/2}$) outperforms most solvers of sparse learning in solving sparse optimization problems with group structure.

SL0	SL0 is an algorithm for finding the sparsest solutions of an underdetermined system of linear equations based on Smoothed ℓ_0 norm (Mohimani et al., 2009). The package is available at http://ee.sharif.edu/~SLzero/
SPGL1	SPGL1 is a Matlab solver for large-scale sparse reconstruction (van den Berg and Friedlander, 2008). The package is available at http://www.cs.ubc.ca/~mpf/spgl1/
YALL1	YALL1 (Your ALgorithm for L1) is a package of Matlab solvers for the ℓ_1 sparse reconstruction, by virtue of the alternating direction method (Deng et al., 2011; Yang and Zhang, 2011). The package is available at http://yall1.blogs.rice.edu/
OMP	Orthogonal Matching Pursuit algorithm for the recovery of a high-dimensional sparse signal (Cai and Wang, 2011). The package is available at MathWorks.
CoSaMP	Compressive Sampling Matched Pursuit algorithms for the recovery of a high-dimensional sparse signal (Needell and Tropp, 2009). The packages are available at MathWorks.
FoBa	Adaptive forward-backward greedy algorithm for sparse learning (Zhang, 2011). The R package is available at https://CRAN.R-project.org/package=foba
MultiFoBa	MultiFoBa is group FoBa for multi-task learning (Tian et al., 2016).
ℓ_1**-Magic**	ℓ_1-Magic is a collection of Matlab routines for solving the convex optimization programs central to compressive sampling, based on standard interior-point methods (Candès et al., 2006a). The package is available at https://statweb.stanford.edu/~candes/l1magic/
ISTA	Iterative Shrinkage/Thresholding Algorithm (Daubechies et al., 2004) for solving the ℓ_1 regularization problem.
GBM	Gradient Based Method for solving the $\ell_{1/2}$ regularization problem (Wu et al., 2014). Suggested by the authors, we choose the initial point as the solution obtained by the ℓ_1-Magic.
LqRecovery	LqRecovery is an iterative algorithm for the ℓ_p norm minimization (Foucart and Lai, 2009). The code is available at http://www.math.drexel.edu/~foucart/software.htm
HardTA	Iterative Hard Thresholding Algorithm (Blumensath and Davies, 2008) for solving the ℓ_0 regularization problem.
HalfTA	Iterative Half Thresholding Algorithm (Xu et al., 2012) for solving the $\ell_{1/2}$ regularization problem.

Table 1: List of the state-of-the-art algorithms for sparse learning.

