Mix Sparse Optimization: Theory and Algorithm

Yaohua Hu[*] Jian Lu[†] Xiaoqi Yang[‡] Kai Zhang[§]

Abstract Structured sparse optimization has been extensively applied in the modeling of many important problems in various disciplines. The mix sparse structure is inherited in a wide class of practical applications, namely, the sparse structure appears as the inter-group and intra-group manners simultaneously. In this paper, we consider the ℓ_0 regularization problem for mix sparse optimization and investigate its mathematical theory and algorithm. In the theoretical part, we first introduce the notions of sparse eigenvalue conditions, one of the weakest regularity conditions in the literature, and then establish the oracle property without any regularity condition and provide a recovery bound for the mix sparse optimization problem under the weak assumption of sparse eigenvalue condition. Moreover, an asymptotic analysis is provided to advance the understanding of the convergence of the ℓ_p regularization to the ℓ_0 regularization. In the algorithmic part, we propose an iterative mix thresholding algorithm with continuation technique (IMTC) to solve the mix sparse optimization problem and present its global convergence theorem and linear convergence rate to a local minimum. The significant advantage of the IMTC is that it has a closed-form expression and low storage requirement, and promotes the mix sparse structure of the solution. Numerical results on simulated data indicate that the IMTC has a strong promoting capability of the mix sparse structure and outperforms several state-of-the-art solvers on both accuracy and robustness, benefiting from the use of the mix sparse structure and the (nonconvex) ℓ_0 regularization. Moreover, we apply the mix sparse optimization to model the differential optical absorption spectroscopy (DOAS) analysis with the wavelength misalignment, and find that the IMTC can exactly and quantitatively predict the existing gases and the factual wavelength misalignment simultaneously within 0.1 second, which may meet the demand of improvement of the DOAS automatic analysis software.

Keywords mix sparse optimization, sparse group Lasso, ℓ_0 regularization, sparse eigenvalue condition, consistency theory, iterative thresholding algorithm, continuation technique

[*]College of Mathematics and Statistics, Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, P. R. China (mayhhu@szu.edu.cn).

[†]College of Mathematics and Statistics, Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Shenzhen University, Shenzhen 518060, P. R. China (jianlu@szu.edu.cn).

[‡]Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (mayangxq@polyu.edu.hk).

[§]Shenzhen Audencia Business School, WeBank Institute of Fintech, Guangdong Laboraty of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen, 518060, P. R. China (kaizhang@szu.edu.cn).

1 Introduction

In the past decade, sparse optimization has become one of the most popular topics in mathematical optimization and gained successful applications in various disciplines, such as compressive sensing [15], image science [5], systems biology [43], and machine learning [3].

Sparse optimization is to find a sparse solution of an underdetermined linear system:

$$b = Ax + \varepsilon,$$

where $A \in \mathbb{R}^{m\times n}$ is a linear transformation matrix, $b \in \mathbb{R}^m$ is an observation vector with an unknown noise $\varepsilon \in \mathbb{R}^m$, and $x \in \mathbb{R}^n$ is the variable to be estimated. In many applications, the dimension of variable is large while the sample size (the number of rows) is small, i.e., $m \ll n$; consequently, the above linear inverse problem is seriously ill-conditioned.

A popular and practical technique for approaching a sparse solution of the linear inverse problem is to solve an associated regularization problem, where a regularizer is used to characterize the sparse structure of the solution. For example, the ℓ_1 regularization problem (also named as Lasso in statistics and basis pursuit in compressive sensing) is the most famous tool for sparse optimization:

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|^2 + \lambda\|x\|_1, \tag{1.1}$$

where $\|x\|_1 := \sum_{i=1}^n |x_i|$ is a sparsity promoting norm, and $\lambda > 0$ is the regularization parameter providing a tradeoff between data fidelity and sparsity.

Benefiting from the convex property, a great deal of attention has been attracted to explore the theoretical property and develop the numerical algorithms for the ℓ_1 regularization problem (1.1). In particular, the nice statistical and theoretical properties, including the oracle property and recovery bound (also named ℓ_2 consistency) of model, have been well studied for the ℓ_1 regularization problem (1.1) under several types of regularity conditions on A, such as the mutual incoherence property (MIP) [11], irrepresentable condition (IC) [36], restricted isometry property (RIP) [12], sparse Riesz condition (SRC) [55] and restricted eigenvalue condition (REC) [7]. Many exclusive and efficient algorithms have been proposed and developed for solving the ℓ_1 regularization problem (1.1), such as the iterative soft thresholding algorithm (ISTA) [18], proximal gradient method (PGM) [5], alternative direction method of multipliers (ADMM) [52], and splitting method [17].

1.1 Structured sparse optimization

Besides the sparse structure, a wide class of application problems usually have certain special structures. In recent years, structured sparse optimization, which adopts the special structure to enhance the structure promoting capability and the sparse recoverability, has been widely applied in the modeling of many important problems in various disciplines. One of the most common structures is the group sparse structure. In this structure, the solution has a natural grouping of its components, and all components within each group are likely to be either zero or nonzero simultaneously. Considering the group sparse structure, Yuan and Lin [54]

proposed an $\ell_{2,1}$ regularization model (named as group Lasso therein) to approach the group variable selection and multi-factor analysis-of-variance problems:

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|^2+\lambda\|x\|_{2,1}, \tag{1.2}$$

where x is of a group structure based on the prior knowledge $x := (x_{\mathcal{G}_1}^\top,\ldots,x_{\mathcal{G}_N}^\top)^\top$, and $\|x\|_{2,1} := \sum_{i=1}^N \|x_{\mathcal{G}_i}\|$ is a group sparsity promoting norm. Employing this group sparsity promoting norm and taking advantage of the group sparsity structure, the $\ell_{2,1}$ regularization problem processes the components in a particular group synchronously, and thus leads to the reduction of freedom degrees in the solution. Consequently, the $\ell_{2,1}$ regularization problem (1.2) is more stable with respect to stochastic noise and results in better recovery performance than the ℓ_1 regularization problem (1.1). Benefitting from these advantages, the group sparse optimization model has been widely applied in practical problems in many fields, such as multi-task machine learning [3], multi-channel Image reconstruction [39] and systems biology [25, 44, 48]. Moreover, the nice statistical and theoretical properties have been well established for the $\ell_{2,1}$ regularization problem (1.2) in [2, 25, 27] and the references therein; many efficient algorithms have been proposed to solve problem (1.2), such as PGM [25], proximal splitting method [53], and proximal point algorithm (PPA) and ADMM [51].

The group sparse optimization model always selects important groups in an "all-in-all-out" fashion, that is, the components within a particular group are either included or not included in the model simultaneously. However, the "all-in-all-out" criterion is too restrictive in many practical applications, that leads to a crucial difficulty in the totally correct definition of the group structure, such as portfolio selection in financial management, supervised learning in machine learning and gene regulatory network inference in systems biology. For example, in portfolio selection [14], investors usually select a few sectors (classified by the types of business) and then identify some most profitable stocks within each sector. In systems biology [22], the gene regulatory network usually has a secondary structure that indicates the regulations of gene clusters by transcription factor (TF) complexes (measured by the cell-specific binding sites) in transcription factories; while in each transcription factory, only a fraction of TFs and genes take functions to regulation network. Therefore, in these problems, the solution always has a mix sparse structure, that is, the sparse structure appears as the inter-group and intra-group manners simultaneously. To deal with the mix sparse structure, Simon et al. [46] introduced a sparse group Lasso (SGL) that blends the $\ell_{2,1}$ and ℓ_1 penalties to achieve sparse effects on the inter-group and intra-group structures, respectively:

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|^2+\lambda\|x\|_{2,1}+\tau\|x\|_1. \tag{1.3}$$

Clearly, SGL (1.3) combines Lasso (1.1) and group Lasso (1.2), and is reduced to Lasso and group Lasso for the special cases when $\lambda = 0$ or $\tau = 0$, respectively. Blending the $\ell_{2,1}$ and ℓ_1 penalties simultaneously, the SGL is more flexible and precise in characterizing the mix sparse structure, and thus arrives at a more stable and accurate optimization solution than Lasso and group Lasso. Benefitting from the mix sparse structure, SGL (1.3) has

gained an increasing amount of successful applications in various disciplines, such as portfolio selection [14], genomic association study [32], visual tracking [59], hyperspectral imaging and differential optical absorption spectroscopy [20]. Furthermore, the statistical consistency theory has been explored for SGL in [32]; several efficient algorithms have been applied to solve SGL, such as fast ISTA [59] and ADMM [20].

However, the ℓ_1 regularization problems of (structured) sparse optimization suffer several frustrations in both theoretical property and practical applications. In particular, it has been revealed by extensive theoretical and empirical studies that the ℓ_1 regularization problem suffers from significant estimation bias when parameters have large absolute values; its induced solutions may be much less sparse than the ground true solution; and it cannot recover a signal or an image with the least measurements when applied to compressive sensing; see, e.g., [13, 21, 25, 44, 26, 50, 56]. Therefore, there is a great demand for developing the alternative (structured) sparse promoting technique that enjoys nice theoretical property and excellent numerical performance.

1.2 Lower-order regularization method

Recently, the lower-order regularization method was developed to overcome these limitations of the ℓ_1-type regularizations. For example, the ℓ_p and $\ell_{2,p}$ regularization problems ($0 < p < 1$) were introduced by [50] and [25] to improve the performance of sparsity recovery and group sparsity recovery, respectively. That is,

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|^2+\lambda\|x\|_p^p \quad \text{and} \quad \min_{x\in\mathbb{R}^n} \|Ax-b\|^2+\lambda\|x\|_{2,p}^p, \tag{1.4}$$

where the sparsity and group sparsity promoting (quasi-)norms are respectively defined by

$$\|x\|_p := (\sum_{i=1}^{n}|x_i|^p)^{\frac{1}{p}} \quad \text{and} \quad \|x\|_{2,p} := (\sum_{i=1}^{N}\|x_{\mathcal{G}_i}\|^p)^{\frac{1}{p}}. \tag{1.5}$$

The ℓ_p-type regularization problems have been found to endow with better theoretical property and numerical performance than the (convex) ℓ_1-type regularization problems. Particularly, it has been shown by many theoretical and empirical studies that the ℓ_p regularization requires a weaker RIP [13] or a weaker REC [25] to guarantee a perfect sparsity recovery or the consistency theory, and allows one to obtain a more sparse solution from fewer linear measurements than that required by the ℓ_1 regularization; the ℓ_p regularization can remove systematic biases and admit a significantly stronger sparsity promoting capability than the ℓ_1 regularization [25, 50]. Benefitting from these advantages, the ℓ_p regularization method has been extensively applied in many fields; see [25, 44, 30, 43, 50] and the references therein. Moreover, it is worth noting that the ℓ_p regularization/penalty method has been investigated for nonlinear constrained optimization problems in [28, 37].

Nevertheless, both the ℓ_1 and ℓ_p ($0 < p < 1$) norms are the relaxations of the sparsity penalty, which have some gap from the original measurement of the sparsity, i.e., the ℓ_0

norm accounts for the cardinality of the support. Indeed, the ℓ_1 and ℓ_p norms penalize the larger components more heavily than the smaller ones, while the ℓ_0 norm forces the unbiased penalization on all components regardless of the magnitude. This is the intrinsical difference between the ℓ_p ($p > 0$) norm and the ℓ_0 norm. By virtue of the exact characterizations of the sparsity and group sparsity, the ℓ_0 and $\ell_{2,0}$ regularization problems were explored in [8, 33] and [40] to approach the sparse solution and group sparse solution, respectively. That is,

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|^2 + \lambda\|x\|_0 \tag{1.6}$$

and

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|^2 + \lambda\|x\|_{2,0}, \tag{1.7}$$

where $\|x\|_0 := \sum_{i=1}^n |x_i|^0$ and $\|x\|_{2,0} := \sum_{i=1}^N \|x_{\mathcal{G}_i}\|^0$ (adopting the convenience that $0^0 = 0$) account for the numbers of nonzero components and nonzero groups, respectively. In despite of its non-continuity and nonconvexity leading to difficulty in designing optimization algorithms, the ℓ_0 regularization provides the most exact sparsity characterization among the lower-order regularizations. However, to the best of our knowledge, there is still no study devoted to investigating the oracle property and consistency theory for the ℓ_0-type regularization problems.

Another major challenge of the lower-order regularization problems is the computational issue. Since the lower-order regularization problem is nonconvex and nonsmooth, it is intractable to find its global solutions. Alternatively, tremendous efforts have been devoted to the development of optimization algorithms for approaching a local minimum or a stationary point of the lower-order regularization problem. Many practical algorithms have been developed to approach the ℓ_0 and ℓ_p regularization problems, such as smoothing method [16], splitting method [31], iterative reweighted minimization method [30], penalty method [35] and difference of convex functions algorithm (DCA) [40]. Besides, one of the most widely applied numerical algorithms for solving the lower-order regularization problems is the class of the iterative thresholding algorithms (ITA), which are of simple formulations and low computational complexity and thus efficient for large-scale structured sparse optimization problems. In particular, iterative hard thresholding algorithm (IHTA) [8] and iterative half thresholding algorithm [50] were proposed to solve the ℓ_0 regularization problem and $\ell_{\frac{1}{2}}$ regularization problem, respectively. The ITA for the lower-order $\ell_{2,p}$ regularization problem was studied in [25] within a unified framework of PGM. Moreover, various variants and extensions of IHTA have been proposed and explored in [4, 9, 29, 34, 58] and the references therein.

The global convergence theorems have been established for the ITA for solving the lower-order regularization problems. The ITA-type algorithms for the ℓ_0 regularization problem are shown to converge to a local minimum of (1.6), and the ones for the ℓ_p regularization problem are proved to converge to a stationary point of (1.4) by virtue of the well-known Kurdyka-Łojasiewicz theory [1].

1.3 Aims of this paper

Inspired by the ℓ_0 regularization method and motivated by the mix sparse structure, we consider the following ℓ_0 regularization problem for mix sparse optimization

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|^2 + \lambda\|x\|_{2,0} + \tau\|x\|_0. \tag{1.8}$$

Clearly, the mix sparse optimization problem (1.8) blends the $\ell_{2,0}$ and ℓ_0 regularizations, and covers the ℓ_0 regularization problem (1.6) and the $\ell_{2,0}$ regularization problem (1.7) as special cases when $\lambda = 0$ or $\tau = 0$, respectively.

In this paper, we will first investigate the oracle property and recovery bound for the mix sparse optimization problem (1.8), which fills the gap of consistency theory for the ℓ_0 regularization problem. For this issue, we will first introduce the notions of sparse eigenvalue conditions (SEC), including the SEC, group SEC and mix SEC, and establish the relationships among them. The SEC [19] is one of the weakest regularity conditions for guaranteeing the nice theoretical property of sparse optimization problems, that is equivalent to the uniquely solvable property. Particularly, sufficient conditions for the SEC-type conditions are provided in terms of the REC-type conditions; see Proposition 2.3. We establish the oracle inequality without any regularity condition on matrix A and provide a recovery bound for the mix sparse optimization problem (1.8) under the weak assumption of mix SEC; see Theorems 2.1 and 2.2. Moreover, by using the notion of epi-convergence in variational analysis [45], we provide an asymptotic approach to validate the epi-convergence of the ℓ_p-type regularizer to the ℓ_0-type regularizer, the continuity of the p-REC to the SEC, and the outer semi-continuity of the solution set of the ℓ_p-type regularization to the one of the ℓ_0-type regularization, as $p \to 0_+$. This helps us to re-establish the oracle property and recovery bound of the ℓ_0-type regularization problem by virtue of the ones of the ℓ_p-type regularization problem in [25]; see Theorem 2.4.

The ITA is one of the most popular and practical numerical algorithms for structured sparse optimization problems, either convex or nonconvex problems. The continuation technique is a widely-applied parameter update strategy for speeding up relevant algorithms in an easily-implemented but efficient way; see, e.g., [24, 23, 29, 49]. Inspired by the ideas of the ITA and the continuation technique, we will propose an iterative mix thresholding algorithm with continuation technique (IMTC) to solve the mix sparse optimization problem (1.8). The IMTC consists of a gradient descent operator, a hard thresholding operator in an individual manner and a group hard thresholding operator in a group manner. The significant advantage of the IMTC is that it has a closed-form expression and low storage requirement, and is able to promote the mix sparse structure of the solution. Hence, the IMTC is particularly attractive and efficient for large-scale strictured sparse optimization problems.

To advance the understanding of our proposed IMTC, we provide a reformulation of the IMTC in terms of proximal operator on mix sparse regularizer and show that it is indeed an application of the PGM (with continuation technique) to solve the mix sparse optimization problem (1.8); see Proposition 3.2. To derive the optimality condition of (1.8) and conver-

gence analysis of the IMTC, we provide the characterization and regular property for the subdifferentials of the ℓ_0 and $\ell_{2,0}$ norms; see Proposition 3.3. In convergence analysis, we prove that the sequence generated by the IMTC globally converges to a local minimum of the mix sparse optimization problem (1.8) at a linear convergence rate under a mild condition on the stepsize; see Theorems 3.2 and 3.3. Moreover, we provide an upper bound for the model error at the limiting point approached by the IMTC, which partially fills a gap between the theoretical and algorithmic studies of the mix sparse optimization problem (1.8). Numerical simulations are conducted to validate the linear convergence rate of the IMTC, to demonstrate the sensitivity analysis on the intra-group sparsity and the inter-group sparsity, and to compare with several state-of-the-art solvers in structured sparse optimization, including FoBa [57], ADMM [52], HalfTA [50], PGM-GSO [25], SGL [46] etc. Numerical results indicate that the IMTC has a strong promoting capability of the mix sparse structure of the solution and outperforms several state-of-the-art solvers on both accuracy and robustness, benefiting from the use of the mix sparse structure and the (nonconvex) ℓ_0 regularization.

