Neighborhood Regularized ℓ^1-Graph

Yingzhen Yang[1], **Jiashi Feng**[2], **Jiahui Yu**[3], **Jianchao Yang**[1], **Pushmeet Kohli**[4], **Thomas S. Huang**[3]
[1] Snap Research
[2] Department of ECE, National University of Singapore, Singapore
[3] Beckman Institute, University of Illinois at Urbana-Champaign
[4] Google DeepMind

Abstract

ℓ^1-Graph, which learns a sparse graph over the data by sparse representation, has been demonstrated to be effective in clustering especially for high dimensional data. Although it achieves compelling performance, the sparse graph generated by ℓ^1-Graph ignores the geometric information of the data by sparse representation for each datum separately. To obtain a sparse graph that is aligned to the underlying manifold structure of the data, we propose the novel Neighborhood Regularized ℓ^1-Graph (NRℓ^1-Graph). NRℓ^1-Graph learns sparse graph with locally consistent neighborhood by encouraging nearby data to have similar neighbors in the constructed sparse graph. We present the optimization algorithm of NRℓ^1-Graph with theoretical guarantee on the convergence and the gap between the suboptimal solution and the globally optimal solution in each step of the coordinate descent, which is essential for the overall optimization of NRℓ^1-Graph. Its provable accelerated version, NRℓ^1-Graph by Random Projection (NRℓ^1-Graph-RP) that employs randomized data matrix decomposition, is also presented to improve the efficiency of the optimization of NRℓ^1-Graph. Experimental results on various real data sets demonstrate the effectiveness of both NRℓ^1-Graph and NRℓ^1-Graph-RP.

This work is supported in part by US Army Research Office grant W911NF-15-1-0317. The work of Jiashi Feng was supported by NUS startup R-263-000-C08-133, MOE R-263-000-C21-112 and IDS R-263-000-C67-646. Pushmeet Kohli was at Microsoft Research during this project.

1 INTRODUCTION

Similarity-based clustering methods, e.g. K-means (Duda et al., 2000), Affinity Propagation (AP) (Frey and Dueck, 2007) and Spectral Clustering (Ng et al., 2001), segment the data based on the similarity measure between data points. In this manner, similarity-based methods alleviate the difficulty of parameter estimation that model-based methods face, such as modeling data distribution by a mixture of parameterized distributions (Fraley and Raftery, 2002). Among various similarity-based clustering methods, sparse graph based methods, which build sparse graph with only a few edges for each vertex and the data similarity serves as edge weight, are demonstrated to be effective, especially for clustering high dimensional data. Examples of sparse graph based clustering methods include ℓ^1-Graph (Yan and Wang, 2009; Cheng et al., 2010) and Sparse Subspace Clustering (SSC) (Elhamifar and Vidal, 2013), which build the sparse graph by reconstructing each datum with all the other data by sparse representation. In the sparse graph produced by ℓ^1-Graph or SSC, the vertices represent the data, and an edge is between two vertices whenever one participates the spare representation of the other. The weight of the edge is the average of the associated elements in the sparse codes corresponding to the two vertices. Throughout this paper, data point or data may also refer to the corresponding vertex or vertices in the sparse graph if no confusion arises. A sparse similarity matrix is then obtained as the weighted adjacency matrix of the constructed sparse graph by ℓ^1-Graph or SSC, and spectral clustering is performed on the sparse similarity matrix to obtain the data clusters. ℓ^1-Graph and SSC have been shown to be robust to noise and capable of producing superior results for high-dimensional data, compared to spectral clustering on similarity computed by the widely used Gaussian kernel. Such sparse graph has also been successfully applied to a novel deep neural network architecture for the first time (Peng et al., 2016).