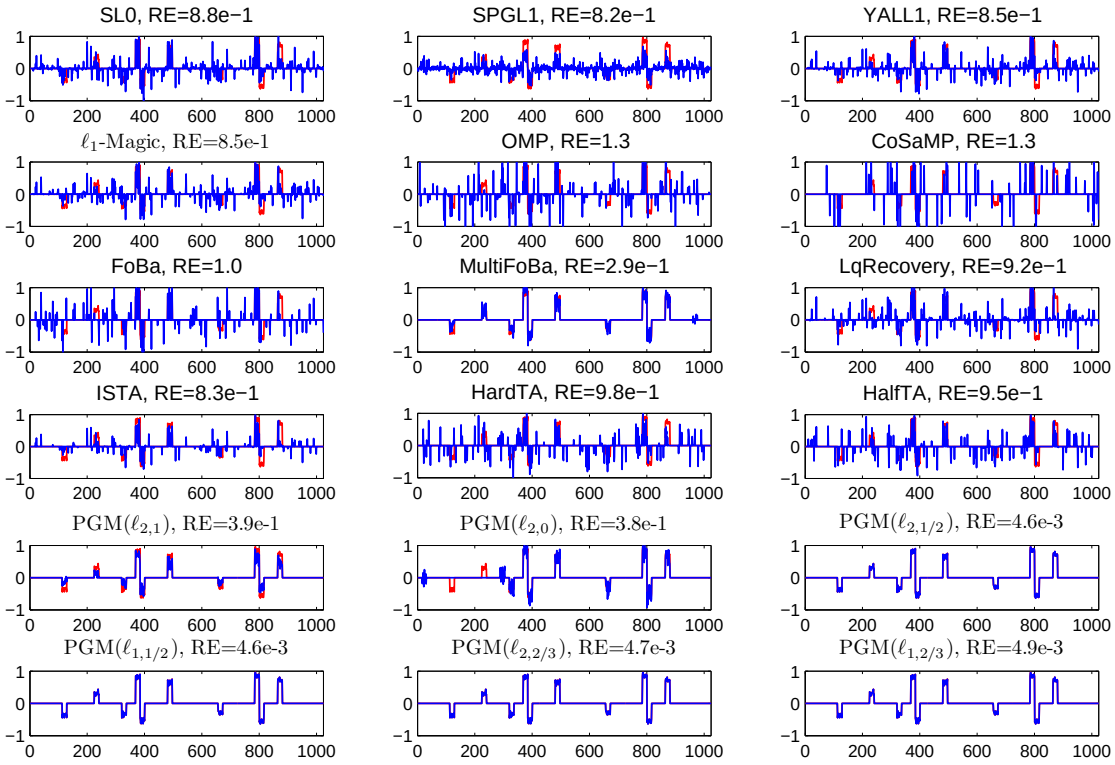


Figure 7: Simulation of the PGM-GSO and several state-of-the-art algorithms.

The fifth experiment is devoted to the phase diagram study (see Donoho, 2006b) of the $\ell_{p,q}$ regularization problem, which further demonstrates the stronger sparsity promoting capability of $\ell_{p,1/2}$ and $\ell_{p,2/3}$ $(p = 1, 2)$ regularization over $\ell_{2,1}$ regularization. In this experiment, we consider a noiseless signal recovery problem with $n = 512$, $G = 4$, $r = 128$ and $sigma = 0$ as a prototype. More specifically, for each fixed m, we vary the number of active groups k from 1 to m/G, that is, the sparsity level varies from $1/r$ to $m/(Gr)$, and then, we increase m from G to n in the way such that 128 equidistributed values $m_j = jG$ are considered. For each specific problem size, we randomly generate the data 500 times and apply the PGM-GSO to solve the $\ell_{p,q}$ regularization problem. For these noiseless data, the recovery is defined as *success* whenever the relative error between the recovered data and the true data is smaller than 10^{-5}, otherwise, it is regarded as *failure*. Also, we embody a pixel blue whenever the point is in the case of *success*, otherwise, red when *failure*. In this way, a phase diagram of an algorithm is plotted in Figure 8, where the color of each cell reflects the empirical recovery rate (scaled between 0 and 1). It is illustrated in Figure 8 that the phase transition phenomenon does appear for the PGM-GSO with the six types of $\ell_{p,q}$ regularization. It is shown that $\ell_{p,1/2}$ and $\ell_{p,2/3}$ $(p = 1, 2)$ regularizations are more robust than that of $\ell_{2,0}$ and $\ell_{2,1}$ in the sense that they allow to achieve higher recovery rates and recover a sparse signal from a smaller amount of samples. It is also revealed by