The motivation of our work also stems from applications. In particular, we propose a novel mathematical method by virtue of the mix sparse structure for differential optical absorption spectroscopy (DOAS) analysis, which is a fundamental and commonly used technique in atmospheric chemistry and computational optics [41]. DOAS aims to quantify the concentrations of trace gases by measuring specific absorption spectrum from optical spectrometers, such as mass spectrum, ultra-violet spectrum (UV) and infrared spectrum (IR). DOAS technique has been accepted as one of the most powerful methods to measure a wide variety of trace gases, and nowadays, it has been widely applied in air quality monitoring [41], tomography [20], satellite-based monitoring [42] and so on.

One of the central aspects of DOAS technique is the analysis of absorption spectra recorded with instruments. However, traditional DOAS analysis methods suffer three major limitations. (i) Only some specific narrow band absorption structures are used and can only quantify the concentrations of certain types of gases, such as nitrous acid (HNO_2), formaldehyde (CH_2O), and ozone (O_3), nitric oxide (NO), nitrogen dioxide (NO_2), bromine dioxide (BrO), sulfur dioxide (SO_2). By virtue of the full absorption spectra from instruments, DOAS can quantitatively monitor much more types of gases and materials, while a novel mathematical method is required and improved for the analysis of full absorption spectra; see [41]. (ii) Traditional DOAS analysis methods divide the identification of involved gases and the quantification of the concentrations of trace gases into two separate steps, which hinder the development of DOAS automatic analysis software; see [41, Chapter 8]. (iii) In the data collection process from optical spectrometers, there are two major types of noise: the additive noise due to scattering and measurement and the basis noise due to wavelength misalignment. The traditional linear least-squares method is designed to deal with the additive noise but does not consider the wavelength misalignment; see [20]. To overcome these three limitations from practical considerations, we propose a novel mathematical method for DOAS automatic analysis by dealing with the full spectra structures and the wavelength misalignment and executing identification and quantification simultaneously. By enlarging the full spectra

data pool to the one with possible candidates and possible wavelength misalignments, we cast the DOAS analysis problem to a mix sparse optimization model (1.8) by virtue of the Lambert-Beer's law and the mix sparsity structure of the solution. Particularly, we treat the concentrations of all candidates associated to a wavelength misalignment as a group of variables; consequently, the group sparsity of the solution equals to 1 (the misalignment occurs for the mixed gases simultaneously), and the sparsity of the solution equals to the number of existing gases. Numerical results show that the IMTC can exactly and quantitatively predict the existing gases and the factual wavelength misalignment simultaneously within 0.1 second, which meets the demand of improvement of the DOAS automatic analysis software proposed in [41, Chapter 8].

1.4 Notations and organization

The notations adopted in this paper are described as follows. We consider the n-dimensional Euclidean space $\mathbb{R}^n$ with inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|\cdot\|$. As usual, the lowercase letters x, y, z are used to denote the vectors, calligraphic letters $\mathcal{I}$, $\mathcal{J}$, $\mathcal{S}$ denote the index sets. For $x \in \mathbb{R}^n$ and $\mathcal{I} \subseteq \{1,\ldots,n\}$, we use $x_{\mathcal{I}}$ to denote the subvector of x corresponding to $\mathcal{I}$. Let $x := (x_{\mathcal{G}_1}^\top, \ldots, x_{\mathcal{G}_N}^\top)^\top$ represent the group structure, where $x_{\mathcal{G}_i} \in \mathbb{R}^{n_i}$ is the i-th group with $\sum_{i=1}^N n_i = n$. Moreover, we write $n_{\max} := \max_{i=1,\ldots,N} n_i$ and $\mathcal{G}_{\mathcal{S}} := \cup_{i\in\mathcal{S}}\mathcal{G}_i$, and we use $\mathbf{I}$ and A_i to denote the identical matrix in $\mathbb{R}^{n\times n}$ and the i-th column of A, respectively.

The rest of this paper is organized as follows. In Section 2, we introduce the notions of SEC, and use it to establish the oracle property and the recovery bound for the mix sparse optimization problem (1.8), as well as its asymptotic analysis when $p \to 0_+$. In Section 3, we propose an iterative mix thresholding algorithm with continuation technique (IMTC) to solve the mix sparse optimization problem (1.8) and establish its global convergence theorem and linear convergence rate to a local minimum of problem (1.8). In Section 4, we conduct numerical experiments to demonstrate the numerical capacity of the IMTC and compare with several state-of-the-art solvers in structured sparse optimization, and apply the IMTC to deal with the DOAS analysis with the full spectra structures and the wavelength misalignment.

2 Oracle property and recovery bound

The aim of this section is to study the oracle property and the recovery bound for the mix sparse optimization problem (1.8). The oracle property is an important statistical property of the Lasso-type estimators, which provides an upper bound on the prediction of loss function plus the violation of incorrect variable selection. The recovery bound provides an upper bound on the error of the estimated solution, which tells how far is the estimator model from the true model and shows the unbiased property in statistics. In order to explore the oracle property and the recovery bound for problem (1.8), we shall introduce some regularity conditions on the linear transform matrix A.

2.1 Sparse eigenvalue conditions

The concept of sparse eigenvalues was first introduced by Donoho [19] to study the perfect recovery property for Lasso. The s-sparse minimal eigenvalue and the s-sparse maximal eigenvalue of $A^\top A$ are denoted respectively by

$$\varphi(s) := \min_{\|x\|_0 \le s} \sqrt{\frac{x^\top A^\top A x}{x^\top x}} \quad \text{and} \quad \varphi_{\max}(s) := \max_{\|x\|_0 \le s} \sqrt{\frac{x^\top A^\top A x}{x^\top x}}. \tag{2.1}$$

One popular regularity condition, in which the ratio of the s-sparse minimal eigenvalue to the s-sparse maximal eigenvalue is assumed to have a positive lower bound, is used to explore the incoherence property of A, and thus to guarantee the nice statistical property of sparse representation; see [7, 36, 55] and the references therein. Here, we introduce a sparse eigenvalue condition (SEC), which is a weaker variant of the regularity condition and only assumes the s-sparse minimal eigenvalue to be positive.

Definition 2.1 (SEC)**.** *The matrix A is said to satisfy the s-sparse eigenvalue condition (in short, SEC(s)) if $\varphi(s) > 0$.*

Equipping with a pre-defined group structure, we introduce the notions of group sparse eigenvalue condition (GSEC) and mix sparse eigenvalue condition (MSEC). To this end, the group sparse minimal eigenvalue and the mix sparse minimal eigenvalue are respectively defined as follows:

$$\Phi(S) := \min_{\|x\|_{2,0} \le S} \sqrt{\frac{x^\top A^\top A x}{x^\top x}}, \tag{2.2}$$

$$\Gamma(S, s) := \min_{\|x\|_{2,0} \le S,\, \|x\|_0 \le s} \sqrt{\frac{x^\top A^\top A x}{x^\top x}}. \tag{2.3}$$

Definition 2.2 (GSEC/MSEC)**.** *The matrix A is said to satisfy*

(i) *the S-group sparse eigenvalue condition (in short, GSEC(S)) if $\Phi(S) > 0$.*

(ii) *the (S, s)-mix sparse eigenvalue condition (in short, MSEC(S, s)) if $\Gamma(S, s) > 0$.*

It was reported in [25, 36] that the SEC is one of the weakest regularity conditions to ensure the nice property of sparse optimization problems. In the following subsections, we relate the SEC-type conditions with the uniquely solvable property and the weak regularity conditions for the compressed matrix.

2.1.1 Relations between SEC and uniquely solvable property

It is well-known that the SEC($2s$) is a sufficient condition for the uniqueness of the s-sparse solution of the linear system $Ax = b$. To advance the understanding on this unique property of (structured) sparse solution, we introduce the notions of the uniquely solvable-type properties for the matrix A.

Definition 2.3 (USP/GUSP/MUSP)**.** *The matrix A is said to satisfy*

(i) *the uniquely solvable property with s-sparsity (in short, USP(s)) if the linear inverse problem $Ax = b$ has a unique sparse solution satisfying $\|x\|_0 \leq s$ for any $b := Az$ with $\|z\|_0 = s$;*

(ii) *the uniquely solvable property with S-group sparsity (in short, GUSP(S)) if the linear inverse problem $Ax = b$ has a unique group sparse solution satisfying $\|x\|_{2,0} \leq S$ for any $b := Az$ with $\|z\|_{2,0} = S$;*

(iii) *the uniquely solvable property with (S, s)-mix sparsity (in short, MUSP(S, s)) if the linear inverse problem $Ax = b$ has a unique mix sparse solution satisfying $\|x\|_0 \leq s$ and $\|x\|_{2,0} \leq S$ for any $b := Az$ with $\|z\|_0 = s$ and $\|z\|_{2,0} = S$.*

The following proposition shows that the SEC($2s$) (resp., GSEC($2S$) and MSEC($2S, 2s$)) is equivalent to the USP(s) (resp., GUSP(S) and MUSP(S, s)) of A. Hence, if the SEC($2s$) (resp., GSEC($2S$) and MSEC($2S, 2s$)) is not satisfied, one has no hope of recovering the ground true s-sparse solution (resp., S-group sparse solution and (S, s)-mix sparse sollution) from noisy observations.

Proposition 2.1. (i) *The matrix A satisfies SEC($2s$) if and only if it satisfies USP(s).*

(ii) *The matrix A satisfies GSEC($2S$) if and only if it satisfies GUSP(S).*

(iii) *The matrix A satisfies MSEC($2S, 2s$) if and only if it satisfies MUSP(S, s).*

Proof. The proofs of assertions (ii) and (iii) adopt a line of analysis similar to that of assertion (i), we omit the details. It remains to prove assertion (i) as follows.

"$\Rightarrow$": Suppose that A satisfies SEC($2s$). Proving by contradiction, we assume that the USP(s) is not satisfied, that is, there exists $b \in \mathbb{R}^m$ such that the linear inverse problem $Ax = b$ has two distinct s-sparse solutions $\hat{x}$ and $\tilde{x}$; consequently, $A\hat{x} = A\tilde{x}$ and $\|\hat{x}\|_0 = \|\tilde{x}\|_0 \leq s$. Let $x := \hat{x} - \tilde{x} \neq 0$. Then one has that $Ax = 0$ and $\|x\|_0 \leq 2s$, and thus by (2.1) that $\phi(2s) = 0$, which contradicts with SEC($2s$). Hence, A satisfies USP(s), as desired.

"$\Leftarrow$": Suppose that A satisfies USP(s). Proving by contradiction, we assume that the SEC($2s$) is not satisfied, that is, $\varphi(2s) = 0$. Then by (2.1), there exists $x \in \mathbb{R}^n$ with $\|x\|_0 \leq 2s$ such that $Ax = 0$. Divide $\mathrm{supp}(x)$ into two disjoint index sets $\mathcal{I}$ and $\mathcal{J}$ with $|\mathcal{I}| = |\mathcal{J}| \leq s$, and let $b := Ax_{\mathcal{I}}$. Then it is easy to check that $x_{\mathcal{I}}$ and $-x_{\mathcal{J}}$ are two distinct s-sparse solution of the linear inverse problem $Ax = b$, which yields a contradiction with the USP(s). Hence, A satisfies SEC($2s$). The proof is complete. □

2.1.2 Joint sparse eigenvalue condition

The MSEC is not sufficient to guarantee the recovery bound of the mix sparse optimization problem (1.8). Alternatively, we introduce the following joint sparse minimal eigenvalue and

joint sparse eigenvalue condition (JSEC) according to the structure of problem (1.8):

$$\Theta(S,s,\lambda,\tau) := \min_{\lambda\|x\|_{2,0}+\tau\|x\|_0\le\lambda S+\tau s} \sqrt{\frac{x^\top A^\top Ax}{x^\top x}}. \tag{2.4}$$

Definition 2.4 (JSEC)**.** *The matrix A is said to satisfy the (S,s,λ,τ)-joint sparse eigenvalue condition (in short, JSEC(S,s,λ,τ)) if $\Theta(S,s,\lambda,\tau)>0$.*

It is interesting to discuss the relationships between the notions of SECs. It is clear by definition that the JSEC is reduced to the SEC and the GSEC when $\lambda = 0$ and $\tau = 0$, respectively. The following proposition provides some sufficient conditions of JSEC in terms of SEC and GSEC and a necessary condition of JSEC in terms of MSEC.

Proposition 2.2. *The matrix A satisfies JSEC(S,s,λ,τ) provided that one of the following conditions is assumed:*

(i) *A satisfies SEC(s) and GSEC(S);*

(ii) *A satisfies SEC$(s+\frac{\lambda}{\tau}S)$;*

(iii) *A satisfies GSEC$(S+\frac{\tau}{\lambda}s)$.*

Moreover, A satisfies MSEC(S,s) if it satisfies JSEC(S,s,λ,τ). Particularly,

$$\Gamma(S,s)\ge\Theta(S,s,\lambda,\tau)\ge\max\left\{\min\{\varphi(s),\Phi(S)\},\ \varphi(s+\frac{\lambda}{\tau}S),\ \Phi(S+\frac{\tau}{\lambda}s)\right\}. \tag{2.5}$$

Proof. By definitions of the SEC-type conditions, the conclusions of this proposition follow by (2.5). Then it remains to validate (2.5). To do this, associated to the MSEC and the JSEC (cf. Definitions 2.2 and 2.4), we denote the constraint sets by

$$C(S,s) := \{x\in\mathbb{R}^n : \|x\|_{2,0}\le S, \|x\|_0\le s\},$$

and

$$C(S,s,\lambda,\tau) := \{x\in\mathbb{R}^n : \lambda\|x\|_{2,0}+\tau\|x\|_0\le\lambda S+\tau s\},$$

respectively. Then one can easily verify that

$$C(S,s)\subseteq C(S,s,\lambda,\tau)\subseteq C(0,s,0,\tau)\cup C(S,0,\lambda,0);$$

$$C(S,s,\lambda,\tau)\subseteq C(0,s+\frac{\lambda}{\tau}S,0,\tau);$$

$$C(S,s,\lambda,\tau)\subseteq C(S+\frac{\tau}{\lambda}s,0,\lambda,0).$$

By definition, the sparse minimal eigenvalues (i.e., (2.1)-(2.4)) are reduced to

$$\varphi(s) = \min_{x\in C(0,s,0,\tau)} \sqrt{\frac{x^\top A^\top Ax}{x^\top x}};$$

$$\Phi(S) = \min_{x \in C(S,0,\lambda,0)} \sqrt{\frac{x^\top A^\top A x}{x^\top x}};$$

$$\Gamma(S,s) = \min_{x \in C(S,s)} \sqrt{\frac{x^\top A^\top A x}{x^\top x}};$$

$$\Theta(S,s,\lambda,\tau) = \min_{x \in C(S,s,\lambda,\tau)} \sqrt{\frac{x^\top A^\top A x}{x^\top x}}.$$

Thus, by the inclusions of constraint sets mentioned above, we obtain that

$$\Gamma(S,s) \geq \Theta(S,s,\lambda,\tau),$$

$$\Theta(S,s,\lambda,\tau) \geq \min\{\varphi(s), \Phi(S)\},$$

$$\Theta(S,s,\lambda,\tau) \geq \varphi(s + \frac{\lambda}{\tau}S),$$

$$\Theta(S,s,\lambda,\tau) \geq \Phi(S + \frac{\tau}{\lambda}s),$$

respectively. That is, (2.5) is shown to hold, and the proof is complete. □

2.2 Oracle property and recovery bound

In this subsection, we establish the oracle property and the recovery bound for the mix sparse optimization problem (1.8). The following theorem presents the oracle property for problem (1.8), which extends the oracle property for the Lasso [7] and the ℓ_p regularization model ($0 < p \leq 1$) [25] to the ℓ_0 regularization model. Our theorem also improves these results in the sense that the REC-type condition is assumed in [7, Theorem 7.2] and [25, Proposition 8], while Theorem 2.1 does not require any regularity condition on matrix A.

Theorem 2.1. *Let $\bar{x}$ be a solution of $Ax = b$ at a group sparsity level S and a sparsity level s simultaneously, and let Λ and $\mathcal{S}$ denote the index set of nonzero groups and the support of $\bar{x}$, respectively. Let x^* be a global minimum of problem* (1.8)*. Then the following oracle inequality holds:*

$$\|Ax^* - A\bar{x}\|^2 + \lambda\|x^*_{\mathcal{G}_{\Lambda^c}}\|_{2,0} + \tau\|x^*_{\mathcal{S}^c}\|_0 \leq \lambda S + \tau s. \tag{2.6}$$

Proof. Since x^* is a global minimum of problem (1.8) and $\bar{x}$ is a solution of $Ax = b$, one has

$$\|Ax^* - b\|^2 + \lambda\|x^*\|_{2,0} + \tau\|x^*\|_0 \leq \lambda\|\bar{x}\|_{2,0} + \tau\|\bar{x}\|_0. \tag{2.7}$$

Then it follows that

$$\begin{aligned}\|Ax^* - A\bar{x}\|^2 + \lambda\|x^*_{\mathcal{G}_{\Lambda^c}}\|_{2,0} + \tau\|x^*_{\mathcal{S}^c}\|_0 &\leq \lambda\|\bar{x}_{\mathcal{G}_\Lambda}\|_{2,0} - \lambda\|x^*_{\mathcal{G}_\Lambda}\|_{2,0} + \tau\|\bar{x}_{\mathcal{S}}\|_0 - \tau\|x^*_{\mathcal{S}}\|_0 \\ &\leq \lambda\|\bar{x}_{\mathcal{G}_\Lambda} - x^*_{\mathcal{G}_\Lambda}\|_{2,0} + \tau\|\bar{x}_{\mathcal{S}} - x^*_{\mathcal{S}}\|_0 \\ &\leq \lambda S + \tau s,\end{aligned}$$

where the second inequality holds due to the triangle inequalities of the ℓ_0 and $\ell_{2,0}$ norms. The proof is complete. □

The following theorem provides the recovery bound for problem (1.8) under the assumption of SEC-type condition, which is weaker than the REC-type conditions (see Proposition 2.3 below) assumed in the establishment of recovery bound for the Lasso and the ℓ_p regularization model ($0 < p \leq 1$) in [7, Theorem 7.2] and [25, Proposition 9], respectively.