While ℓ^1-Graph demonstrates compelling performance for clustering, it performs sparse representation for each datum independently, and the sparse codes of nearby data lack smoothness in accordance with the geometric information of the data. Regularized ℓ^1-Graph (Yang et al., 2014) employs a manifold assumption which imposes local smoothness on the sparse codes of nearby data, namely nearby data are encouraged to have similar sparse codes in the sense of ℓ^2-distance. This method is also termed ℓ^2-Rℓ^1-Graph in this paper. Moreover, various regularized sparse coding methods, such as (Liu et al., 2010; He et al., 2011; Zheng et al., 2011; Gao et al., 2013), also utilize manifold assumption (Belkin et al., 2006) to obtain locally smooth sparse codes by ℓ^2-distance based graph regularization term.

ℓ^2-Rℓ^1-Graph imposes locally smoothness or consistency on the sparse codes of the data (i.e. the vertices of the graph), not directly on the local structure of the sparse graph. Consistency of the sparse codes indicates that the codes are similar to each other. In this paper, we propose a novel Neighborhood Regularized ℓ^1-Graph, abbreviated as NRℓ^1-Graph, which learns locally consistent neighborhood in the sparse graph, i.e. nearby points have similar neighborhoods. In this manner, manifold assumption is employed directly on the structure of the graph so as to obtain the sparse graph that accounts for the local manifold structure of the data. NRℓ^1-Graph embodies the local consistency on the neighborhood by the novel neighborhood regularization term in the objective function. It should be emphasized that the widely used ℓ^2 graph regularization term cannot represent the neighborhood consistency. Another benefit of local consistency on the neighborhood in the sparse graph is that, instead of choosing neighbors by itself, each data point is encouraged to coordinate with its nearby points on the data manifold (usually specified by its K nearest neighbors by Euclidean distance) to choose its neighbors in the sparse graph (See Figure 1). This makes the sparse graph more robust to outliers while preserving the freedom in the spare representation of data without constraints on the magnitude of the sparse codes. We present the efficient proximal gradient descent (PGD) style iterative method for the optimization problem of NRℓ^1-Graph with theoretical guarantee on the convergence and bounded gap between the sub-optimal solution and the globally optimal solution for each step of coordinate descent for the optimization. Furthermore, NRℓ^1-Graph by Random Projection (NRℓ^1-Graph-RP) is proposed to accelerate the optimization of NRℓ^1-Graph by randomized rank-k approximation of the data matrix. Such low rank approximation reduces the time complexity of the gradient descent step in the PGD-style iterative method from $\mathcal{O}(nd)$ to $\mathcal{O}(nk + dk)$, where d and n are dimension and size of the data, leading to the significantly improved efficiency when $k \ll \min\{d, n\}$ compared to the original NRℓ^1-Graph.

We use bold letters for matrices and vectors, and regular lower letter for scalars throughout this paper. The bold letter with superscript indicates the corresponding column of a matrix, e.g. $\mathbf{Z}^i$ indicates the i-th column of the matrix $\mathbf{Z}$, and the bold letter with subscript indicates the corresponding element of a matrix or vector. $\|\cdot\|_F$ and $\|\cdot\|_p$ denote the Frobenius norm and the ℓ^p-norm, and $\mathrm{diag}(\cdot)$ indicates the diagonal elements of a matrix. $\sigma_{\max}(\cdot)$ and $\sigma_{\max}(\cdot)$ indicate the maximum and minimum singular value of a matrix.

2 PRELIMINARIES: ℓ^1-GRAPH, ℓ^2-Rℓ^1-GRAPH

ℓ^1-Graph (Yan and Wang, 2009; Cheng et al., 2010) and SSC (Elhamifar and Vidal, 2009, 2013) apply the idea of sparse coding where the data similarity is represented by sparse codes. Given data $\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, ℓ^1-Graph and SSC solve the following optimization problem to obtain a sparse representation for each data point

$$\min_{\mathbf{Z}^i\in\mathbb{R}^n, \mathbf{Z}^i_i=0} \|\mathbf{Z}^i\|_1 \quad s.t.\ \mathbf{x}_i = \boldsymbol{X}\mathbf{Z}^i \tag{1}$$