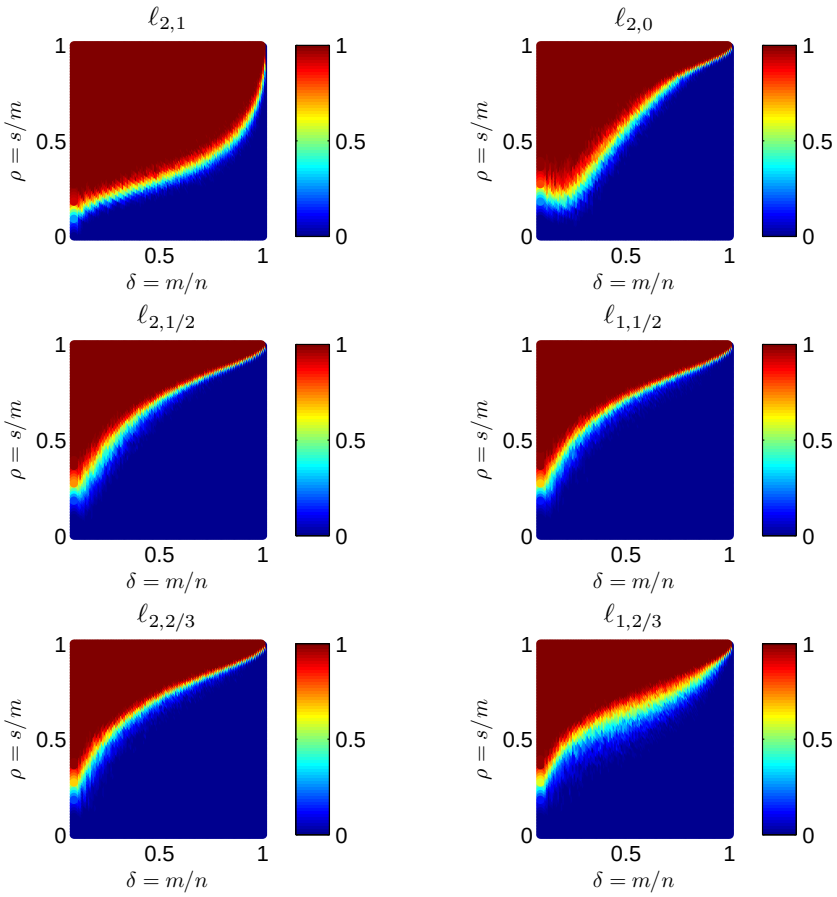


Figure 8: Phase diagram study of $\ell_{p,q}$ regularization of group sparse optimization. Blue denotes perfect recovery in all experiments, and red denotes failure for all experiments.

Figure 8 that $\ell_{2,1/2}$ regularization is the most robust one of these six regularizations, which achieves the same conclusion as the first and third experiments.

Even though some global optimization method, such as the filled function method (Ge, 1990), can find the global solution of the lower-order regularization problem as in Example 2, however, it does not work for the large-scale sparse optimization problems. Because, in the filled function method, all the directions need to be searched or compared in each iteration, which costs a large amount of time and hampers the efficiency for solving the large-scale problems.

4.2 Real Data in Gene Transcriptional Regulation

Gene transcriptional regulation is the process that a combination of transcription factors (TFs) act in concert to control the transcription of the target genes. Inferring gene regulatory network from high-throughput genome-wide data is still a major challenge in systems biology, especially when the number of genes is large but the number of experimental samples is small. In large genomes, such as human and mouse, the complexity of gene regulatory system dramatically increases. Thousands of TFs combine in different ways to regulate tens of thousands target genes in various tissues or biological processes. However, only a few TFs collaborate and usually form complexes (groups of cooperative TFs) to control the expression of a specific gene in a specific cell type or developmental stage. Thus, the prevalence of TF complex makes the solution of gene regulatory network have a group structure, and the gene regulatory network inference in such large genomes becomes a group sparse optimization problem, which is to search a small number of TF complexes (or TFs) from a pool of thousands of TF complexes (or TFs) for each target gene based on the dependencies between the expression of TF complexes (or TFs) and the targets. Even though TFs often work in the form of complexes (Xie et al., 2013), and TF complexes are very important in the control of cell identity and diseases (Hnisz et al., 2013), current methods to infer gene regulatory network usually consider each TF separately. To take the grouping information of TF complexes into consideration, we can apply the group sparse optimization to gene regulatory network inference with the prior knowledge of TF complexes as the pre-defined grouping.