Theorem 2.2. *Suppose that A satisfies JSEC$(2S, 2s, \lambda, \tau)$. Let $\bar{x}$ be a (unique) solution of $Ax = b$ at a group sparsity level S and a sparsity level s simultaneously, and let Λ and $\mathcal{S}$ denote the index set of nonzero groups and the support of $\bar{x}$, respectively. Let x^* be a global minimum of problem* (1.8)*. Then the following recovery bound holds:*

$$\|x^* - \bar{x}\|^2 \leq (\lambda S + \tau s)/\Theta^2(2S, 2s, \lambda, \tau). \tag{2.8}$$

Proof. By assumptions, it follows from (2.7) that

$$\lambda\|x^*\|_{2,0} + \tau\|x^*\|_0 \leq \lambda\|\bar{x}\|_{2,0} + \tau\|\bar{x}\|_0 \leq \lambda S + \tau s.$$

Consequently, one has by the triangle inequality that

$$\begin{aligned}\lambda\|x^* - \bar{x}\|_{2,0} + \tau\|x^* - \bar{x}\|_0 &\leq \lambda(\|x^*\|_{2,0} + \|\bar{x}\|_{2,0}) + \tau(\|x^*\|_0 + \|\bar{x}\|_0)\\ &\leq 2\lambda S + 2\tau s.\end{aligned}$$

By the assumption of JSEC$(2S, 2s, \lambda, \tau)$, we obtain

$$\|x^* - \bar{x}\| \leq \|A(x^* - \bar{x})\|/\Theta(2S, 2s, \lambda, \tau),$$

which, together with (2.6), implies that (2.8). The proof is complete. □

As direct applications of Theorems 2.1 and 2.2, we can derive oracle inequalities and recovery bounds for the ℓ_0 regularization problem (1.6) and the $\ell_{2,0}$ regularization problem (1.7) as order $\mathcal{O}(\lambda s)$ and $\mathcal{O}(\lambda S)$ under the assumption of SEC$(2s)$ and GSEC$(2S)$, respectively. In particular, applied to the case when $\tau = 0$, we obtain the following special case of Theorems 2.1 and 2.2. In Section 2.3.2, we will revisit this result by using an alternative approach.

Corollary 2.1. *Let $\bar{x}$ be a solution of $Ax = b$ at a group sparsity level S, and let Λ denote the index set of nonzero groups of $\bar{x}$. Let x^* be a global minimum of problem* (1.7)*. Then the following assertions are true.*

(i) *The following oracle inequality holds*

$$\|Ax^* - A\bar{x}\|^2 + \lambda\|x^*_{\mathcal{G}_{\Lambda^c}}\|_{2,0} \leq \lambda S. \tag{2.9}$$

(ii) *Suppose that A satisfies GSEC$(2S)$. Then the following recovery bound holds*

$$\|x^* - \bar{x}\|^2 \leq \lambda S/\Phi^2(2S). \tag{2.10}$$

Theorem 2.2 exhibits that the recovery bound

$$\|x^* - \bar{x}\|^2 \leq \mathcal{O}(\lambda S + \tau s)$$

is satisfied for any $b := A\bar{x}$ with $\|\bar{x}\|_0 = s$ and $\|\bar{x}\|_{2,0} = S$ and an arbitrary global solution x^* of the mix sparse optimization problem (1.8) under the JSEC assumption of A. We define this sparse recovery property (SRP) as follows.

Definition 2.5. *Let $\phi : \mathbb{R}^n \to \mathbb{R}_+$ be a regularizer function and denote the associated regularization problem of the linear system $Ax = b$ by*

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|^2 + \lambda\phi(x). \tag{2.11}$$

The regularizer ϕ is said to satisfy

(i) *the sparse recovery property with s-sparsity (in short, SRP(s)) if an arbitrary global solution x^* of the regularization problem (2.11) satisfies $\|x^* - \bar{x}\|^2 \leq \mathcal{O}(\lambda s)$ for any $b := A\bar{x}$ with $\|\bar{x}\|_0 = s$;*

(ii) *the group sparse recovery property with S-group sparsity (in short, GSRP(S)) if an arbitrary global solution x^* of the regularization problem (2.11) satisfies $\|x^* - \bar{x}\|^2 \leq \mathcal{O}(\lambda S)$ for any $b := A\bar{x}$ with $\|\bar{x}\|_{2,0} = S$;*

(iii) *the mix sparse recovery property with (S, s)-mix sparsity (in short, MSRP(S, s)) if an arbitrary global solution x^* of the regularization problem (2.11) satisfies $\|x^* - \bar{x}\|^2 \leq \mathcal{O}(\lambda S + \lambda s)$ for any $b := A\bar{x}$ with $\|z\|_0 = s$ and $\|z\|_{2,0} = S$.*

By Proposition 2.1 and Theorem 2.1, we present in the following corollary the equivalence between SRP (resp., GSRP) and SEC (resp., GSEC); this is because USP is a necessary condition of SRP (if the linear inverse problem $Ax = b$ has two distinct s-solution, SRP must fail for small λ). However, JSEC is only a sufficent condition of MSRP, which suffices MUSP; none of these three properties is equivalent to one of them.

Corollary 2.2. *The following assertions hold.*

(i) *The ℓ_0 norm penalty satisfies SRP(s) if and only if A satisfies SEC($2s$).*

(ii) *The $\ell_{2,0}$ norm penalty satisfies GSRP(S) if and only if A satisfies GSEC($2S$).*

2.3 Relative to ℓ_p regularization as $p \to 0_+$

The lower-order regularization method is also a popular nonconvex regularization technique for (structured) sparse optimization; see [13, 25, 26, 43, 50] and the references therein. Particularly, the lower-order $\ell_{2,p}$ regularization problem ($0 < p < 1$) for group sparse optimization was presented in [25] and is formulated as

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|^2 + \lambda\|x\|_{2,p}^p, \tag{2.12}$$

where the $\ell_{2,p}$ (quasi-)norm is defined by (1.5). In [25], the authors proposed a notion of lower-order group restricted eigenvalue condition (GREC)[1] and used it to investigate the oracle property and the recovery bound for problem (2.12).

We first recall the notion of the GREC and the recovery bound results from [25, Definition 4 and Theorem 9], respectively. To proceed, write $0 \le p \le 1$, $S \le T \ll N$, $\mathcal{J} \subseteq \{1, \ldots, N\}$, and use $\mathcal{J}(x;T)$ to denote the index set of the first T largest groups in value of $\|x_{\mathcal{G}_i}\|$ among $\{\|x_{\mathcal{G}_j}\|,\, j \in \mathcal{J}^c\}$. The $(2,p)$-group restricted minimal eigenvalue is defined as

$$\Phi_p(S,T) := \min\left\{ \frac{\|Ax\|}{\|x_{\mathcal{G}_{\mathcal{J}\cup\mathcal{J}(x;T)}}\|} : |\mathcal{J}| \le S,\ \|x_{\mathcal{G}_{\mathcal{J}^c}}\|_{2,p} \le \|x_{\mathcal{G}_{\mathcal{J}}}\|_{2,p} \right\}. \tag{2.13}$$

Definition 2.6 (GREC)**.** *The matrix A is said to satisfy the $(2,p)$-(S,T)-group restricted eigenvalue condition (in short, $(2,p)$-GREC(S,T)) if $\Phi_p(S,T) > 0$.*

Theorem 2.3. *Let $\bar{x}$ be a solution of $Ax = b$ at a group sparsity level S, and let Λ denote the index set of nonzero groups of $\bar{x}$. Let x^* be a global minimum of problem* (2.12)*. Suppose that the $(2,p)$-GREC(S,S) is satisfied. Then the following assertions are true.*

(i) *The following oracle inequality holds*

$$\|Ax^* - A\bar{x}\|^2 + \lambda \|x^*_{\mathcal{G}_{\Lambda^c}}\|^p_{2,p} \le \lambda^{\frac{2}{2-p}} S / \Phi_p^{\frac{2p}{2-p}}(S,S).$$

(ii) *The following recovery bound holds*

$$\|x^* - \bar{x}\|^2 \le 2\lambda^{\frac{2}{2-p}} S / \Phi_p^{\frac{4}{2-p}}(S,S).$$

The preceding subsections provide the recovery bounds for ℓ_0-type regularization problems under the SEC-type assumptions, while Theorem 2.3 (i.e., [25, Theorem 9]) exhibited the recovery bounds for $\ell_{2,p}$ regularization problems under the $(2,p)$-GREC assumptions as $0 < p < 1$. This subsection aims to provide an asymptotic approach for the recovery bound of the $\ell_{2,0}$ regularization problem (1.7) as in Corollary 2.1 by virtue of the ones for the $\ell_{2,p}$ regularization problems (2.12) in [25] as $p \to 0_+$.

2.3.1 Relation between GSEC and GREC

This section aims to provide a further understanding on the relation between the GSEC and GREC, that is to investigate the continuity of the $(2,p)$-GREC to the GSEC as $p \to 0$.

Let $p \in [0,1]$. Associated to the $(2,p)$-GREC, we denote its constraint set by

$$C_p(S) := \{x \in \mathbb{R}^n : \|x_{\mathcal{G}_{\mathcal{J}^c}}\|_{2,p} \le \|x_{\mathcal{G}_{\mathcal{J}}}\|_{2,p} \text{ for some } \mathcal{J} \text{ satisfying } |\mathcal{J}| \le S\}. \tag{2.14}$$

[1]The lower-order GREC [25] extends the classical restricted eigenvalue condition (REC) [7] to the lower-order and group settings. The classical SEC is weaker than the well-known regularity conditions, such as the mutual incoherence property (MIP) [11] and the restricted isometry property (RIP) [12]; see [7].

By noting that

$$C_0(S) = \{x : \|x\|_{2,0} \leq 2S\}, \tag{2.15}$$

it is clear to derive the following equivalence:

$$\text{GSEC}(2S) \quad \Leftrightarrow \quad \text{0-GREC}(S, S). \tag{2.16}$$

It was revealed in [25, Proposition 5] that the lower-order GREC is less restrictive than the classical GREC (as $p = 1$); particularly, the smaller the $p \in (0, 1]$, the weaker the $(2, p)$-GREC. The following proposition completes the implication when $p \in [0, 1]$ and shows that the GSEC is a weaker condition than the GREC,.

Proposition 2.3. *Let $p \in [0, 1]$. If $(2, p)$-GREC(S, S) holds, then GSEC$(2S)$ holds.*

Proof. Fix $p \in [0, 1]$. By (2.16), the implication to be shown is equivalent to

$$(2, p)\text{-GREC}(S, S) \quad \Rightarrow \quad \text{(2,0)-GREC}(S, S).$$

By Definitions 2.2 and 2.6 and (2.14), it remains to prove that

$$C_0(S) \subseteq C_p(S). \tag{2.17}$$

To this end, let $x \in C_0(S)$, and $\mathcal{J}_*$ denote the index set of the first S largest groups in value of $\|x_{\mathcal{G}_i}\|$. Then it follows from the definition of $C_0(S)$ (2.15) that

$$\|x_{\mathcal{G}_{\mathcal{J}_*^c}}\|_{2,0} \leq \|x_{\mathcal{G}_{\mathcal{J}_*}}\|_{2,0} \leq S. \tag{2.18}$$

By the construction of $\mathcal{J}_*$, one has $\|x_{\mathcal{G}_i}\| \geq \|x_{\mathcal{G}_j}\|$ for each $(i, j) \in \mathcal{J}_* \times \mathcal{J}_*^c$. This, together with (2.18), implies that $\sum_{j\in\mathcal{J}_*^c} \|x_{\mathcal{G}_j}\|^p \leq \sum_{i\in\mathcal{J}_*} \|x_{\mathcal{G}_i}\|^p$; consequently, $\|x_{\mathcal{G}_{\mathcal{J}_*^c}}\|_{2,p} \leq \|x_{\mathcal{G}_{\mathcal{J}_*}}\|_{2,p}$. This says that $x \in C_p(S)$, and hence, (2.17) is proved. The proof is complete. □

It is well-known that the REC is a weaker regularity condition than the RIP and the MIP; see [7]. This, together with Proposition 2.3, claims that the SEC-type conditions are the weakest ones among these regularity conditions.

Proposition 2.3 shows the outer semicontinuity of the $(2, p)$-GREC to the GSEC as $p \to 0_+$. To provide a clear understanding, we below explore the continuity between them. To this end, we recall a notion of epi-convergence from [45]. For $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, the epigraph of f is the set consisting of all the points of $\mathbb{R}^{n+1}$ lying on or above the graph of f; that is,

$$\text{epi}f := \{(x, w) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq w\}.$$

Characterized via the epigraph, the notion of epi-convergence is taken from [45, Definition 7.1].

Definition 2.7. *A sequence of functions $\{f_i : \mathbb{R}^n \to \bar{\mathbb{R}}\}$ is said to epi-converge to f if their epigraphs $\{\text{epi}f_i\}$ graphically converge to epi$f$; namely,*

$$e\text{-}\lim f_i = f \quad \Leftrightarrow \quad \lim_{i\to\infty} \text{epi}f_i = \text{epi}f.$$

By virtue of the concept of epi-convergence and Proposition 2.3, we can show the continuity of the $(2,p)$-GREC to the GSEC as $p \to 0_+$. An interesting result can be derived by Proposition 2.4(ii) that, if A satisfies the GSEC($2S$), then it also satisfies the $(2,p)$-GREC(S,S) for sufficiently small $p > 0$.

Proposition 2.4. *The following assertions hold.*

(i) $C_0(S) = \lim_{p\to 0_+} C_p(S) = \cap_{p>0} C_p(S)$.

(ii) $\lim_{p\to 0_+} \Phi_p(S,S) = \Phi(2S)$.

Proof. (i) It is clear by (2.17) that $C_0(S) \subseteq \cap_{p>0} C_p(S)$. On the other hand, let $x \in \cap_{p>0} C_p(S)$ and use $\mathcal{J}_S$ to denote the index set of the first S largest groups in value of $\|x_{\mathcal{G}_i}\|$. Consequently, by the construction of $\mathcal{J}_S$, one has by (2.14) that

$$\|x_{\mathcal{G}_{\mathcal{J}_S^c}}\|_{2,p}^p \le \|x_{\mathcal{G}_{\mathcal{J}_S}}\|_{2,p}^p \quad \text{for each } p > 0.$$

Noting that $\lim_{p\to 0_+} \|x\|_{2,p}^p = \|x\|_{2,0}$ for each x, we derive that $\|x_{\mathcal{G}_{\mathcal{J}_S^c}}\|_{2,0} \le \|x_{\mathcal{G}_{\mathcal{J}_S}}\|_{2,0} \le S$; that is, $x \in C_0(S)$. Hence, assertion (i) of this proposition is achieved.

(ii) Let $\mathbb{S}$ denote the unit sphere centered at the origin in $\mathbb{R}^n$. Note by the definition of $C_p(S)$ in (2.14) that $C_0(S)$ and $C_p(S)$ are cones and $\|x_{\mathcal{G}_{\mathcal{J}\cup\mathcal{J}(x;S)}}\| \le \|x_{\mathcal{G}_{\mathcal{J}_{2S}}}\|$ for each $|\mathcal{J}| \le S$. Note further by the definitions of $\Phi(S)$ and $\Phi_p(S,S)$ in (2.2) and (2.13) that the objective functions associated to the GSEC and the $(2,p)$-GREC are zero-order homogeneous[2]. Therefore, we derive the reformulations that

$$\Phi(2S) = \min\left\{\frac{\|Ax\|}{\|x_{\mathcal{G}_{\mathcal{J}_{2S}}}\|} : x \in C_0(S)\cap\mathbb{S}\right\} = \min_x \frac{\|Ax\|}{\|x_{\mathcal{G}_{\mathcal{J}_{2S}}}\|} + \delta_{C_0(S)\cap\mathbb{S}}, \tag{2.19}$$

and

$$\Phi_p(S,S) = \min\left\{\frac{\|Ax\|}{\|x_{\mathcal{G}_{\mathcal{J}_{2S}}}\|} : x \in C_p(S)\cap\mathbb{S}\right\} = \min_x \frac{\|Ax\|}{\|x_{\mathcal{G}_{\mathcal{J}_{2S}}}\|} + \delta_{C_p(S)\cap\mathbb{S}}. \tag{2.20}$$

Below, we will show the convergence of minimal values by virtue of the notion of epi-convergence. Indeed, by assertion (i) of this proposition, we obtain by [45, Proposition 7.4(e)] that $e\text{-}\lim_{p\to 0_+} \delta_{C_p(S)\cap\mathbb{S}} = \delta_{C_0(S)\cap\mathbb{S}}$, and then by [45, Exercise 7.8(a)] that

$$e\text{-}\lim_{p\to 0_+} f + \delta_{C_p(S)\cap\mathbb{S}} = f + \delta_{C_0(S)\cap\mathbb{S}}.$$

Moreover, $\{f + \delta_{C_p(S)\cap\mathbb{S}}\}_p$ is proper, lower semicontinuous and level-bounded (since $\mathbb{S}$ is bounded), then [45, Theorem 7.33] is applicable to concluding that

$$\lim_{p\to 0_+} \min_x f(x) + \delta_{C_p(S)\cap\mathbb{S}}(x) = \min_x f(x) + \delta_{C_0(S)\cap\mathbb{S}}(x).$$

This, together with (2.19) and (2.20), implies assertion (ii). The proof is complete. □

[2] f is said to be zero-order homogeneous if $f(\alpha x) = f(x)$ for each $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$.