SSC (Elhamifar and Vidal, 2009, 2013) also proves that the above sparse representation recovers the underlying independent or disjoint subspaces from which the data are generated when certain conditions on the geometric properties of the subspaces, such as the principle angle between different subspaces, hold. Allowing some tolerance for inexact representation and robustness to noise (Wang and Xu, 2013; Soltanolkotabi et al., 2014), the following Lasso-type problem is solved instead of (1):

$$\min_{\mathbf{Z}^i\in\mathbb{R}^n, \mathbf{Z}^i_i=0} \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^i\|_2^2 + \lambda^{(\ell^1)}\|\mathbf{Z}^i\|_1 \quad i = 1, \ldots, n \tag{2}$$

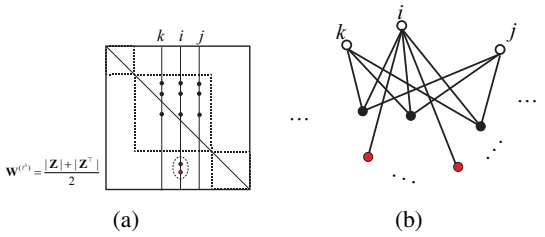
for some weighting parameter $\lambda > 0$, and $\mathbf{Z}^i \in \mathbb{R}^n$, $\mathbf{Z} = [\mathbf{Z}^1, \ldots, \mathbf{Z}^n] \in \mathbb{R}^{n\times n}$ is the sparse code matrix with the element $\mathbf{Z}_{ij} = \mathbf{Z}^j_i$. The diagonal elements of $\mathbf{Z}$ are enforced to be zero, i.e. $\mathbf{Z}_{ii} = 0$ for $1 \le i \le n$, so as to avoid trivial solution $\mathbf{Z} = \mathbf{I}_n$ where $\mathbf{I}_n$ is a $n \times n$ identity matrix.

ℓ^1-Graph constructs the sparse graph $G = (\boldsymbol{X}, \mathbf{W}^{(\ell^1)})$ where $\boldsymbol{X}$ is the set of vertices, $\mathbf{W}$ is the weighted adjacency matrix of G and $\mathbf{W}_{ij}$ indicates the edge weight, or the similarity, between $\mathbf{x}_i$ and $\mathbf{x}_j$. $\mathbf{W}$ is set by the sparse codes:

$$\mathbf{W}^{(\ell^1)}_{ij} = (|\mathbf{Z}_{ij}| + |\mathbf{Z}_{ji}|)/2 \quad 1 \le i, j \le n \tag{3}$$

And there is an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ if and only if $\mathbf{W}^{(\ell^1)}_{ij} \neq 0$, i.e. either $\mathbf{x}_i$ chooses $\mathbf{x}_j$ as its neighbor

by nonzero $\mathbf{Z}_{ji}$, or $\mathbf{x}_j$ chooses $\mathbf{x}_i$ as its neighbor by nonzero $\mathbf{Z}_{ij}$. ℓ^1-Graph then performs spectral clustering on $\mathbf{W}^{(\ell^1)}$ to obtain the data clusters, with better performance than spectral clustering on the similarity matrix produced by the widely used Gaussian kernel. ℓ_1-Graph features robustness to data noise and adaptive neighborhood, specified by the non-zero entries in the sparse codes. It should be emphasized that the above sparse graph construction method is used for almost all the sparse graph based clustering methods (Yan and Wang, 2009; Cheng et al., 2010; Elhamifar and Vidal, 2009, 2013, 2011) while the sparse codes are learnt in different ways.



(a) (b)

Figure 1: (a) Illustration of the weighted adjacency matrix of the sparse graph shown in (b), where the three black dots indicate three common neighbors of points $\mathbf{x}_k$, $\mathbf{x}_i$ and $\mathbf{x}_j$. The inner dashed box specifies the scope of correct neighbors, i.e. the ones in the same ground truth cluster. The locally consistent neighborhood encourages $\mathbf{x}_i$ to abandon the wrong neighbors marked with red dots encompassed by the dashed ellipse. (b) The sparse graph corresponding to the weighted adjacency matrix in (a).