4.2.1 Materials

Chromatin immunoprecipitation (ChIP) coupled with next generation sequencing (ChIP-seq) identifies *in vivo* active and cell-specific binding sites of a TF. They are commonly used to infer TF complexes recently. Thus, we manually collect ChIP-seq data in mouse embryonic stem cells (mESCs), as shown in Table 2. Transcriptome is the gene expression profile of the whole genome that is measured by microarray or RNA-seq. The transcriptome data in mESCs for gene regulatory network inference are downloaded from Gene Expression Omnibus (GEO). 245 experiments under perturbations in mESC are collected from three papers Correa-Cerro et al. (2011); Nishiyama et al. (2009, 2013). Each experiment produced transcriptome data with or without overexpression or knockdown of a gene, in which the control and treatment have two replicates respectively. Gene expression fold changes between control samples and treatment samples of 12488 target genes in all experiments are $\log 2$ transformed and form matrix $B \in \mathbb{R}^{245\times 12488}$ (Figure 9A). The known TFs are collected from four TF databases, TRANSFAC, JASPAR, UniPROBE and TFCat, as well as literature. Let matrix $H \in \mathbb{R}^{245\times 939}$ be made up of the expression profiles of 939 known TFs, and matrix $Z \in \mathbb{R}^{939\times 12488}$ describe the connections between these TFs and targets. Then, the regulatory relationship between TFs and targets can be represented approximately by a linear system

$$HZ = B + \epsilon.$$

The TF-target connections defined by ChIP-seq data are converted into an initial matrix Z^0 (see Qin et al., 2014). Indeed, if TF i has a binding site around the gene j promoter within a defined distance (10 kbp), a non-zero number is assigned on Z^0_{ij} as a prior value.

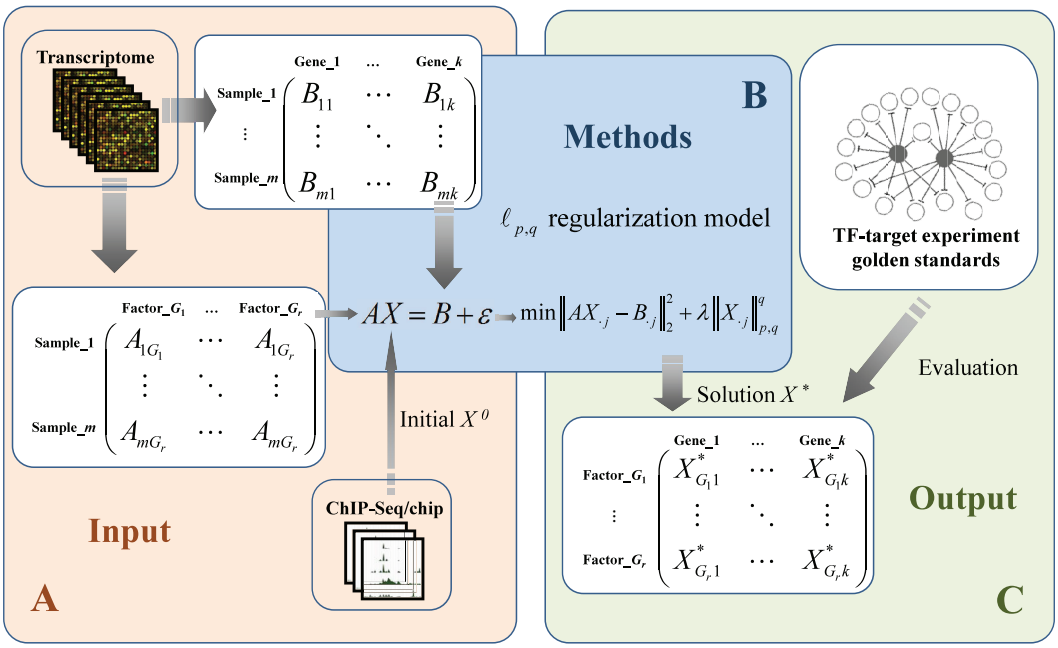


Figure 9: Workflow of gene regulatory network inference with $\ell_{p,q}$ regularization.

Now we add the grouping information (TF complexes) into this linear system. The TF complexes are inferred from ChIP-seq data (Table 2) via the method described in Giannopoulou and Elemento (2013). Let the group structure of Z be a matrix $W \in \mathbb{R}^{2257\times 939}$ (actually, the number of groups is 1439), whose Moore-Penrose pseudoinverse is denoted by W^+ (Horn and Johnson., 1985). We further let $A := HW^+$ and $X := WZ$. Then the linear system can be converted into

$$AX = B + \epsilon,$$

where A denotes expression profiles of TF complexes, and X represents connections between TF complexes and targets (Figure 9A).