2.3.2 An asymptotic approach for the recovery bound

It is shown in Corollary 2.1(ii) that the $\ell_{2,0}$ regularization problem (1.7) has a recovery bound of order $\mathcal{O}(\lambda S)$ under the GSEC assumption. This recovery bound can be understood as an asymptotic result of the recovery bounds of the $\ell_{2,p}$ regularization problem (2.12), which are of orders $\mathcal{O}(\lambda^{\frac{2}{2-p}} S)$ (see Theorem 2.3), as $p \to 0_+$. For this purpose, we first show in the following proposition that the convergence in minimization of $\ell_{2,p}$ regularization problem (2.12) to that of $\ell_{2,0}$ regularization problem (1.8) as $p \to 0_+$.

Proposition 2.5. *Let $\lambda > 0$. Then*

$$\limsup_{p\to 0_+} (\arg\min_{x\in\mathbb{R}^n} \|Ax-b\|^2 + \lambda\|x\|_{2,p}^p) \subseteq \arg\min_{x\in\mathbb{R}^n} \|Ax-b\|^2 + \lambda\|x\|_{2,0}.$$

Proof. We will claim that

$$e\text{-}\lim_{p\to 0_+} \|x\|_{2,p}^p = \|x\|_{2,0}. \tag{2.21}$$

Granting this, we obtain by [45, Exercise 7.8(a)] that

$$e\text{-}\lim_{p\to 0_+} \|Ax-b\|^2 + \lambda\|x\|_{2,p}^p = \|Ax-b\|^2 + \lambda\|x\|_{2,0}.$$

Noting further that $\|Ax-b\|^2 + \lambda\|x\|_{2,p}^p$ is proper, lower semicontinuous and level-bounded, [45, Theorem 7.33] is applicable to guaranteeing the conclusion.

To show (2.21), by the equivalent characterization of epi-convergence in [45, Proposition 7.2], it suffices to validate the following relations for each $x \in \mathbb{R}^n$:

$$\begin{cases} \liminf\limits_{i\to\infty,\, p\to 0_+} \|x^i\|_{2,p}^p \geq \|x\|_{2,0} & \text{for each sequence } x^i \to x, \\ \limsup\limits_{i\to\infty,\, p\to 0_+} \|x^i\|_{2,p}^p \leq \|x\|_{2,0} & \text{for some sequence } x^i \to x. \end{cases} \tag{2.22}$$

Let $x \in \mathbb{R}^n$ and consider the following two cases.

Case 1. Let $\{x^i\} \subseteq \mathbb{R}^n$ be such that $\lim_{i\to\infty} x^i = x$. Note that

$$\lim_{p\to 0_+} |t|^p = 1 \quad \text{for each } t \neq 0, \tag{2.23}$$

and $|t|^p \geq 0$ for each $t \in \mathbb{R}$. Then we obtain that

$$\liminf_{i\to\infty,\, p\to 0_+} \|x^i\|_{2,p}^p \geq \|x\|_{2,0}. \tag{2.24}$$

Case 2. Let $\{x^i\} \subseteq \mathbb{R}^n$ be such that $\lim_{i\to\infty} x^i = x$ and $\mathrm{supp}(x^i) = \mathrm{supp}(x)$ for each $i \in \mathbb{N}$. By the construction of $\{x^i\}$ and by (2.23), one can check that

$$\lim_{i\to\infty,\, p\to 0_+} \|x^i\|_{2,p}^p = \|x\|_{2,0}.$$

This, together with (2.24), validates (2.22), as desired. The proof is complete. □

In the following, we re-establish the oracle property and the recovery bound of the $\ell_{2,0}$ regularization problem (1.7) (as in Corollary 2.1) as an asymptotic conclusion of the recovery bound results of the $\ell_{2,p}$ regularization problems (see Theorem 2.3 or [25, Theorem 9]). To be honest, Corollary 2.1 improves Theorem 2.4 in the sense that (i) all global minima of problem (1.7) own the oracle property and recovery bound; (ii) the oracle property does not need the GSEC assumption; and (iii) the recovery bound in (2.10) is tighter than the one below.

Theorem 2.4. *Let $\bar{x}$ be a solution of $Ax = b$ at a group sparsity level S, and let Λ denote the index set of nonzero groups of $\bar{x}$. Suppose that A satisfies GSEC$(2S)$. Then there exists a global minimum x^* of problem* (1.7) *satisfying the following properties.*

(i) *The oracle inequality* (2.9) *holds.*

(ii) *The following recovery bound holds*

$$\|x^* - \bar{x}\|^2 \leq 2\lambda S/\Phi^2(2S).$$

Proof. Suppose that the GSEC$(2S)$ is satisfied. By Proposition 2.4, there exists $p' > 0$ such that the $(2,p)$-GREC(S,S) is satisfied for each $0 < p < p'$. Fix $0 < p < p'$. Hence Theorem 2.3 is applicable; particularly, letting $x^*(p)$ be a global minimum of the the $\ell_{2,p}$ regularization problem (2.12). Thus we have that

$$\|Ax^*(p) - A\bar{x}\|^2 + \lambda\|x^*_{\mathcal{G}_{\Lambda^c}}\|^p_{2,p} \leq \lambda^{\frac{2}{2-p}} S/\Phi_p^{\frac{2p}{2-p}}(S,S), \tag{2.25}$$

$$\|x^*(p) - \bar{x}\|^2 \leq 2\lambda^{\frac{2}{2-p}} S/\Phi_p^{\frac{4}{2-p}}(S,S). \tag{2.26}$$

Noting that $\lim_{p\to 0_+} \|x\|^p_{2,p} = \|x\|_{2,0}$ for each x, one sees that $\{x^*(p)\}$ is bounded and must have a cluster point x^*. Consequently, by applying Proposition 2.5, x^* is a global minimum x^* of the $\ell_{2,0}$ regularization problem (1.7). Since (2.25) and (2.26) hold for arbitrary $0 < p < p'$, taking $p \to 0_+$ and by Proposition 2.4, they are reduced to assertions (i) and (ii), respectively. The proof is complete. □

Remark 2.1. *As a special case when $\max|\mathcal{G}_i| = 1$, group sparse optimization is reduced to sparse optimization. Hence all results of group sparse optimization are true for sparse optimization, e.g., the continuity of p-REC to SEC, the outer semicontinuity of the solution sets of the ℓ_p regularization to the one of the ℓ_0 regularization, and the asymptotic approach of the recovery bounds of the ℓ_p regularization to the one of the ℓ_0 regularization, as $p \to 0_+$.*

3 Iterative mix thresholding algorithm with continuation

In recent years, tremendous efforts have been devoted to the development of optimization algorithms for structured sparse optimization problems. In modern applications of sparse optimization, the number of variables and data are typical of large-scale, and thus drive

the requirement for first-order numerical methods that are of simple formulations and low computational complexity. In particular, the iterative thresholding algorithms (ITA) are a class of popular and efficient first-order numerical methods for sparse optimization. For example, the iterative soft thresholding algorithm (ISTA) [18], iterative hard thresholding algorithm (IHTA) [8] and iterative half thresholding algorithm [50] were proposed to solve the ℓ_1 regularization problem, the ℓ_0 regularization problem and $\ell_{\frac{1}{2}}$ regularization problem, respectively. The ITA for the group sparse optimization problem was studied in [25] within a unified framework of the proximal gradient method (PGM).

The regularization parameter plays an important role in the numerical performance of relevant algorithms. According to the oracle property and the recovery bound theory established in Theorems 2.1 and 2.2 (also see [7, 25]), the regularization parameters λ and τ should be quite small to guarantee the perfect recovery; however, the computational mathematics theory and a great number of numerical experiments show that the quite small parameter will result in the ill-posedness of the subproblems and the convergence rate will be faster if themaintain parameter is properly large. To inherit both advantages in theoretical and numerical aspects, the continuation technique is a widely-applied parameter update strategy for speeding up the relevant algorithm in an easily-implemented but efficient way; see, e.g., [23, 24, 29, 49]. The main idea of the continuation technique uses a decreasing sequence of parameters starting at a pair of large parameters to instead the fixed pair.

Inspired by the ideas of the ITA and the continuation technique, we propose an iterative mix thresholding algorithm with continuation technique (IMTC) to solve the mix sparse optimization problem (1.8) and establish its convergence theory in this section. In particular, by selecting a decreasing sequence of parameter pairs $\{(\lambda_k, \tau_k)\} \searrow (\lambda, \tau)$, the iteration of IMTC consists of a gradient descent operator (3.1), a hard thresholding operator in an individual manner (3.2) and a group hard thresholding operator in a group manner (3.3). Formally, the IMTC is presented as follows.

Algorithm 3.1. Select an initial point $x^0 \in \mathbb{R}^n$, a sequence of stepsizes $\{v_k\} \subseteq (0, +\infty)$, and a decreasing sequence of parameter pairs $\{(\lambda_k, \tau_k)\} \searrow (\lambda, \tau)$. The sequence $\{x^k\} \subseteq \mathbb{R}^n$ are generated via the iterations

$$z^k := x^k - 2v_k A^\top (Ax^k - b), \tag{3.1}$$

$$y^k := \mathbf{H}(z^k; \sqrt{2v_k\tau_k}), \tag{3.2}$$

$$x^{k+1}_{\mathcal{G}_i} := \mathbf{H}_{\mathcal{G}_i}\left(y^k_{\mathcal{G}_i}; \sqrt{2v_k(\lambda_k + \tau_k\|y^k_{\mathcal{G}_i}\|_0)}\right), \quad \text{for } i = 1, \ldots, N, \tag{3.3}$$

where $\mathbf{H} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ and $\mathbf{H}_{\mathcal{G}_i} : \mathbb{R}^{n_i} \times \mathbb{R} \to \mathbb{R}^{n_i}$ are defined respectively by

$$\mathbf{H}(z; \alpha) := \left\{ y \in \mathbb{R}^n \middle| \begin{array}{ll} y_i = z_i, & \text{if } |z_i| > \alpha \\ y_i = 0, & \text{if } |z_i| \le \alpha \end{array} \right\}, \tag{3.4}$$

$$\mathbf{H}_{\mathcal{G}_i}(y; \beta) := \left\{ x \in \mathbb{R}^{n_i} \middle| \begin{array}{l} x = y, \text{ if } \|y\| > \beta \\ x = 0, \text{ if } \|y\| \le \beta \end{array} \right\}. \tag{3.5}$$

In view of Algorithm 3.1, the IMTC has a closed-form expression and low storage requirement; hence it is particularly attractive for the structured sparse optimization problems, especially for large-scale problems. Moreover, the IMTC alternately implements the hard thresholding operators in an individual manner (3.2) and in a group manner (3.3), which is able to promote the mix sparse structure of the solution.

To advance the understanding of the IMTC, we provide a reformulation of the IMTC in terms of proximal operator. For a proper and lower semicontinuous function $f : \mathbb{R}^n \to (-\infty, +\infty]$, its proximal operator $\mathrm{prox}_f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued mapping defined by

$$\mathrm{prox}_f(z) := \arg\min_{x\in\mathbb{R}^n} \{f(x) + \frac{1}{2}\|x - z\|^2\}. \tag{3.6}$$

The following proposition recalls the closed-form expressions of proximal operators of $\tau\|\cdot\|_0$ and $\lambda\|\cdot\|_{2,0}$; see, e.g., [8] and [25, Proposition 18].

Proposition 3.1. (i) *The proximal operator of $\tau\|\cdot\|_0$ has a separable formulation as*

$$(\mathrm{prox}_{\tau\|\cdot\|_0}(z))_i = \begin{cases} z_i, & |z_i| > \sqrt{2\tau}, \\ 0 \text{ or } z_i, & |z_i| = \sqrt{2\tau}, \\ 0, & |z_i| < \sqrt{2\tau}, \end{cases} \quad \textit{for } i = 1, \ldots, n.$$

(ii) *The proximal operator of $\lambda\|\cdot\|_{2,0}$ has a group separable formulation as*

$$(\mathrm{prox}_{\lambda\|\cdot\|_{2,0}}(z))_{\mathcal{G}_i} = \begin{cases} z_{\mathcal{G}_i}, & \|z_{\mathcal{G}_i}\| > \sqrt{2\lambda}, \\ 0 \text{ or } z_{\mathcal{G}_i}, & \|z_{\mathcal{G}_i}\| = \sqrt{2\lambda}, \\ 0, & \|z_{\mathcal{G}_i}\| < \sqrt{2\lambda}, \end{cases} \quad \textit{for } i = 1, \ldots, N.$$

One can see from Proposition 3.1 that the thresholding operators $\mathbf{H}$ and $\mathbf{H}_{\mathcal{G}_i}$ in Algorithm 3.1 are the analytical solutions of proximal operators of $\tau\|\cdot\|_0$ and $\lambda\|\cdot\|_{2,0}$, respectively.

It is well-known that the ISTA and the IHTA could be understood as the applications of the PGM to solve the ℓ_1 regularization problem and the ℓ_0 regularization problem, respectively; see, e.g., [5, 25]. In the following proposition, we show that Algorithm 3.1 is indeed an application of the PGM (with the continuation technique) to solve the mix sparse optimization problem (1.8). More precisely, compounding (3.2) and (3.3) presents a closed-form expression of the proximal operator of the $\ell_{2,0}$ plus the ℓ_0 penalties.

Proposition 3.2. *Let $\{x^k\}$ be a sequence generated by Algorithm 3.1. Then it holds that*

$$x^{k+1} \in \mathrm{prox}_{v_k\lambda_k\|\cdot\|_{2,0}+v_k\tau_k\|\cdot\|_0}(x^k - 2v_kA^\top(Ax^k - b)) \quad \textit{for each } k \in \mathbb{N}.$$

Proof. Fix $k \in \mathbb{N}$. It suffices to show

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \{v_k\lambda_k\|x\|_{2,0} + v_k\tau_k\|x\|_0 + \frac{1}{2}\|x - z^k\|^2\}. \tag{3.7}$$

Note that problem (3.7) is of a group separable structure. The solution of problem (3.7) can be achieved parallelly at each group, and thus it remains to verify the solutions of a cycle of low dimensional proximal optimization subproblems ($i = 1, \ldots, N$):

$$x_{\mathcal{G}_i}^{k+1} \in \arg\min_{x \in \mathbb{R}^{n_i}} \Psi_i(x; z_{\mathcal{G}_i}^k), \tag{3.8}$$

where $\Psi_i : \mathbb{R}^{n_i} \times \mathbb{R}^{n_i} \to \mathbb{R}$ is defined by $\Psi_i(x; y) := v_k\lambda_k\|x\|^0 + v_k\tau_k\|x\|_0 + \frac{1}{2}\|x - y\|^2$.

Fix $i \in \{1, \ldots, N\}$. To validate (3.8), we shall find the minimum of $\Psi_i(\cdot; z_{\mathcal{G}_i}^k)$ over $\mathbb{R}^{n_i} \setminus \{0\}$ and then compare the obtained minimal value with $\Psi_i(0; z_{\mathcal{G}_i}^k)$. For each $x \in \mathbb{R}^{n_i} \setminus \{0\}$, one has

$$\Psi_i(x; z_{\mathcal{G}_i}^k) = v_k\lambda_k + v_k\tau_k\|x\|_0 + \frac{1}{2}\|x - z_{\mathcal{G}_i}^k\|^2,$$

which is the objective function of the proximal operator of $v_k\tau_k\|\cdot\|_0$ (plus a constant). Then by Proposition 3.1(i), we have that a minimum of $\Psi_i(\cdot; z_{\mathcal{G}_i}^k)$ over $\mathbb{R}^{n_i} \setminus \{0\}$ is given by

$$y_j^k = \begin{cases} z_j^k, & \text{if } |z_j^k| > \sqrt{2v_k\tau_k}, \\ 0, & \text{if } |z_j^k| \le \sqrt{2v_k\tau_k}, \end{cases} \quad \text{for each } j \in \mathcal{G}_i, \tag{3.9}$$

which is exactly the same as (3.2). According to (3.9), one can check that

$$\Psi_i(y_{\mathcal{G}_i}^k; z_{\mathcal{G}_i}^k) = v_k\lambda_k + v_k\tau_k\|y_{\mathcal{G}_i}^k\|_0 + \frac{1}{2}(\|z_{\mathcal{G}_i}^k\|^2 - \|y_{\mathcal{G}_i}^k\|^2);$$

while $\Psi_i(0; z_{\mathcal{G}_i}^k) = \frac{1}{2}\|z_{\mathcal{G}_i}^k\|^2$. Hence, by comparing $\Psi_i(y_{\mathcal{G}_i}^k; z_{\mathcal{G}_i}^k)$ and $\Psi_i(0; z_{\mathcal{G}_i}^k)$, we verify that a minimum of $\Psi_i(\cdot; z_{\mathcal{G}_i}^k)$ over $\mathbb{R}^{n_i}$ is given by (3.3). That is, (3.8) is verified, and the proof is complete. □

As a consequence of Proposition 3.2, for the special cases when $\lambda_k \equiv 0$ or $\tau_k \equiv 0$, the IMTC is reduced to the IHTA for the ℓ_0 regularization problem [8] or the PGM for the $\ell_{2,0}$ regularization problem [25] with continuation technique, respectively.