The widely adopted manifold assumption assumes that high-dimensional data always lie on or close to a submanifold of low intrinsic dimension, and clustering the data according to its underlying manifold structure is important and challenging in machine learning. While ℓ^1-Graph demonstrates better performance than many traditional similarity-based clustering methods, it performs sparse representation for each datum independently without considering the geometric information and manifold structure of the entire data. On the other hand, in order to obtain the embedding of the data that accounts for the geometric information and manifold structure of the data, manifold assumption (Belkin et al., 2006) is usually employed (Liu et al., 2010; He et al., 2011; Zheng et al., 2011; Gao et al., 2013). Interpreting the sparse code of a data point as its embedding, most existing methods that employ manifold assumption in the sparse representation literature require that if two points $\mathbf{x}_i$ and $\mathbf{x}_j$ are close in the intrinsic geometry of the manifold, their corresponding sparse codes $\mathbf{Z}^i$ and $\mathbf{Z}^j$ are also expected to be similar to each other in the sense of ℓ^2-distance (Zheng et al., 2011; Gao et al., 2013). In other words, $\mathbf{Z}$ varies smoothly along the geodesics in the intrinsic geometry.

Based on the spectral graph theory (Chung, 1997), extensive literature uses graph Laplacian to impose local smoothness of the embedding and preserve the local manifold structure (Belkin et al., 2006; Zheng et al., 2011; Gao et al., 2013). Given a proper symmetric similarity matrix $\mathbf{S}$ that encodes the intrinsic manifold structure of the data, the sparse code $\mathbf{Z}$ in accordance with the manifold assumption by graph Laplacian minimizes the following ℓ^2 regularization term below:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{ij}\|\mathbf{Z}^i-\mathbf{Z}^j\|_2^2=\mathrm{Tr}(\mathbf{Z}\mathbf{L}_\mathbf{S}\mathbf{Z}^\top) \tag{4}$$

where the ℓ^2-norm is used to measure the distance between sparse codes. $\mathbf{L}_\mathbf{S}=\mathbf{D}_\mathbf{S}-\mathbf{S}$ is the graph Laplacian using the adjacency matrix $\mathbf{S}$, the degree matrix $\mathbf{D}_\mathbf{S}$ is a diagonal matrix with each diagonal element being the sum of the elements in the corresponding row of $\mathbf{S}$, namely $(\mathbf{D}_\mathbf{S})_{ii}=\sum_{j=1}^{n}\mathbf{S}_{ij}$. To the best of our knowledge, such ℓ^2 regularization is employed by most methods that use graph regularization for sparse representation. *Incorporating the ℓ^2 graph regularization term into the objective of ℓ^1-Graph, the optimization problem of ℓ^2-Rℓ^1-Graph*, is presented below:

$$\begin{aligned}\min_{\mathbf{Z}\in\mathbb{R}^{n\times n},\mathrm{diag}(\mathbf{Z})=\mathbf{0}}&\sum_{i=1}^{n}\|\mathbf{x}_i-\boldsymbol{X}\mathbf{Z}^i\|_2^2+\lambda^{(\ell^2)}\|\mathbf{Z}^i\|_1\\&+\gamma^{(\ell^2)}\mathrm{Tr}(\mathbf{Z}\mathbf{L}_\mathbf{S}\mathbf{Z}^\top)\end{aligned} \tag{5}$$