A literature-based golden standard (low-throughput golden standard) TF-target pair set from biological studies (Figure 9C), including 97 TF-target interactions between 23 TFs and 48 target genes, is downloaded from iScMiD (Integrated Stem Cell Molecular Interactions Database). Each TF-target pair in this golden standard data set has been verified by biological experiments. Another more comprehensive golden standard mESC network is constructed from high-throughput data (high-throughput golden standard) by ChIP-Array (Qin et al., 2011) using the methods described in Qin et al. (2014). It contains 40006 TF-target pairs between 13092 TFs or targets (Figure 9C). Basically, each TF-target pair in the network is evidenced by a cell-type specific binding site of the TF on the target's promoter and the expression change of the target in the perturbation experiment of the TF,

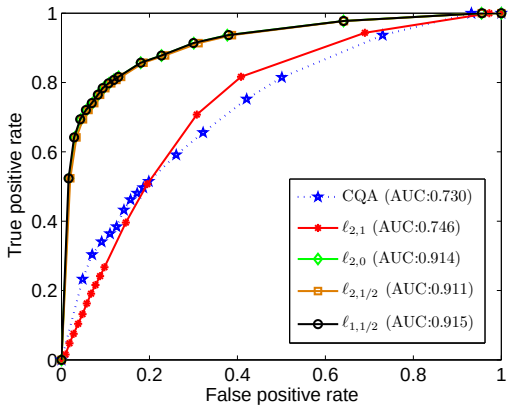
Factor	GEO accession	Pubmed ID	Factor	GEO accession	Pubmed ID
Atf7ip	GSE26680	-	Rad21	GSE24029	21589869
Atrx	GSE22162	21029860	Rbbp5	GSE22934	21477851
Cdx2	GSE16375	19796622	Rcor1	GSE27844	22297846
Chd4	GSE27844	22297846	Rest	GSE26680	-
Ctcf	GSE11431	18555785	Rest	GSE27844	22297846
Ctcf	GSE28247	21685913	Rnf2	GSE13084	18974828
Ctr9	GSE20530	20434984	Rnf2	GSE26680	-
Dpy30	GSE26136	21335234	Rnf2	GSE34518	22305566
E2f1	GSE11431	18555785	Setdb1	GSE17642	19884257
Ep300	GSE11431	18555785	Smad1	GSE11431	18555785
Ep300	GSE28247	21685913	Smad2	GSE23581	21731500
Esrrb	GSE11431	18555785	Smarca4	GSE14344	19279218
Ezh2	GSE13084	18974828	Smc1a	GSE22562	20720539
Ezh2	GSE18776	20064375	Smc3	GSE22562	20720539
Jarid2	GSE18776	20064375	Sox2	GSE11431	18555785
Jarid2	GSE19365	20075857	Stat3	GSE11431	18555785
Kdm1a	GSE27844	22297846	Supt5h	GSE20530	20434984
Kdm5a	GSE18776	20064375	Suz12	GSE11431	18555785
Klf4	GSE11431	18555785	Suz12	GSE13084	18974828
Lmnb1	GSE28247	21685913	Suz12	GSE18776	20064375
Med1	GSE22562	20720539	Suz12	GSE19365	20075857
Med12	GSE22562	20720539	Taf1	GSE30959	21884934
Myc	GSE11431	18555785	Taf3	GSE30959	21884934
Mycn	GSE11431	18555785	Tbp	GSE30959	21884934
Nanog	GSE11431	18555785	Tbx3	GSE19219	20139965
Nipbl	GSE22562	20720539	Tcfcp2l1	GSE11431	18555785
Nr5a2	GSE19019	20096661	Tet1	GSE26832	21451524
Pou5f1	GSE11431	18555785	Wdr5	GSE22934	21477851
Pou5f1	GSE22934	21477851	Whsc2	GSE20530	20434984
Prdm14	GSE25409	21183938	Zfx	GSE11431	18555785

Table 2: ChIP-seq data for TF complex inference.