3.1 Subdifferential property

This subsection is devoted to characterizing the subdifferentials of $\|\cdot\|_0$ and $\|\cdot\|_{2,0}$ and their regularity property, which will be useful in characterizing the optimality condition of problem (1.8) and the convergence analysis of the IMTC. The definitions of subdifferentials and subdifferential regularity of nonconvex functions are taken from [45, Definitions 8.3 and 7.25], respectively.

Definition 3.1. *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper and lower semicontinuous function, and let $x \in \mathrm{dom}(f)$.*

(i) *The regular subdifferential of f at x is defined by*

$$\hat{\partial} f(x) := \left\{ u \in \mathbb{R}^n : \liminf_{y \neq x, y \to x} \frac{f(y) - f(x) - \langle u, y - x\rangle}{\|y - x\|} \ge 0 \right\}.$$

(ii) *The (limiting) subdifferential of f at x is defined by*

$$\partial f(x) := \left\{ u \in \mathbb{R}^n : \exists x^k \xrightarrow{f} x, u^k \in \hat{\partial} f(x^k) \text{ with } u^k \to u \right\}.$$

(iii) *The horizon subdifferential of f at x is defined by*

$$\partial^\infty f(x) := \left\{ u \in \mathbb{R}^n : \exists x^k \xrightarrow{f} x, u^k \in \hat{\partial} f(x^k), \kappa^k \downarrow 0, \text{ with } \kappa^k u^k \to u \right\}.$$

f is said to be subdifferentially regular at x if epi$f$ *is Clarke regular at $(x, f(x))$.*

A point x is said to be a stationary point of f if $0 \in \partial f(x)$, which is also a necessary optimality condition of x being a local minimum of f. Below, we characterize the subdifferentials of $\|\cdot\|_0$ and $\|\cdot\|_{2,0}$.

Proposition 3.3. *The functions $\|\cdot\|_0$ and $\|\cdot\|_{2,0}$ are subdifferentially regular at each $x \in \mathbb{R}^n$. Particularly,*

$$\partial \|x\|_0 = \{y \in \mathbb{R}^n : y_i = 0 \text{ for each } i \text{ such that } x_i \neq 0\}, \tag{3.10}$$

$$\partial \|x\|_{2,0} = \{y \in \mathbb{R}^n : y_{\mathcal{G}_i} = 0 \text{ for each } i \text{ such that } \|x_{\mathcal{G}_i}\| \neq 0\}. \tag{3.11}$$

Moreover, if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathbb{R}^n$, then

$$\partial(f + \lambda\|\cdot\|_{2,0} + \tau\|\cdot\|_0)(x) = \nabla f(x) + \partial\|x\|_{2,0} + \partial\|x\|_0. \tag{3.12}$$

Proof. Define an indicator function $c : \mathbb{R} \to \mathbb{R}$ by assigning 0 if $t = 0$, and 1 otherwise. Then the ℓ_0 norm can be reformulated as

$$\|x\|_0 = \sum_{i=1}^{n} c(x_i). \tag{3.13}$$

It follows directly from Definition 3.1 that

$$\hat{\partial} c(t) = \mathbb{R} \text{ if } t = 0, \text{ and } \{0\} \text{ otherwise}, \tag{3.14}$$

and thus one can check that

$$\hat{\partial} c(t) = \partial c(t) = \partial^\infty c(t) = (\hat{\partial} c(t))^\infty \quad \text{for each } t \in \mathbb{R}, \tag{3.15}$$

where the horizon cone of S is defined by $S^\infty := \{x : \exists x^k \in S, \kappa^k \downarrow 0, \text{ with } \kappa^k x^k \to x\}$. Hence, we obtain by [45, Corollary 8.11] that $c(\cdot)$ is subdifferentially regular at each $t \in \mathbb{R}$, and then get the subdifferential regularity of $\|\cdot\|_0$ at each $x \in \mathbb{R}^n$ by the decomposition formulation (3.13) and [45, Proposition 10.5].

For the $\ell_{2,0}$ norm, we define a sequence of group indicator functions $C_i : \mathbb{R}^{n_i} \to \mathbb{R}$ by $C_i(x) = c(\|x\|)$, namely, returning 0 if $x = 0$, and 1 otherwise, for $i = 1, \ldots, N$. Consequently, $\|x\|_{2,0} = \sum_{i=1}^{N} C_i(x_{\mathcal{G}_i})$. Similar to the arguments do for $c(\cdot)$, we obtain that

$$\hat{\partial} C_i(x) = \partial C_i(x) = \partial^\infty C_i(x) = (\hat{\partial} C_i(x))^\infty = \begin{cases} \mathbb{R}^{n_i}, & \text{if } \|x\| = 0, \\ \{0\}, & \text{otherwise}, \end{cases} \tag{3.16}$$

and thus $C_i(\cdot)$ is subdifferentially regular at each $x \in \mathbb{R}^{n_i}$ for $i = 1, \ldots, N$. Hence, it follows from [45, Proposition 10.5] that $\|\cdot\|_{2,0}$ is subdifferentially regular at each $x \in \mathbb{R}^n$.

Finally, by the obtained subdifferential regularity and [45, Proposition 10.5], the formulas (3.10) and (3.11) follow from (3.14) and (3.16), respectively; (3.12) follows from the obtained subdifferential regularity and [45, Corollary 10.9]. The proof is complete. □

3.2 Characterization of local minima

The structure of local minima usually provides an important and useful information for the numerical study of the nonconvex regularized sparse optimization problem; see, e.g., [16, 25, 26, 50]. This subsection aims to study the characterization and structure of the local minimum of problem (1.8). The following theorem provides an equivalent characterization for the local minimum of problem (1.8) in terms of the global minimum of a least squares problem over the subspace associated to its support.

Theorem 3.1. *Let $x^* \in \mathbb{R}^n$ and $X := \{x \in \mathbb{R}^n : \mathrm{supp}(x) \subseteq \mathrm{supp}(x^*)\}$. Then x^* is a local minimum of problem* (1.8) *if and only if x^* is a global optimal solution of*

$$\min_{x \in X} \|Ax - b\|^2. \tag{3.17}$$

Proof. "⇒": Suppose that x^* is a local minimum of problem (1.8). By the optimality condition and the sum rule of subdifferential (3.12), we obtain

$$0 \in \partial F(x^*) = 2A^\top(Ax^* - b) + \lambda \partial \|x^*\|_{2,0} + \tau \partial \|x^*\|_0.$$

Then one has by (3.10) and (3.11) that the following optimality condition holds

$$A_i^\top(Ax^* - b) = 0 \quad \text{for each } i \in \mathrm{supp}(x^*). \tag{3.18}$$

Note that x^* is an (relative) interior point of subspace X and (3.18) is also the optimality condition of problem (3.17). Since (3.17) is a convex optimization problem, we conclude that x^* is a global optimal solution of problem (3.17), as desired.

"⇐": Suppose that x^* is a global optimal solution of problem (3.17). Define

$$\epsilon := \min\left\{ \frac{\tau}{2\|A\|\|Ax^* - b\|}, \min_{i \in \mathrm{supp}(x^*)} |x_i^*| \right\}. \tag{3.19}$$

Then it suffices to show that, for each $x \in \mathrm{B}(x^*, \epsilon)$,

$$F(x) \geq F(x^*). \tag{3.20}$$

Fix $x \in \mathrm{B}(x^*, \epsilon)$. Without loss of generality, we assume that $\mathrm{supp}(x) \neq \mathrm{supp}(x^*)$; otherwise, (3.20) follows directly from the optimality of x^* to (3.17). Noting by (3.19) that

$\epsilon \leq \min_{i\in\mathrm{supp}(x^*)} |x_i^*|$, one has $\mathrm{supp}(x) \supseteq \mathrm{supp}(x^*)$. This, together with the inconsistency of the supports, implies that $\|x\|_0 \geq \|x^*\|_0 + 1$ and $\|x\|_{2,0} \geq \|x^*\|_{2,0}$. Then we obtain

$$\begin{aligned} F(x) - F(x^*) &\geq \|Ax - b\|^2 - \|Ax^* - b\|^2 + \tau \\ &= \langle x - x^*, 2A^\top(Ax^* - b)\rangle + \|A(x - x^*)\|^2 + \tau \\ &\geq -2\epsilon\|A\|\|Ax^* - b\| + \tau \\ &\geq 0 \end{aligned}$$

(thanks to (3.19)). Hence, (3.20) is shown to hold for each $x \in \mathrm{B}(x^*, \epsilon)$, and the proof is complete. □

Without loss of generality, it is always assumed that the rows of A are linearly independent; otherwise, we can remove the redundant rows. As a byproduct of Theorem 3.1, we show in the following proposition that the number of local minima of problem (1.8) with cardinality being no more than the sample size m is finite. We use $\mathrm{LM}(A, \lambda, \tau)$ to denote the set of local minima of problem (1.8) and use $\#(s)$ to denote the set of vectors in $\mathbb{R}^n$ whose cardinality is no more than s; namely,

$$\#(s) := \{x \in \mathbb{R}^n : \|x\|_0 \leq s\}.$$

Proposition 3.4. *Let $\lambda + \tau > 0$. Then $|\mathrm{LM}(A, \lambda, \tau) \cap \#(m)| \leq \sum_{i=1}^m C_n^i$.*

Proof. Let $x^* \in \mathrm{LM}(A, \lambda, \tau)$ and let $I := \mathrm{supp}(x^*)$. Then one has by Theorem 3.1 that x^* is a global optimal solution of (3.17) with $X := \{x \in \mathbb{R}^n : \mathrm{supp}(x) \subseteq I\}$. By the blanket assumption that the rows of A are linearly independent, the global optimal solution (3.17) with $X := \{x \in \mathbb{R}^n : \mathrm{supp}(x) \subseteq I\}$ is unique whenever $|I| \leq m$; particularly,

$$x_I^* = (A_I^T A_I)^{-1} A_I^T b \quad \text{and} \quad x_{I^c}^* = 0.$$

Therefore, the number of local minima of problem (1.8) within $\#(m)$ is no more than the number of index sets with $|I| \leq m$. Hence, we obtain the conclusion, and the proof is complete. □

As corollaries of Proposition 3.4 as $\lambda = 0$ or $\tau = 0$, we can respectively conclude that the numbers of local minima of the ℓ_0 regularized sparse optimization problem (1.6) and the $\ell_{2,0}$ regularized group sparse optimization problem (1.7) with cardinality being no more than the sample size m are finite.

3.3 Convergence theory of IMTC

This subsection aims to establish the convergence theory of the IMTC, including the global convergence theorem and the linear convergence rate. For the remainder of this section, we make the following blanket assumption that

- $\{x^k\}$ is a sequence generated by Algorithm 3.1 with $\{v_k\}$ satisfying

$$0 < \inf_{k\in\mathbb{N}} v_k \le \sup_{k\in\mathbb{N}} v_k < \frac{1}{2\|A\|^2}. \tag{3.21}$$

For the sake of simplicity, we use $F : \mathbb{R}^n \to \mathbb{R}$ to denote the objective function of the mix sparse optimization problem (1.8), namely,

$$F(\cdot) := \|A\cdot - b\|^2 + \lambda\|\cdot\|_{2,0} + \tau\|\cdot\|_0, \tag{3.22}$$

and define a sequence of functions $\{F_k : \mathbb{R}^n \to \mathbb{R}\}$ by

$$F_k(\cdot) := \|A\cdot - b\|^2 + \lambda_k\|\cdot\|_{2,0} + \tau_k\|\cdot\|_0. \tag{3.23}$$

By the continuation technique rule that $\{(\lambda_k, \tau_k)\}$ is a decreasing sequence of positive parameter pairs converging to (λ, τ), one easily gets that

$$F_k(x) \ge F_{k+1}(x) \ge F(x) \quad \text{for each } k \in \mathbb{N} \text{ and } x \in \mathbb{R}^n. \tag{3.24}$$

To establish the convergence theory of the IMTC, we begin with several lemmas that describe the useful properties of the sequences $\{x^k\}$ and $\{F(x^k)\}$. The following lemma provides a lower bound for the absolute value of nonzero entries of $\{x^k\}$, which directly follows from (3.2) in Algorithm 3.1 and (3.21). We write $\mathrm{supp}(x)$ to denote the support of x.

Lemma 3.1. *$|x_j^k| > \sqrt{2\tau \inf_{k\in\mathbb{N}} v_k}$ for each $k \in \mathbb{N}$ and $j \in \mathrm{supp}(x^k)$.*

Recall from Proposition 3.2 that x^{k+1} could be understood as an iterate generated by one PGM step for minimizing F_k (defined by (3.23)) starting at x^k. Hence, the following lemma directly follows from the descent property of the PGM step; see, e.g., [10, Lemma 2].

Lemma 3.2. *$F_k(x^{k+1}) - F_k(x^k) \le -\frac{1}{2}(\frac{1}{v_k} - 2\|A\|^2)\|x^{k+1} - x^k\|^2$ for each $k \in \mathbb{N}$.*

The following lemma shows a decreasing property of $\{F_k(x^k)\}$ and the vanishing property of $\{\|x^{k+1} - x^k\|\}$.

Lemma 3.3. *$\{F_k(x^k)\}$ is decreasing and convergent, and $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$.*

Proof. By Lemma 3.2, we have by (3.24) and (3.21) that

$$F_{k+1}(x^{k+1}) - F_k(x^k) \le F_k(x^{k+1}) - F_k(x^k) \le -\frac{1}{2}(\frac{1}{v_k} - 2\|A\|^2)\|x^{k+1} - x^k\|^2 \le 0,$$

which shows that $\{F_k(x^k)\}$ is decreasing and convergent (as $\{F_k(x^k)\}$ are positive). This also indicates that

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 \le 2\left(\frac{1}{\sup_{k\in\mathbb{N}} v_k} - 2\|A\|^2\right)^{-1} F_0(x^0) < +\infty;$$

consequently, $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$. The proof is complete. □

The following lemma displays a consistent feature of the IMTC that $\{x^k\}$ share a consistent support when k is sufficiently large, which is useful for providing a uniform decomposition of $\{x^k\}$ in convergence analysis.

Lemma 3.4. *There exists $K \in \mathbb{N}$ such that $\mathrm{supp}(x^k) = \mathrm{supp}(x^{k+1})$ for each $k \geq K$.*

Proof. By Lemma 3.3 that $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$, there exists $K \in \mathbb{N}$ such that

$$\|x^{k+1} - x^k\|^2 < 2\tau \inf_{k\in\mathbb{N}} v_k \quad \text{for each } k \geq K. \tag{3.25}$$

Proving by contradiction, we assume, without loss of generality, that there exists $i \in \{1, \ldots, n\}$ such that $x_i^k = 0$ and $x_i^{k+1} \neq 0$. Then one has by Lemma 3.1 that $|x_i^{k+1}|^2 > 2\tau \inf_{k\in\mathbb{N}} v_k$, and thus

$$\|x^{k+1} - x^k\|^2 \geq (x_i^{k+1} - x_i^k)^2 > 2\tau \inf_{k\in\mathbb{N}} v_k,$$

which yields a contradiction with (3.25). The proof is complete. □

The following lemma shows that the IMTC enjoys a descent property of $\{F(x^k)\}$.

Lemma 3.5. *There exists $K \in \mathbb{N}$ such that*

$$F(x^{k+1}) - F(x^k) \leq -\frac{1}{2}(\frac{1}{v_k} - 2\|A\|^2)\|x^{k+1} - x^k\|^2 \quad \textit{for each } k \geq K.$$

Proof. By Lemma 3.4, there exists $K \in \mathbb{N}$ such that $\mathrm{supp}(x^k) = \mathrm{supp}(x^{k+1})$, and thus $\|x^k\|_{2,0} = \|x^{k+1}\|_{2,0}$ and $\|x^k\|_0 = \|x^{k+1}\|_0$, for each $k \geq K$. Then we obtain that

$$F(x^{k+1}) - F(x^k) = \|Ax^{k+1} - b\|^2 - \|Ax^k - b\|^2 = F_k(x^{k+1}) - F_k(x^k).$$

Consequently, the conclusion follows from Lemma 3.2. The proof is complete. □

The main theorem of this subsection is presented as follows, in which we establish the global convergence of the IMTC to a local minimum of the mix sparse optimization problem (1.8).

Theorem 3.2. *Let $\{x^k\}$ be a sequence generated by Algorithm 3.1 with* (3.21) *being satisfied. Then $\{x^k\}$ converges to a local minimum of problem* (1.8).

Proof. By Proposition 3.2, we obtain that, for each $k \in \mathbb{N}$,

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \left\{ \lambda_k \|x\|_{2,0} + \tau_k \|x\|_0 + \frac{1}{2v_k} \|x - (x^k - 2v_k A^\top (Ax^k - b))\|^2 \right\}. \tag{3.26}$$

It follows from Lemma 3.4 that there exists $K \in \mathbb{N}$ such that $\mathrm{supp}(x^k) = \mathrm{supp}(x^{k+1})$ for each $k \geq K$. Fix $k \geq K$ and define a subspace $X := \{x \in \mathbb{R}^n : \mathrm{supp}(x) \subseteq \mathrm{supp}(x^K)\}$. Then one has $x^{k+1} \in X$, and also, $\|x^{k+1}\|_{2,0} \geq \|x\|_{2,0}$ and $\|x^{k+1}\|_0 \geq \|x\|_0$ for each $x \in X$. Hence, (3.26) is reduced to

$$x^{k+1} = \arg\min_{x\in X} \|x - (x^k - 2v_k A^\top (Ax^k - b))\|^2. \tag{3.27}$$

This shows that $\{x^k\}_{k=K}^{\infty}$ is a sequence generated by the projected gradient method for solving the convex constrained least squares problem

$$\min_{x\in X} \|Ax-b\|^2. \tag{3.28}$$

We will show that $\{x^k\}$ converges to a global optimal solution of problem (3.28). Granting this, by Theorem 3.1, we conclude that $\{x^k\}$ converges to a local minimum of problem (1.8), as desired.