where $\gamma^{(\ell^2)}>0$ is the weighting parameter for the ℓ^2 graph regularization term. Regularized ℓ^1-Graph (Yang et al., 2014) employs the formulation with the same form as (5. Following the representative ℓ^2 graph regularized sparse coding methods (Zheng et al., 2011; Gao et al., 2013), $\mathbf{S}$ is typically chosen as the adjacency matrix of K-Nearest-Neighbor (KNN) graph to represent the local manifold structure of the data, i.e. $\mathbf{S}_{ij}=1$ if and only if $\mathbf{x}_i$ is among the K nearest neighbors of $\mathbf{x}_j$. Note that KNN is extensively used in the manifold learning literature, such as Locally Linear Embedding (Roweis and Saul, 2000), Laplacian Eigenmaps (Belkin and Niyogi, 2003) and Sparse Manifold Clustering and Embedding (Elhamifar and Vidal, 2011), to reveal the local structure in the manifold. Although $\mathbf{S}$ is not symmetric, letting $\mathbf{S}'=\frac{\mathbf{S}+\mathbf{S}^\top}{2}$, then a symmetric adjacency matrix can be used in the graph regularization term without changing its value: $\mathrm{Tr}(\mathbf{Z}\mathbf{L}_{\mathbf{S}'}\mathbf{Z}^\top)=\mathrm{Tr}(\mathbf{Z}\mathbf{L}_\mathbf{S}\mathbf{Z}^\top)$.

In the next section, we propose NRℓ^1-Graph which learns locally consistent neighborhood in the sparse graph by a novel neighborhood regularization term rather than the ℓ^2 graph regularization term with superior clustering performance.

Algorithm 1 Learning NRℓ^1-Graph

Input:
The data set $\boldsymbol{X} = \{\mathbf{x}_i\}_{i=1}^n$, the number of clusters c, the parameter $\lambda^{(\ell^2)}$, $\gamma^{(\ell^2)}$, K for NRℓ^1-Graph, maximum iteration number M_c for coordinate descent, and maximum iteration number M_p for the iterative proximal method, stopping threshold ε.

$r = 1$, initialize the sparse code matrix as $\mathbf{Z}^{(0)} = \mathbf{Z}^{(\ell^1)}$.
while $r \leq M_c$ **do**
  Obtain $\mathbf{Z}^{(r)}$ from $\mathbf{Z}^{(r-1)}$ by coordinate descent. In i-th ($1 \leq i \leq n$) step of the r-th iteration of coordinate descent, solve (8) using the PGD-style iterative method (11), (12) and (13) to update $\mathbf{Z}^i$ in each iteration of the iterative proximal method.
  if $|L(\mathbf{Z}^{(r)}) - L(\mathbf{Z}^{(r-1)})| < \varepsilon$ **then**
    break
  else
    $r = r + 1$.
  end if
end while
Obtain the sub-optimal sparse code matrix $\mathbf{Z}^*$ when the above iterations converge or maximum iteration number is achieved.
Build the pairwise similarity matrix by symmetrizing $\mathbf{Z}^*$: $\mathbf{W}^* = \frac{|\mathbf{Z}^*|+|\mathbf{Z}^*|^\top}{2}$

Output: The sparse graph whose weighted adjacency matrix is $\mathbf{W}^*$.

3 NEIGHBORHOOD REGULARIZED ℓ^1-GRAPH

In this section, we propose Neighborhood Regularized ℓ^1-Graph (NRℓ^1-Graph) which learns locally consistent neighborhood in the sparse graph. Instead of imposing local smoothness on the sparse codes in the existing regularized ℓ^1-Graph (Yang et al., 2014), NRℓ^1-Graph learns locally consistent neighborhood so as to capture the local manifold structure of the data in the construction of the sparse graph. In addition, NRℓ^1-Graph benefits from robustness to noise or outliers by encouraging each point to choose its neighbors in the sparse graph via coordinating with its nearby points on the manifold. Note that ℓ^2-distance based graph regularization cannot enjoy this benefit since small ℓ^2-distance between the sparse codes of nearby data points does not guarantee their consistent neighborhood in the sparse graph. The optimization problem of NRℓ^1-Graph is presented below:

$$\min_{\mathbf{Z}\in\mathbb{R}^{n\times n},\mathrm{diag}(\mathbf{Z})=\mathbf{0}} L(\mathbf{Z}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^i\|_2^2 + \lambda\|\mathbf{Z}^i\|_1 + \gamma\mathbf{R}_{\mathbf{S}}(\mathbf{Z}) \tag{6}$$