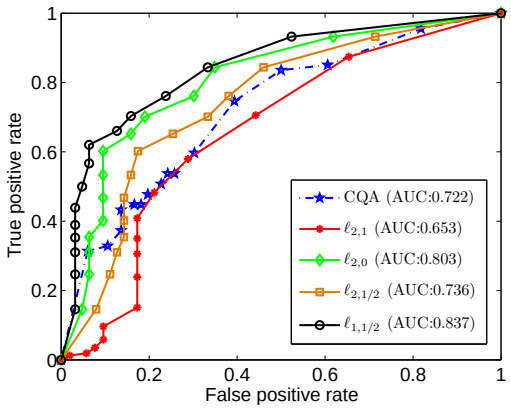
which is generally accepted as a true TF-target regulation. These two independent golden standards are both used to validate the accuracy of the inferred gene regulatory networks.

4.2.2 Numerical Results

We apply the PGM-GSO, starting from the initial matrix $X^0 := WZ^0$, to the gene regulatory network inference problem (Figure 9B), and compare with the CQ algorithm (CQA), which is shown in Wang et al. (2017) an efficient solver for gene regulatory network inference based on the group structure of TF complexes. The area under the curve (AUC) of a receiver operating characteristic (ROC) curve is widely recognized as an important index of the overall classification performance of an algorithm (see Fawcett, 2006). Here, we apply AUC to evaluate the performance of the PGM-GSO with four types of $\ell_{p,q}$ regularization, $(p,q) = (2,1), (2,0), (2,1/2)$ and $(1,1/2)$, as well as the CQA. A series of numbers of predictive TF complexes (or TFs), denoted by k, from 1 to 100 (that is, the sparsity level varies from about 0.07% to 7%) are tested. For each k and each pair of TF complex (or TF) i and target j, if the $X^{(k)}_{G_i j}$ is non-zero, this TF complex (or TF) is regarded as a potential regulator of this target in this test. In biological sense, we only concern about whether the true TF is predicted, but not the weight of this TF. We also expect that the TF complexes (or TFs) which are predicted in a higher sparsity level should be more important than those that are only reported in a lower sparsity level. Thus, when calculating the AUC, a score



(a) Evaluation with high-throughput golden standard.



(b) Evaluation with literature-based low-throughput golden standard.

Figure 10: ROC curves and AUCs of the PGM-GSO on mESC gene regulatory network inference.

Score_{ij} is applied as the predictor for TF i on target j:

$$\text{Score}_{ij} := \begin{cases} \max_k\{1/k\}, & X_{ij}^{(k)} \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Both high-throughput and low-throughput golden standards are used to draw the ROC curves of the PGM-GSO with four types of $\ell_{p,q}$ regularization and the CQA in Figure 10 to compare their accuracy. When matched with the high-throughput golden standard, it is illustrated from Figure 10(a) that $\ell_{2,1/2}$, $\ell_{1,1/2}$ and $\ell_{2,0}$ perform almost the same (as indicated by the almost same AUC value), and significantly outperform $\ell_{2,1}$ and CQA. With the low-throughput golden standard, it is demonstrated from Figure 10(b) that $\ell_{1,1/2}$ is slightly better than $\ell_{2,1/2}$, $\ell_{2,0}$ and CQA, and these three regularizations perform much better than $\ell_{2,1}$. These results are basically consistent with the results from simulated data. Since the golden standards we use here are obtained from real biological experiments, which are well-accepted as true TF-target regulations, the higher AUC, the more biologically accurate the result gene regulatory network is. Thus, our results indicate that the $\ell_{p,1/2}$ and $\ell_{p,0}$ regularizations are applicable to gene regulatory network inference in biological researches that study higher organisms but generate transcriptome data for only a small number of samples, which facilitates biologists to analyze gene regulation in a system level.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable suggestions and remarks which helped to improve the quality of the paper. We are also thankful to Professor Marc Teboulle for providing the reference Bolte et al. (2013) and the suggestion that the global convergence of the PGM-GSO can be established by using the so-called Kurdyka-

Łojasiewicz theory, and Professor Junwen Wang for providing the gene expression data and biological golden standards for inferring gene regulatory network of mouse embryonic stem cell. Hu's work was supported in part by the National Natural Science Foundation of China (11601343), Natural Science Foundation of Guangdong (2016A030310038) and Foundation for Distinguished Young Talents in Higher Education of Guangdong (2015KQNCX145). Li's work was supported in part by the National Natural Science Foundation of China (11571308, 11371325, 91432302). Meng's work was supported in part by the National Natural Science Foundation of China (11671329, 31601066, 71601162). Qin's work was supported in part by the National Natural Science Foundation of China (41606143). Yang's work was supported in part by the Research Grants Council of Hong Kong (PolyU 152167/15E) and the National Natural Science Foundation of China (11431004).