To complete the proof, it remains to show that $\{x^k\}$ converges to a global optimal solution of problem (3.28). To this end, we use X^* to denote the set of global optimal solutions of problem (3.28). Fix $k \geq K$ and write $g^k := 2A^\top(Ax^k - b)$. We obtain by the optimality condition of problem (3.27) that

$$\langle x^{k+1} - (x^k - v_k g^k), x - x^{k+1}\rangle \geq 0 \quad \text{for each } x \in X.$$

Fix $x \in X$. Then it follows that

$$\|x - (x^k - v_k g^k)\|^2 \geq \|x^{k+1} - (x^k - v_k g^k)\|^2 + \|x - x^{k+1}\|^2;$$

equivalently,

$$\|x - x^k\|^2 + 2v_k\langle x - x^k, g^k\rangle \geq \|x^{k+1} - x^k\|^2 + 2v_k\langle x^{k+1} - x^k, g^k\rangle + \|x - x^{k+1}\|^2. \tag{3.29}$$

Note by definition of g^k that

$$\langle x - x^k, g^k\rangle = \langle x - x^k, 2A^\top(Ax^k - b)\rangle = \|Ax - b\|^2 - \|Ax^k - b\|^2 - \|Ax - Ax^k\|^2.$$

Substituting this into (3.29), we derive that

$$\begin{aligned}
&\|Ax-b\|^2 + \frac{1}{2v_k}\|x - x^k\|^2 \\
&\geq \|Ax^{k+1} - b\|^2 + \frac{1}{2v_k}\|x - x^{k+1}\|^2 + \frac{1}{2v_k}\|x^{k+1} - x^k\|^2 - \|Ax^{k+1} - Ax^k\|^2 \qquad (3.30)\\
&\geq \|Ax^{k+1} - b\|^2 + \frac{1}{2v_k}\|x - x^{k+1}\|^2
\end{aligned}$$

(due to (3.21)). Note that $x \in X$ is arbitrary. Taking $x := x^k$ in (3.30), one has

$$\|Ax^{k+1} - b\|^2 \leq \|Ax^k - b\|^2 - \frac{1}{2v_k}\|x^k - x^{k+1}\|^2,$$

i.e., $\{\|Ax^k - b\|^2\}$ is decreasing. Taking $x := x^* \in X^*$ in (3.30), we obtain

$$0 \leq \|Ax^{k+1} - b\|^2 - \|Ax^* - b\|^2 \leq \frac{1}{2v_k}(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2). \tag{3.31}$$

This shows that $\{\|x^k - x^*\|^2\}$ is decreasing for each $x^* \in X^*$, and so, $\{x^k\}$ is bounded and must have a cluster point $\bar{x}$. By the decreasing property of $\{\|Ax^k - b\|^2\}$, we inductively obtain by (3.31) and (3.21) that

$$\|Ax^k - b\|^2 - \|Ax^* - b\|^2 \leq \frac{1}{k}\sum_{i=1}^{k}(\|Ax^i - b\|^2 - \|Ax^* - b\|^2) \leq \frac{1}{2k\inf_{i\in\mathbb{N}} v_i}\|x^0 - x^*\|^2.$$

By letting $k \to \infty$, we have $\|A\bar{x} - b\|^2 = \|Ax^* - b\|^2$, and thus $\bar{x} \in X^*$. Since $\{\|x^k - \bar{x}\|^2\}$ is decreasing and $\bar{x}$ is a cluster point of $\{x^k\}$, then $\{x^k\}$ converges to such $\bar{x} \in X^*$. The proof is complete. □

In addition to the global convergence theorem, the establishment of convergence rate is important to guarantee the numerical performance of relevant algorithms. We establish the linear convergence rate of the IMTC in the following theorem.

Theorem 3.3. *Let $\{x^k\}$ be a sequence generated by Algorithm 3.1 with* (3.21) *being satisfied. Then $\{x^k\}$ converges linearly to a local minimum of problem* (1.8).

Proof. Theorems 3.2 shows that $\{x^k\}$ globally converges to a local minimum x^* of problem (1.8). Hence, the optimality condition (3.18) is satisfied. Moreover, by Lemmas 3.1 and 3.4, there exists $K \in \mathbb{N}$ such that

$$\operatorname{supp}(x^k) = \operatorname{supp}(x^*) \quad \text{for each } k \geq K. \tag{3.32}$$

Fix $k \geq K$. In view of Algorithm 3.1 and by (3.18), we obtain that, for each $i \in \operatorname{supp}(x^*)$,

$$\begin{aligned} x_i^{k+1} - x_i^* &= x_i^k - 2v_k A_i^\top (Ax^k - b) - x_i^* \\ &= x_i^k - x_i^* - v_k A_i^\top ((Ax^k - b) - (Ax^* - b)) \\ &= (\mathbf{I}_i^\top - 2v_k A_i^\top A)(x^k - x^*). \end{aligned}$$

Then one has by (3.32) that

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \sum_{i\in\operatorname{supp}(x^*)} (x_i^{k+1} - x_i^*)^2 \\ &= \sum_{i\in\operatorname{supp}(x^*)} \Big((\mathbf{I}_i^\top - 2v_k A_i^\top A)(x^k - x^*)\Big)^2 \\ &\leq \|\mathbf{I} - 2v_k A^\top A\|^2 \|x^k - x^*\|^2. \end{aligned}$$

By assumption (3.21), there exists $\eta \in (0, 1)$ such that $\|\mathbf{I} - 2v_k A^\top A\| < \eta$. This, together with the above inequality, shows the linear convergence rate of $\{x^k\}$ to x^*. The proof is complete. □

There is still a gap between the theoretical and algorithmic studies of the mix sparse optimization problem (1.8). In particular, Theorems 3.2 and 3.3 present the linear convergence

rate of the IMTC to a local minimum of problem (1.8); while Theorems 2.1 and 2.2 provide the oracle property and the recovery bound for the global minimum of problem (1.8). To fill this gap in some sense, we end this section by the following theorem, which provides an upper bound for the model error at the limiting point approached by the IMTC. Indeed, the upper bound (3.34) has the same order of $\mathcal{O}(\lambda+\tau)$ as the ones in the oracle inequality (2.6) and the recovery bound (2.8). Recall that n_i is the size of the i-th group and $n_{\max} := \max_{i=1,\ldots,N} n_i$.

Theorem 3.4. *Let $\{x^k\}$ be a sequence generated by Algorithm 3.1 with* (3.21) *being satisfied. Let x^* be the limiting point of $\{x^k\}$. Suppose that*

$$\psi(A) := \inf_{u\in\mathbb{R}^m} \frac{\|A^\top u\|_\infty^2}{\|u\|^2} > 0^3. \tag{3.33}$$

Then

$$\|Ax^* - b\|^2 \le \frac{\lambda + \tau n_{\max}}{2\psi(A)\inf_{k\in\mathbb{N}} v_k}. \tag{3.34}$$

Proof. Since x^* is the limiting point of $\{x^k\}$, in view of Algorithm 3.1 and by (3.21), there exists $v \in [\inf_{k\in\mathbb{N}} v_k, \sup_{k\in\mathbb{N}} v_k]$ such that

$$z^*_{\mathcal{G}_i} = x^*_{\mathcal{G}_i} - 2vA^\top_{\mathcal{G}_i}(Ax^* - b),\ y^*_{\mathcal{G}_i} = \mathbf{H}(z^*_{\mathcal{G}_i}, \sqrt{2v\tau}),\ x^*_{\mathcal{G}_i} = \mathbf{H}_{\mathcal{G}_i}\left(y^*_{\mathcal{G}_i}; \sqrt{2v(\lambda+\tau\|y^*_{\mathcal{G}_i}\|_0)}\right) \tag{3.35}$$

for $i = 1,\ldots,N$. Let $\Lambda := \{i : \|x^*_{\mathcal{G}_i}\| \neq 0\}$ and Λ^c be its complement set. We first claim that

$$\|A^\top_{\mathcal{G}_i}(Ax^* - b)\|^2 \le \frac{\lambda + \tau n_i}{2\inf_{k\in\mathbb{N}} v_k} \quad \text{for each } i \in \Lambda^c, \tag{3.36}$$

and

$$(A_j^\top(Ax^* - b))^2 \le \frac{\tau}{2\inf_{k\in\mathbb{N}} v_k} \quad \text{for each } j \in \mathcal{G}_\Lambda \setminus \mathrm{supp}(x^*). \tag{3.37}$$

To show (3.36), we fix $i \in \Lambda^c$. This says $x^*_{\mathcal{G}_i} = 0$. Consequently, we derive by the first two equalities in (3.35) and the definition of $\mathbf{H}$ in (3.4) that

$$\|2vA^\top_{\mathcal{G}_i}(Ax^* - b)\|^2 = \|z^*_{\mathcal{G}_i}\|^2 \le \|y^*_{\mathcal{G}_i}\|^2 + 2v\tau(n_i - \|y^*_{\mathcal{G}_i}\|_0), \tag{3.38}$$

and by the last equality of (3.35) and the definition of $\mathbf{H}_{\mathcal{G}_i}$ in (3.5) that

$$\|y^*_{\mathcal{G}_i}\|^2 \le 2v(\lambda + \tau\|y^*_{\mathcal{G}_i}\|_0).$$

This, together with (3.38), deduces (3.36).

To prove (3.37), fix $j \in \mathcal{G}_i \setminus \mathrm{supp}(x^*)$ with $i \in \Lambda$. This says that $x^*_j = 0$ and $x^*_{\mathcal{G}_i} \neq 0$. Then we obtain by (3.35) that $|2vA_j^\top(Ax^* - b)| = |z^*_j| \le \sqrt{2v\tau}$, and thus (3.37) is achieved.

By (3.18), (3.36) and (3.37), we conclude that

$$\|A^\top(Ax^* - b)\|_\infty^2 \le \frac{\lambda + \tau n_{\max}}{2\inf_{k\in\mathbb{N}} v_k}.$$

Then we obtain by assumption (3.33) (with $Ax^* - b$ in place of u) that

$$\|Ax^* - b\|^2 \le \frac{1}{\psi(A)}\|A^\top(Ax^* - b)\|_\infty^2 \le \frac{\lambda + \tau n_{\max}}{2\psi(A)\inf_{k\in\mathbb{N}} v_k}.$$

The proof is complete. □

[3] A testable sufficient condition for (3.33) is that A is of full row rank and each $\|A_i\| > 0$.

4 Numerical experiments

In this section, we carry out numerical experiments to illustrate the numerical performance of the IMTC, and compare with several state-of-the-art solvers in structured sparse optimization. All numerical experiments are implemented in Matlab R2014a and executed on a personal desktop (Intel Core Duo i7-8550, 1.80 GHz, 8.00 GB of RAM).

4.1 Numerical simulations

In the numerical experiments, the simulation data are generated via the standard process of compressive sensing with the mix sparse solution. In details, we randomly generate an i.i.d. Gaussian ensemble $A \in \mathbb{R}^{m\times n}$ satisfying $A^\top A = I$, and a mix sparse solution $\bar{x} \in \mathbb{R}^n$ via randomly splitting its components into N equi-size groups and randomly picking S of them as nonzero groups, each of which is set as an r-sparse vector with an i.i.d. Gaussian ensemble. Consequently, the inter-group sparsity of the solution is $\frac{S}{N}$, and the intra-group sparsity is $\frac{rN}{n}$. We define

$$\gamma := \frac{r}{S},$$

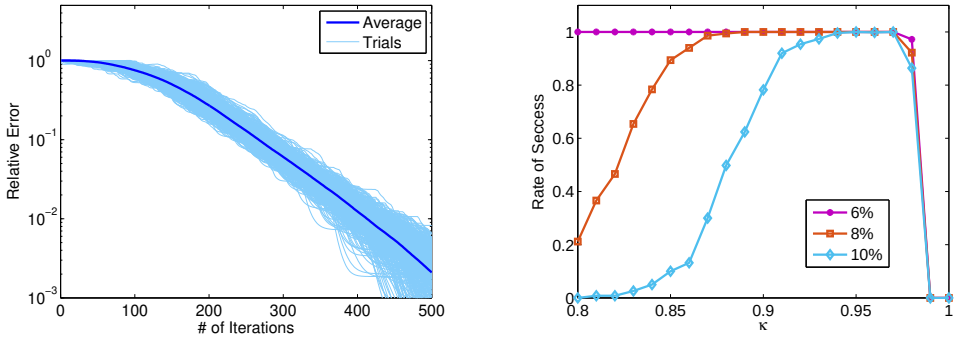
which is proportional to the ratio of the intra-group sparsity to inter-group sparsity when the variable dimension and the group size are fixed. Given the transform matrix A and solution $\bar{x}$, the observation data b is generated via the Matlab script

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

where *sigma* is the standard deviation of additive Gaussian noise. In the numerical experiments, the problem size is set as $m = 256$ and $n = 1024$, the number of groups $N = 64$ (unless otherwise specified), and the noise level is set as $sigma = 0.1\%$.

In the implementation of the IMTC (cf. Algorithm 3.1), the initial point and parameters are selected as follows. The initial point is $x^0 := 0$, and we set the constant stepsize $v := 0.5$, regularization parameters $(\lambda, \tau) := (10^{-4}, 10^{-6})$ and the decreasing sequence of parameter pairs in the continuation technique as $(\lambda_k, \tau_k) = (\max\{\kappa\lambda_{k-1}, \lambda\}, \max\{\kappa\tau_{k-1}, \tau\})$ with $\kappa := 0.96$. Two key criteria to measure the performance of the solvers are the relative error $\mathrm{RE} = \frac{\|x-\bar{x}\|}{\|\bar{x}\|}$ and the successful recovery rate, where the recovery is regarded as *success* if $\mathrm{RE} < 0.5\%$; otherwise, it is regarded as *failure*. Five numerical experiments will be performed, in each of which we will conduct 500 trials with randomly simulated data (unless otherwise specified) to show the efficiency and stability of the IMTC, comparing with several state-of-the-art solvers.

The first experiment aims to validate the linear convergence rate of the IMTC by conducting extensive simulations. Figure 1 plots the relative error of the estimation along the number of iterations in 500 random trials for different sparsity levels 5% and 10%, respectively. As shown in Figure 1, the IMTC can successfully recover the ground true solution and stably behave a linear convergence rate, which is consistent with Theorem 3.3. In this



(a) Sparsity level 10%. (b) Continuation parameter.

Figure 1: Linear convergence of IMTC.

experiment, we also note that the IMTC is very fast, whose CUP time is about 0.1 second per 500 iterations.

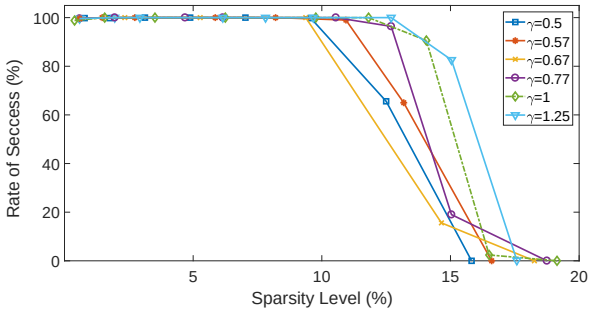


Figure 2: Sensitivity analysis on intra- and inter-group sparsity.

The second experiment is to illustrate the sensitivity analysis of the IMTC on the intra-group sparsity and the inter-group sparsity by varying the ratio γ from 0.5 to 1.25. Figure 2 draws the successful recovery rate of the IMTC along the sparsity level for different γ. It indicates that the IMTC can achieve a high successful recovery rate in the scenario of mix sparse structure, and particularly, the IMTC performs better for a larger γ, e.g., when $\gamma \geq 1$.

The third experiment is implemented to compare the IMTC with several state-of-the-art solvers in structured sparse optimization. The solvers can be divided into three types of structures: the sparse solvers, the group sparse solvers, and the mix sparse solvers. The sparse solvers include YALL1 [52], ADMM (ℓ_0), OMP [47], CoSaMP [38], SPGL1 [6], FoBa [57], HalfTA [50], HardST [8] and ISTA [18]; the group sparse solvers include PGM-GSO for different $\ell_{p,q}$ regularizations [25], ADMM ($\ell_{2,1}$) [51]; the mix sparse solvers include SGL [46], ADMMSGL [20] and PGASGL [59].

In this experiment, we first illustrate in Figure 3 the structures of the signals obtained by these solvers in a random trial with $S = 15$ and $r = 8$ (i.e., the total sparsity level is

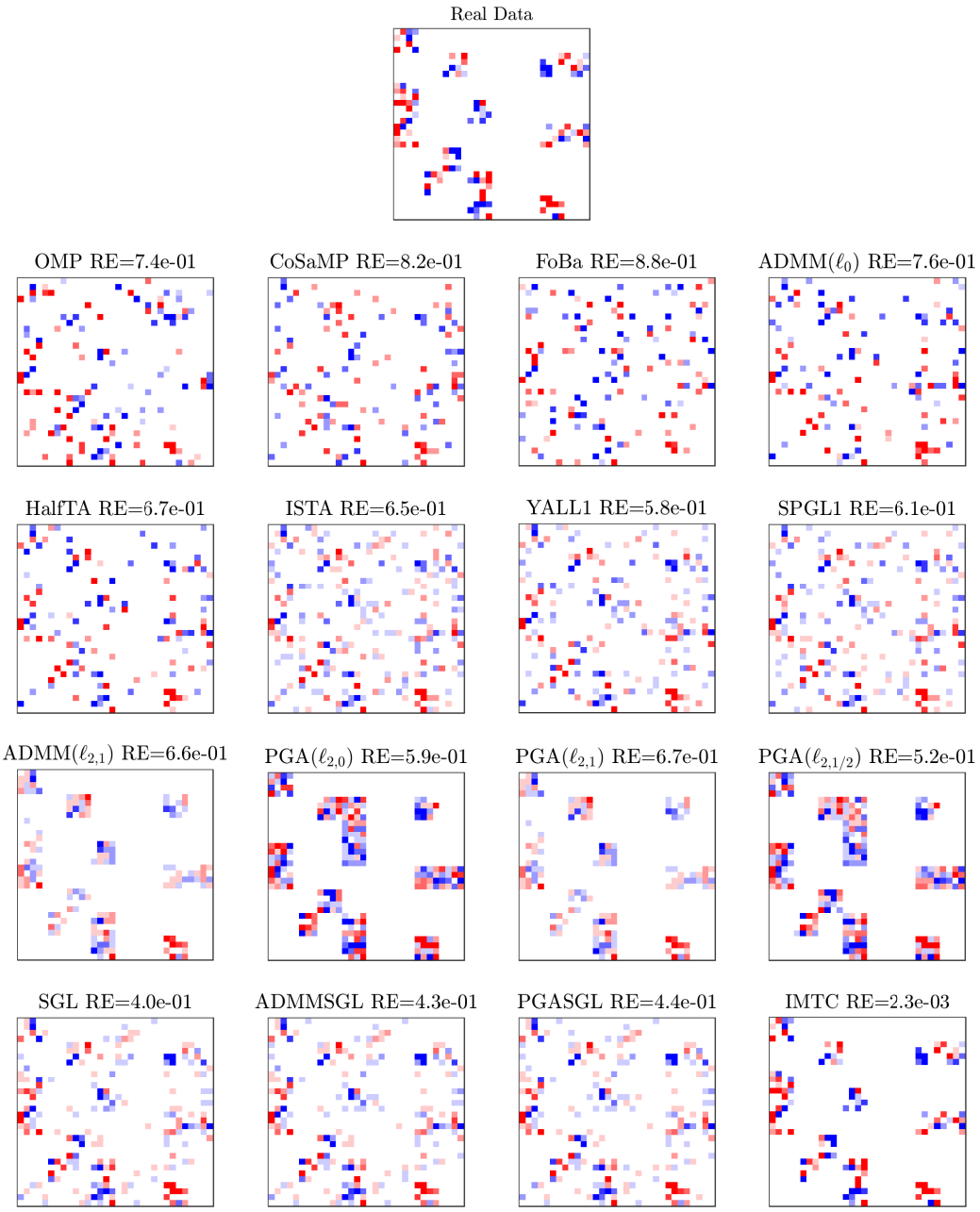
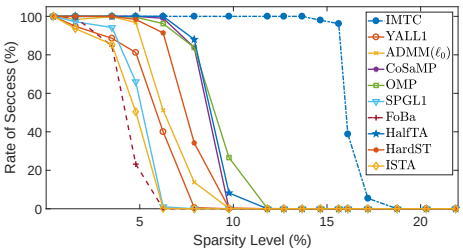
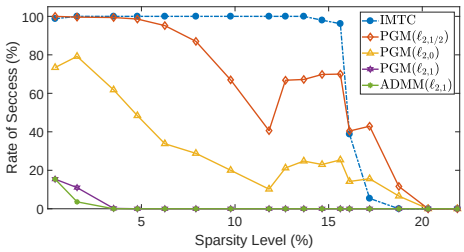


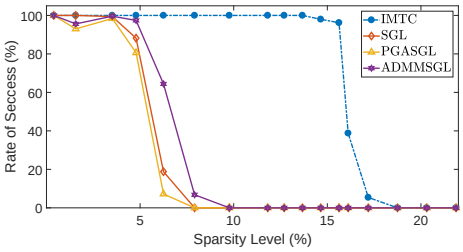
Figure 3: Structures obtained by IMTC and several state-of-the-art solvers.



(a) Compare with sparse solvers.



(b) Compare with group sparse solvers.

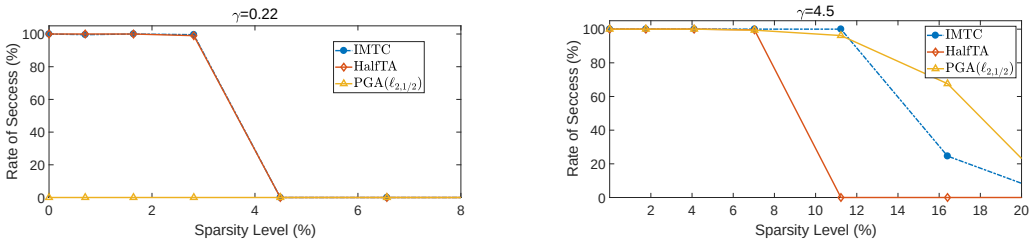


(c) Compare with mix sparse solvers.

Figure 4: Comparison between IMTC and several state-of-the-art solvers.

12%). We reshape the signal to a 32×32 matrix for convenience of viewing, in which a group structure is denoted by a 4×4 sub-matrix. In Figure 3, the top picture plots the heat-map of the ground true signal, the second and third (resp., the fourth and the last) rows draw the solutions obtained by the sparse (resp., group sparse and mix sparse) solvers. We can clearly observe from Figure 3 the structures of the solutions approached by these three types of solvers. (i) The sparse solvers obtain the sparse solutions without any group structure; (ii) the group sparse solvers get the sparse solutions with group structure, in which the components within each group are either all zeros or all nonzeros; and (iii) the mix sparse solvers achieve the solutions with the intra-group sparse and inter-group sparse structures simultaneously. In particular, the IMTC performs the best in approaching the ground true signal (RE$\sim$ 2e-3) among the mix sparse solvers.

Furthermore, we compare the overall performance of the IMTC and these three types of solvers by conducting extensive simulations with mix sparse structure. Figure 4(a)-(c) plot the successful recovery rates on different sparsity levels compared with the sparse, group sparse and mix sparse solvers, respectively. Figure 4 indicates that the IMTC can achieve a higher successful recovery rate than the other solvers (with or without mix sparse structure), by exploiting the mix sparse structure and (nonconvex) ℓ_0 regularization. In Figure 4(b), it should be pointed out that the group sparse solvers perform unstable because of the high sensitivity on γ (cf. Figure 5). From this experiment, it is revealed that the IMTC outperforms most solvers of structured sparse optimization in solving sparse optimization problems with mix sparse structure.



(a) Sparse solution without group structure.

(b) Group sparse solution.

Figure 5: Performance of IMTC at two extreme situations.

The forth experiment is to explore the performance of the IMTC at two extreme situations: (a) γ is very small ($\gamma = 0.22$), in which case the solution is sparse but has no group structure (there is only one nonzero component in each nonzero group); and (b) γ is reasonably large ($\gamma = 4.5$), in which case the solution is of group sparse structure (the components are all nonzeros in each nonzero group). One representative is selected from each type of structured sparse solvers: HalfTA for the sparse solvers, PGA ($\ell_{2,1/2}$) for the group sparse solvers, and IMTC for the mix sparse solvers. It is illustrated by Figure 5(a) that the IMTC works as well as the HalfTA in the situation of sparse solution (without group structure), while the PGA ($\ell_{2,1/2}$) fails in this situation because the group sparse solvers always induce a solution with group structure. It is observed from Figure 5(b) that the IMTC works almost as well as the PGA ($\ell_{2,1/2}$), and outperforms the HalfTA in the situation of group sparse solution. This is because that the sparse solvers (without group or mix structure) are not able to promote the group structure of the solution. From this experiment, it is displayed that the IMTC performs well and stably at these two extreme situations of sparse or group sparse structure.

The last experiment is devoted to the phase transition study of the IMTC and to further demonstrate its strong promoting capability of mix sparse structure. The phase diagram is a popular and powerful tool in studying the stability and sensitivity of the relevant solvers; see [25, 50] and the references therein. In this experiment, we set $m = 256$, $n = 1024$, $N = 64$ and $sigma = 0.1\%$ as default, and vary the inter-group sparsity and the intra-group sparsity from [0,1]. For each circumstance of mix sparse structure, we randomly generate the simulation data 500 times and apply the relevant solvers to approach the solution. Each pixel in the phase diagram is embodied as yellow whenever its successful recovery rate is 100%; otherwise, blue when the successful recovery rate is 0. In this way, the phase diagrams of six representative solvers, including IMTC, HalfTA, PGA ($\ell_{2,1/2}$), PGASGL, ADMMSGL and SGL, are exhibited in Figure 6, in which the color of each cell reflects the empirical recovery rate (scaled between 0 and 1) and the red curve denotes the contour line of total sparsity 15%.

Observed from Figure 6, we find that the phase transition phenomenon does appear for these six solvers of structured sparse optimization. It is displayed in Figure 6 that the yellow area of the IMTC is closest to the red curve and is much larger than the other solvers'. This indicates that the IMTC is more robust (on the inter- and intra-group sparsity simultane-

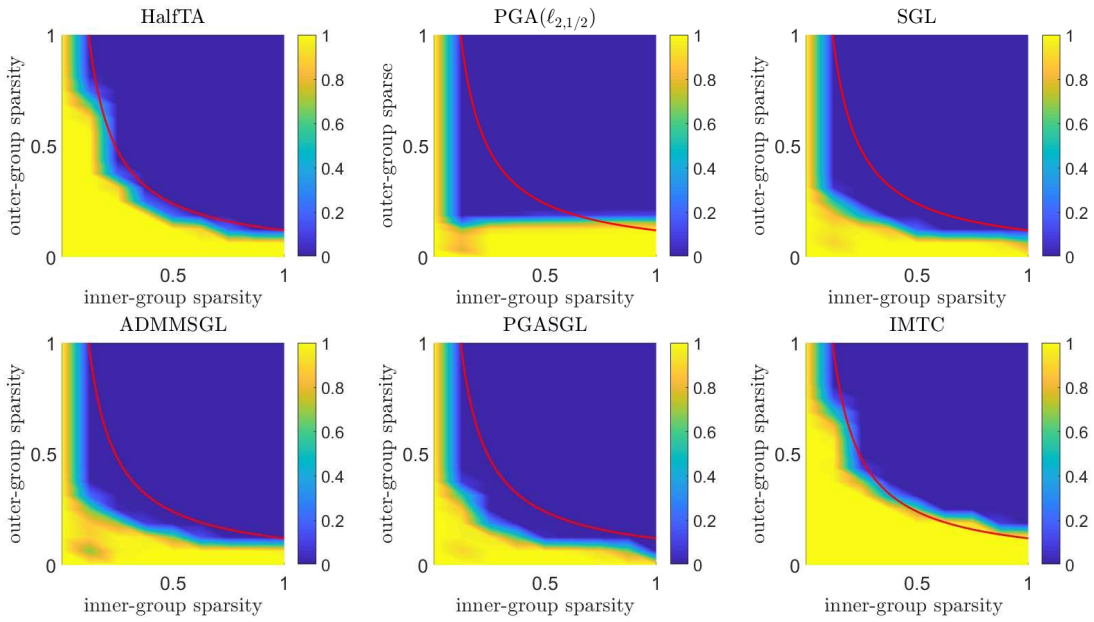


Figure 6: Phase diagram study of structured sparse optimization solvers. Yellow and blue denote success and failure in all experiments, respectively.

ously) than the other solvers in the sense that it allows to reach a higher successful recovery rate, which achieves the same conclusion as the third experiment.

In conclusion, the numerical experiments with simulated data show that the IMTC has a strong promoting capability of the mix sparse structure and outperforms several state-of-the-art solvers on both accuracy and robustness, benefiting from the use of the mix sparse structure and (nonconvex) ℓ_0 regularization.

4.2 Differential optical absorption spectroscopy

Differential optical absorption spectroscopy (DOAS) analysis is a fundamental and commonly used technique in atmospheric chemistry and computational optics [41]. The goal of DOAS is to identify the gases or materials in a mixture and quantify their concentrations by measuring the reduction over a range of wavelengths in terms of the intensity of light shined through it. When the light shines through a gas with path length L and concentration $x(\cdot)$, the light intensity will be absorbed and attenuated; Lambert-Beer's law [41, Chapter 8] relates the process of light absorption as

$$A(\omega) = \log \frac{I_0(\omega)}{I(\omega)} = \theta(\omega) \int_0^L x(l)dl \tag{4.1}$$

where ω is the wavelength, $A(\cdot)$, $I(\cdot)$ and $I_0(\cdot)$ are the absorbed, transmitted and the initial intensity of the light, respectively, and $\theta(\cdot)$ is the characteristic absorption spectra for the absorbing gas. For simplicity, assuming that the gas is equable at the path line, the

concentration is a constant x, and then Lambert-Beer's law (4.1) is reduced to

$$A(\omega) = Lx\theta(\omega). \tag{4.2}$$

When multiple absorbing gases $\mathcal{M}$ are present at the path line and the reduction of the light intensity due to scattering is considered, $x\theta(\cdot)$ can be replaced by a linear combination of the characteristic absorption spectra of involved gases plus an additive noise of light scattering, and hence Lambert-Beer's law (4.2) can be rewritten as

$$A(\omega) = L \sum_{i \in \mathcal{M}} x_i \theta_i(\omega) + \epsilon(\omega). \tag{4.3}$$

Collected the reference characteristic absorption spectra $\{\theta_i\}$ from database and measured the absorption spectra of the mixture of gases A from the optical spectrometer, the DOAS analysis aims to estimate the concentrations of gases $\{x_i\}_{i \in \mathcal{M}}$ via (4.3). Considering only the involved gases and dealing with the additive white Gaussian noise ϵ, one of the most classical and common mathematical methods for DOAS analysis is the linear least-squares method; see [41, Chapter 8].

Two issues should be considered in the improvement of the DOAS analysis technique. The first issue is the identification of involved gases $\mathcal{M}$, which was assumed to be known in advance (or identified by other methods) as in the traditional DOAS analysis technique. However, as shown in [41, Chapter 8], this prior information requirement may hinder the development and applications of DOAS automatic analysis technique. Secondly, besides the additive noise due to scattering, a challenging complication in practice is the wavelength misalignment, which is caused by the installment of the optical spectrometer and leads to the basis noise; see [20]. In particular, caused by the installment or environment of optical spectrometers, the nominal wavelengths in the measurement of $A(\omega)$ may not correspond exactly to those in the basis $\{\theta_i(\omega)\}$, but aligned with an approximately linear deformation $\{\theta_i(\omega + d(\omega))\}$. Hence, equipped with the wavelength misalignment, (4.3) is turned into

$$A(\omega) = L \sum_{i \in \mathcal{M}} x_i \theta_i(\omega + d(\omega)) + \epsilon(\omega). \tag{4.4}$$

However, the quantity of linear deformation $d(\cdot)$ is usually unknown in practice. Therefore, given the data $A(\cdot)$ and reference spectra $\{\theta_i(\cdot)\}$, the DOAS automatic analysis aims to identify the involved gases and quantify their concentrations $\{x_i\}$ and the deformation $\{d(\cdot)\}$ from the linear model (4.4).

To deal with these two issues, inspired by the ideas of variable selection and using a set of modified bases [20], we construct a dictionary by enlarging characteristic absorption spectra database $\{\theta_i\}_{i \in \mathcal{M}}$ to that of a set of possible (common but concerned) gases $\{\theta_i\}_{i \in \mathcal{P}}$ with $\mathcal{M} \subseteq \mathcal{P}$ and deforming each θ_i with a set of possible deformations $\mathcal{D}$. In particular, the deformations can be approximated by linear functions, i.e.,

$$\mathcal{D} := \{d(\omega) : d(\omega) = u\omega + v, u \in \mathcal{U}, v \in \mathcal{V}\}, \tag{4.5}$$

from two predetermined sets $\mathcal{U}$ and $\mathcal{V}$. Let Θ_j denote a matrix whose columns are the reference spectra $\{\theta_i\}_{i\in\mathcal{P}}$ of j-th deformation, i.e., $\Theta_j := [\theta_i(\omega + d_j(\omega))]_{i\in\mathcal{P}}$ for $d_j \in \mathcal{D}$, and let $x_{d_j} \in \mathbb{R}^{|\mathcal{P}|}$ denote the concentrations of possible gases at j-th deformation. Then with the enlarged sets of candidates $\mathcal{P}$ and deformations $\mathcal{D}$, (4.4) can be rewritten as

$$A = L\left[\Theta_1, \ldots, \Theta_{|\mathcal{D}|}\right]\begin{bmatrix} x_{d_1} \\ \vdots \\ x_{d_{|\mathcal{D}|}} \end{bmatrix} + \epsilon. \tag{4.6}$$

Note that the solution of the linear system (4.6) enjoys a mix sparse structure. Specifically, there are in general only a few gases involved in a mixture (contrast with the number of possible candidates), and the deformation for the involved gases is unique and consistent (tested in an experiment). Consequently, with the group structure $x := (x_{d_1}^\top, \ldots, x_{d_{|\mathcal{D}|}}^\top)^\top$, the ground true solution of the DOAS has 1-group sparsity (inter-group sparsity) and the nonzero group is also sparse (intra-group sparsity). Therefore, writing $n := |\mathcal{P}| \times |\mathcal{D}|$, the mix sparse optimization model for the DOAS automatic analysis with wavelength misalignment (4.6) can be formulated as

$$\min_{x\in\mathbb{R}^n} \|\Theta x - \frac{A}{L}\|^2 + \lambda\|x\|_{2,0} + \tau\|x\|_0, \tag{4.7}$$

where $\|x\|_{2,0}$ is to promote the consistent deformation, $\|x\|_0$ is to promote the sparsity of candidate selection.