References

H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

F. R. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Cambridge, 1999.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14:629–654, 2008.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2013.

K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14(5):813–837, 2008.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

T. T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57:4680–4688, 2011.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006a.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8): 1207–1223, 2006b.

R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24:1–14, 2008.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43:129–159, 2001.

X. Chen, F. Xu, and Y. Ye. Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization. *SIAM Journal on Scientific Computing*, 32(5):2832–2852, 2010.

X. Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Mathematical Programming*, 134(1):71–99, 2012.

P. Combettes and V. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.

L. S. Correa-Cerro, Y. Piao, A. A. Sharov, A. Nishiyama, J. S. Cadet, H. Yu, L. V. Sharova, L. Xin, H. G. Hoang, M. Thomas, Y. Qian, D. B. Dudekula, E. Meyers, B. Y. Binder, G. Mowrer, U. Bassey, D. L. Longo, D. Schlessinger, and M. S. Ko. Generation of mouse ES cell lines engineered for the forced induction of transcription factors. *Scientific Reports*, 1:167, 2011.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

W. Deng, W. Yin, and Y. Zhang. Group sparse optimization by alternating direction method. Technical report, Rice University, 2011.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(8): 1289–1306, 2006a.

D. L. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete and Computational Geometry*, 35(4):617–652, 2006b.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3): 395–407, 2009.

P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for Kurdyka–Lojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

D. Ge, X. Jiang, and Y. Ye. A note on complexity of L_p minimization. *Mathematical Programming*, 129:285–299, 2011.

R. Ge. A filled function method for finding a global minimizer of a function of several variables. *Mathematical Programming*, 46(1):191–204, 1990.

E. G. Giannopoulou and O. Elemento. Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome Research*, 23(8):1295–1306, 2013.

P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *International Conference on Machine Learning (ICML)*, 2013.

E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A. Sigova, H. A. Hoke, and R. A. Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4): 934–947, 2013.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

Y. Hu, C. Li, and X. Yang. On convergence rates of linearized proximal algorithms for convex composite optimization with applications. *SIAM Journal on Optimization*, 26(2): 1207–1235, 2016.

X. Huang and X. Yang. A unified augmented Lagrangian approach to duality and exact penalization. *Mathematics of Operations Research*, 28(3):533–552, 2003.

M.-J. Lai and J. Wang. An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems. *SIAM Journal on Optimization*, 21(1):82–101, 2011.

M.-J. Lai, Y. Xu, and W. Yin. Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization. *SIAM Journal on Numerical Analysis*, 51(2):927–957, 2013.

G. Li and T. K. Pong. Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming*, 159(1):1–31, 2015a.

G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015b.

G. Li and T. K. Pong. Calculus of the exponent of Kurdyka–Lojasiewicz inequality and its applications to linear convergence of first-order methods. *https://arXiv:1602.02915*, 2016.

Z. Lu. Iterative reweighted minimization methods for lp regularized unconstrained nonlinear programming. *Mathematical Programming*, 147(1):277–307, 2014.

Z.-Q. Luo, J. S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, 1996.

J. Mairal. Optimization with first-order surrogate functions. *International Conference on Machine Learning (ICML)*, 2013.

L. Meier, S. A. van de Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70:53–71, 2008.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.

H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed l_0 norm. *IEEE Transactions on Signal Processing*, 57 (1):289–301, 2009.