For the database of numerical experiments, we choose 15 concerned materials as possible candidates (listed in Table 1) and collect their reference characteristic absorption infrared (IR) spectra $\{\theta_i\}_{i\in\mathcal{P}}$ at National Institute of Standards and Technology (NIST) Chemistry WebBook[4]. Then we construct a dictionary Θ by deforming each collected reference IR spectrum by a set of linear deformations (4.5) with $\mathcal{U} = \{-0.15 + 0.05k : k = 1, \ldots, 5\}$ and $\mathcal{V} = \{-3 + k : k = 1, \ldots, 5\}$; see Figure 7 for the reference IR spectra with some deformed examples. Therefore, there are 15 possible materials and 25 possible deformations, and each spectrum has 554 sample intensity along with wavelength; consequently, the dictionary $\Theta \in \mathbb{R}^{554\times 375}$.

In the numerical experiments, the simulation data are generated via the following process. Firstly, the dictionary Θ is constructed as mentioned above with each θ_i being normalized. Then we randomly pick several materials in $\mathcal{P}$ with their concentrations being i.i.d. Gaussian ensembles and randomly select a deformation in $\mathcal{D}$ with a uniform distribution, that is, the ground true solution $\bar{x}$ is randomly generated. With the generated Θ and $\bar{x}$, the observation data (absorption spectra) A is generated via (4.6) with ϵ being an additive Gaussian noise with the standard deviation being 0.1%.

In the implementation of the IMTC (cf. Algorithm 3.1), the initial point and parameters are selected as follows. The initial point is $x^0 := 0$, and we set the constant stepsize $v := \frac{1}{2\|\Theta\|^2}$,

[4]https://webbook.nist.gov/chemistry/

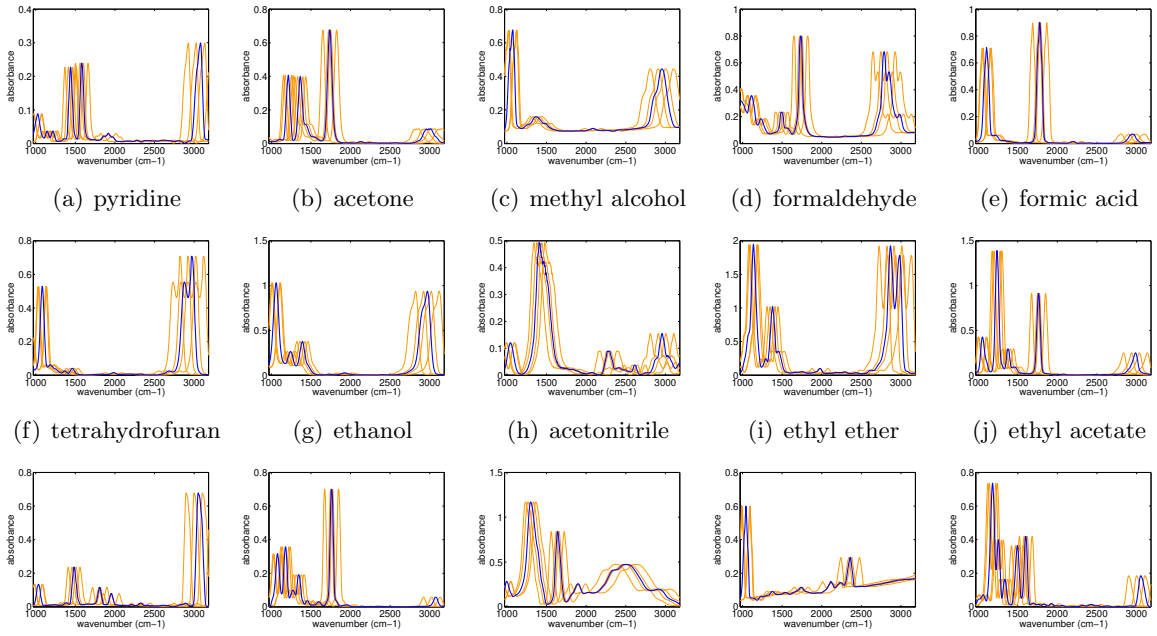
Table 1: Materials in reference dataset.

Number	Name	Formula	Number	Name	Formula
1	pyridine	C5H5N	9	ethyl ether	C4H10O
2	acetone	C3H6O	10	ethyl acetate	C4H8O2
3	methyl alcohol	CH4O	11	benzene	C6H6
4	formaldehyde	CH2O	12	benzoic acid	C7H6O2
5	formic acid	CH2O2	13	sodium bicarbonate	CHNaO3
6	tetrahydrofuran	C4H8O	14	ozone	O3
7	ethanol	C2H6O	15	phenol	C6H6O
8	acetonitrile	C2H3N			

regularization parameters $(\lambda, \tau) := (10^{-4}, 10^{-6})$ and the decreasing sequence of parameter pairs in the continuation technique as $(\lambda_k, \tau_k) = (\max\{\kappa\lambda_{k-1}, \lambda\}, \max\{\kappa\tau_{k-1}, \tau\})$ with $\kappa := 0.96$. Two key criteria to measure the performance of the solvers are the relative error (on observation or solution) and the true positive rate of solution (on materials or misalignment). In the numerical experiments, we compare the numerical performance on DOAS analysis of the IMTC with several state-of-the-art solvers, including the sparse solvers: OMP, ISTA, HardTA, HalfTA, the group sparse solvers: PGA ($\ell_{2,1}$), PGA ($\ell_{2,0}$), and the mix sparse solver: PGASGL.

In the first experiment, we aim to show the spectrum reconstruction, the materials identification and quantification capability of the algorithms. Figure 8 exhibits the materials and the misalignment predicted by algorithms, as well as their relative error of concentrations, and Figure 9 displays the absorbtion spectra reconstructed by the algorithms, as well as their relative error of spectra, at a random trial. It is illustrated from Figure 8 that the solutions obtained by the sparse solvers have different misalignments and thus have no physical sense; the solutions obtained by the group sparse solvers can exactly predict the misalignment but cannot predict the involved materials; and the IMTC can exactly predict the misalignment and the involved materials simultaneously (the relative error is quite small), and outperform the PGASGL. It is revealed from Figure 9 the IMTC can exactly reconstruct the absorbtion spectra and outperform other solvers.

In the second experiment, we aim to show the stability of the IMTC on predicting the materials and misalignment, comparing with other algorithms. Figure 10 plots the true positive rate (TPR) of the predicted materials and misalignment and the relative error of solutions obtained by the algorithms along with the number of involved materials (from 1 to 15) at 500 random simulations. It is demonstrated from Figures 10(a) and 10(b) that the IMTC has a much higher true positive rate than other algorithms and from Figure 10(c) that the solution of the IMTC is much more precise than the ones obtained by other algorithms. Moreover, the CPU time of the IMTC is less than 0.1 second. Therefore, the numerical results show that the IMTC can quickly, stably, exactly and quantitatively predict the existing materials and the factual wavelength misalignment simultaneously, which meets the demand of improvement of the DOAS automatic analysis software proposed in [41, Chapter 8].



(a) pyridine (b) acetone (c) methyl alcohol (d) formaldehyde (e) formic acid

(f) tetrahydrofuran (g) ethanol (h) acetonitrile (i) ethyl ether (j) ethyl acetate

(k) benzene (l) benzoic acid (m) bicarbonate (n) ozone (o) phenol

Figure 7: For each material, the reference spectrum is plotted in blue, while four deformed spectra are in orange.

References

[1] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137:91–129, 2013.

[2] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(2):1179–1225, 2008.

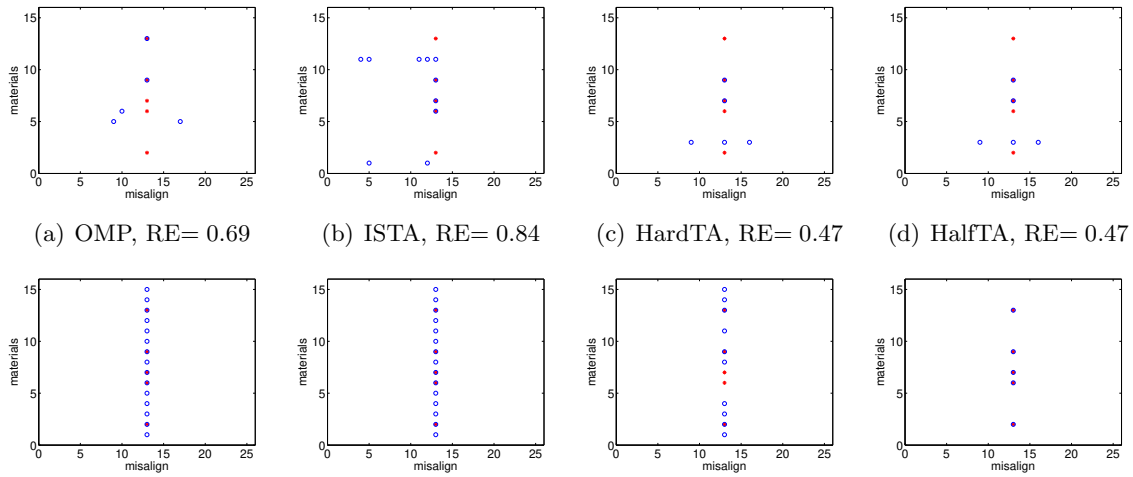
[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[4] S. Bahmani, B. Raj, and P. T. Boufounos. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14:807–841, 2013.

[5] 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.

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

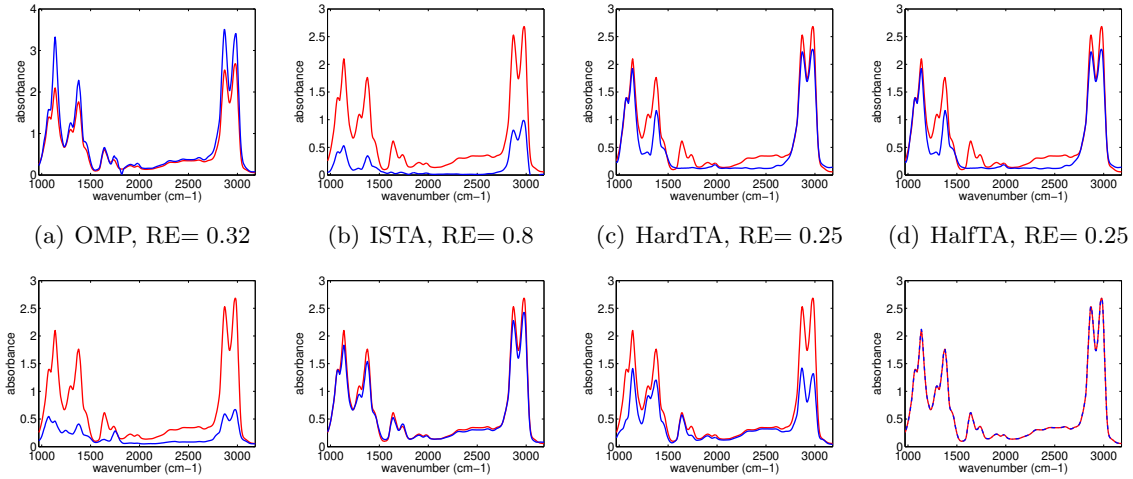
[7] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.



(a) OMP, RE= 0.69 (b) ISTA, RE= 0.84 (c) HardTA, RE= 0.47 (d) HalfTA, RE= 0.47

(e) PGA ($\ell_{2,1}$), RE= 0.82 (f) PGA ($\ell_{2,0}$), RE= 0.26 (g) PGASGL, RE= 0.42 (h) IMTC, RE= 0.008

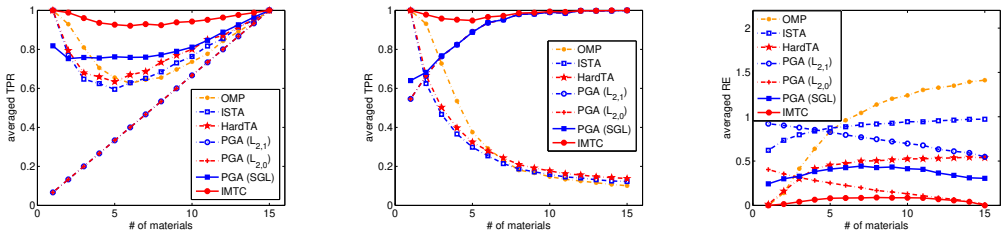
Figure 8: Identification and quantification of algorithms at a trial.



(a) OMP, RE= 0.32 (b) ISTA, RE= 0.8 (c) HardTA, RE= 0.25 (d) HalfTA, RE= 0.25

(e) PGA ($\ell_{2,1}$), RE= 0.74 (f) PGA ($\ell_{2,0}$), RE= 0.10 (g) PGASGL, RE= 0.45 (h) IMTC, RE= 0.006

Figure 9: Spectrum reconstruction of algorithms at a trial.



(a) TPR of materials (b) TPR of misalign (c) RE of prediction

Figure 10: Identification and quantification of algorithms in 500 simulations.

[8] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5):629–654, 2008.

[9] T. Blumensath and M. E. Davies. Normalized iterative hard thresholding: Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):298–309, 2010.

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

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

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

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

[14] J. Chen, G. Dai, and N. Zhang. An application of sparse-group Lasso regularization to equity portfolio optimization and sector selection. *Annals of Operations Research*, 2019.

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

[16] 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.

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

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

[19] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[20] E. Esser, Y. Lou, and J. Xin. A method for finding structured sparse solutions to non-negative least squares problems with applications. *SIAM Journal on Imaging Sciences*, 6(4):2010–2046, 2013.

[21] 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.

[22] M. J. Fullwood and et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462:58–64, 2009.

[23] D. Goldfarb and S. Ma. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210, 2011.

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

[25] Y. Hu, C. Li, K. Meng, J. Qin, and X. Yang. Group sparse optimization via $\ell_{p,q}$ regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017.

[26] Y. Hu, C. Li, K. Meng, and X. Yang. Linear convergence of inexact descent methods and inexact proximal gradient algorithms for lower-order regularization problems. *Journal of Global Optimization*, 79(4):853–883, 2021.

[27] J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4):1978–2004, 2010.

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

[29] Y. Jiao, B. Jin, and X. Lu. Iterative soft/hard thresholding with homotopy continuation for sparse recovery. *IEEE Signal Processing Letters*, 24(6):784–788, 2017.

[30] M. 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.

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

[32] Y. Li, B. Nan, and J. Zhu. Multivariate sparse group Lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363, 2015.

[33] J. Lu, K. Qiao, X. Li, Y. Zou and Z. Lu. ℓ_0-minimization methods for image restoration problems based on wavelet frames. *Inverse Problems*, 35:064001, 2019.

[34] Z. Lu. Iterative hard thresholding methods for l_0 regularized convex cone programming. *Mathematical Programming*, 147(1):125–154, 2014.

[35] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.

[36] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

[37] K. W. Meng and X. Q. Yang. Optimality conditions via exact penalty functions. *SIAM Journal on Optimization*, 20(6):3208–3231, 2010.

[38] 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.

[39] L. Pan and X. Chen. Group sparse optimization for images recovery using capped folded concave functions. *SIAM Journal on Imaging Sciences*, 14(1):1–25, 2021.

[40] D. N. Phan and H. A. L. Thi. Group variable selection via $\ell_{p,0}$ regularization and application to optimal scoring. *Neural Networks*, 118:220–234, 2019.

[41] U. Platt and J. Stutz. *Differential Optical Absorption Spectroscopy: Principles and Applications*. Springer-Verlag, Berlin, Heidelberg, 2008.

[42] U. Platt and T. Wagner. Satellite mapping of enhanced BrO concentrations in the troposphere. *Nature*, 395(6701):486–490, 1998.

[43] 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.

[44] J. Qin, Y. Hu, J.-C. Yao, R. W. T. Leung, Y. Zhou, Y. Qin, and J. Wang. Cell fate conversion prediction by group sparse optimization method utilizing single-cell and bulk OMICs data. *Briefings in Bioinformatics*, 22(6):bbab311 2021.

[45] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer Science & Business Media, Berlin, 2009.

[46] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[47] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[48] J. Wang, Y. Hu, C. Li, and J.-C. Yao. Linear convergence of CQ algorithms and applications in gene regulatory network inference. *Inverse Problems*, 33(5):055017, 2017.

[49] 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.

[50] 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.

[51] J. Yang, D. Sun, and K.-C. Toh. A proximal point algorithm for log-determinant optimization with group Lasso regularization. *SIAM Journal on Optimization*, 23(2):857–893, 2013.

[52] 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.

[53] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group Lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2104–2116, 2013.

[54] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2007.

[55] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 08 2008.

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

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

[58] Y.-B. Zhao. Optimal k-thresholding algorithms for sparse optimization problems. *SIAM Journal on Optimization*, 30(1):31–55, 2020.

[59] Y. Zhou, J. H. Han, X. H. Yuan, Z. C. Wei, and R. C. Hong. Inverse sparse group Lasso model for robust object tracking. *IEEE Transactions on Multimedia*, 19(8):1798–1810, 2017.