B. S. Mordukhovich. *VariationalAnalysis and Generalized Differentiation I: Basic Theory*. Springer, Berlin, 2006.

B. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

A. Nishiyama, L. Xin, A. A. Sharov, M. Thomas, G. Mowrer, E. Meyers, Y. Piao, S. Mehta, S. Yee, Y. Nakatake, C. Stagg, L. Sharova, L. S. Correa-Cerro, U. Bassey, H. Hoang, E. Kim, R. Tapnio, Y. Qian, D. Dudekula, M. Zalzman, M. Li, G. Falco, H. T. Yang, S. L. Lee, M. Monti, I. Stanghellini, M. N. Islam, R. Nagaraja, I. Goldberg, W. Wang, D. L. Longo, D. Schlessinger, and M. S. Ko. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell*, 5: 420–433, 2009.

A. Nishiyama, A. A. Sharov, Y. Piao, M. Amano, T. Amano, H. G. Hoang, B. Y. Binder, R. Tapnio, U. Bassey, J. N. Malinou, L. S. Correa-Cerro, H. Yu, L. Xin, E. Meyers, M. Zalzman, Y. Nakatake, C. Stagg, L. Sharova, Y. Qian, D. Dudekula, S. Sheer, J. S. Cadet, T. Hirata, H. T. Yang, I. Goldberg, M. K. Evans, D. L. Longo, D. Schlessinger, and M. S K.o. Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Scientific Reports*, 3:1390, 2013.

J. Qin, M. J. Li, P. Wang, M. Q. Zhang, and J. Wang. ChIP-Array: Combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Research*, 39:W430–436, 2011.

J. Qin, Y. Hu, F. Xu, H. K. Yalamanchili, and J. Wang. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, 67(3):294–303, 2014.

M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23 (2):1126–1153, 2013.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer-Verlag, Berlin, 1998.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1976.

W. T. Short. Hyperbolic solution of the cubic equation. *National Mathematics Magazine*, 12(3):111–114, 1937.

S. Tao, D. Boley, and S. Zhang. Local linear convergence of ISTA and FISTA on the LASSO problem. *SIAM Journal on Optimization*, 26(1):313–336, 2016.

L. Tian, P. Xu, and Q. Gu. Forward backward greedy algorithms for multi-task learning with faster rates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 735–744, 2016.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1994.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009.

M. Usman, C. Prieto, T. Schaeffter, and P. G. Batchelor. *k-t* Group sparse: A method for accelerating dynamic MRI. *Magnetic Resonance in Medicine*, 66(4):1163–1176, 2011.

S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

E. van den Berg, M. Schmidt, M. P. Friedlander, and K. Murphy. Group sparsity via linear-time projection. Technical report, University of British Columbia, 2008.

J. Wang, Y. Hu, C. Li, and J.-C. Yao. Linear convergence of CQ algorithms and applications in gene regulatory network inference. *Inverse Problems*, to appear, 2017.

S. J. Wright, R. D. Nowak, and M. A T Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

L. Wu, Z. Sun, and D.-H. Li. A gradient based method for the l_2-$l_{1/2}$ minimization and application to compressive sensing. *Pacific Journal of Optimization*, 10(2):401–414, 2014.

L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

D. Xie, A. P. Boyle, L. Wu, J. Zhai, T. Kawli, and M. Snyder. Dynamic trans-acting factor colocalization in human cells. *Cell*, 155(3):713–724, 2013.

Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

Z. Xu, X. Chang, F. Xu, and H. Zhang. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1013–1027, 2012.

H. Yang, Z. Xu, I. King, and M. R. Lyu. Online learning for group Lasso. *International Conference on Machine Learning (ICML)*, 2010.

J. Yang and Y. Zhang. Alternating direction algorithms for ℓ_1-problems in compressive sensing. *SIAM Journal on Scientific Computing*, 33(1):250–278, 2011.

X. Yang and X. Huang. A nonlinear Lagrangian approach to constrained optimization problems. *SIAM Journal on Optimization*, 11(4):1119–1144, 2001.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society: Series B*, 68:49–67, 2006.

J. Zeng, S. Lin, and Z. Xu. Sparse regularization: Convergence of iterative jumping thresholding algorithm. *https://arxiv.org/abs/1402.5744*, 2015.

T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011.

T. Zhang. Some sharp performance bounds for least squares regression with L_1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11:1081–1107, 2010.