where $\mathbf{R}_{\mathbf{S}}(\mathbf{Z}) = \sum_{i,j=1}^{n} \mathbf{S}_{ij} d(\mathbf{Z}^i, \mathbf{Z}^j)$ is the novel neighborhood regularization term, $\mathbf{S}$ is the adjacency matrix of the KNN graph, $\gamma > 0$ is the weighting parameter. $d(\mathbf{Z}^i, \mathbf{Z}^j)$ indicates the neighborhood distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ in terms of the sparse code matrix $\mathbf{Z}$ through the weighted adjacency matrix $\mathbf{W}$ (3), which measures the number of different neighbors these two points have in the sparse graph:

$$d(\mathbf{Z}^i, \mathbf{Z}^j) = \sum_{1\leq k\leq n, k\neq i,j} (\mathbb{1}_{\mathbf{W}_{ki}=0,\mathbf{W}_{kj}\neq 0} + \mathbb{1}_{\mathbf{W}_{ki}\neq 0,\mathbf{W}_{kj}=0}) \tag{7}$$

where $\mathbf{W} = \frac{|\mathbf{Z}|+|\mathbf{Z}^\top|}{2}$ and $\mathbb{1}$ is the indicator function. Indices i and j are excluded when computing the neighborhood distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ since any point is not a neighbor of itself in the sparse graph, so it is not necessary to impose penalty on $\mathbf{x}_i$ in the calculation of $d(\mathbf{Z}^i, \mathbf{Z}^j)$ if it does choose $\mathbf{x}_j$ as its neighbor, and vice versa.

3.1 OPTIMIZATION ALGORITHM

We use coordinate descent to optimize (6). In each iteration of coordinate descent, the optimization is performed with respect to $\mathbf{Z}^i$ sequentially for $1 \leq i \leq n$, while fixing all the other sparse codes $\{\mathbf{Z}^j\}_{j\neq i}$. In the i-th step of each iteration of coordinate descent, the optimization problem for $\mathbf{Z}^i$ is below:

$$\min_{\mathbf{Z}^i\in\mathbb{R}^n,\mathbf{Z}_i^i=0} F(\mathbf{Z}^i) = \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^i\|_2^2 + \lambda\|\mathbf{Z}^i\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i) \tag{8}$$

where $\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i) = \sum_{j=1}^{n} \tilde{\mathbf{S}}_{ij} d(\mathbf{Z}^i, \mathbf{Z}^j)$, $\tilde{\mathbf{S}} = \mathbf{S} + \mathbf{S}^\top$.

We employ proximal gradient descent method (PGD) to optimize the nonconvex problem (8) inspired by the proximal linearized method (Bolte et al., 2014). Although the proximal mapping is typically associated with a lower semicontinuous function (Bolte et al., 2014) and it can be verified that $\mathbf{R}_{\tilde{\mathbf{S}}}$ is not always lower semicontinuous, we can still derive a PGD-styple iterative method to optimize (8).

Define $\mathbf{F}^{\tilde{\mathbf{S}}} \in \mathbb{R}^{n\times n}$ as $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} = \sum_{j\neq k} \tilde{\mathbf{S}}_{ij}\mathbb{1}_{\mathbf{W}_{kj}=0} - \sum_{j\neq k} \tilde{\mathbf{S}}_{ij}\mathbb{1}_{\mathbf{W}_{kj}\neq 0}$ where $\mathbb{1}$ is the indicator function, then $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}$ indicates the degree to which $\mathbf{Z}_{ki}$ is discouraged to be nonzero and it can be verified that up to a constant irrelevant to $\mathbf{Z}^i$,

$$\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i) = \sum_{1\leq k\leq n,k\neq i} \mathbb{1}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}\mathbb{1}_{\mathbf{Z}_{ki}\neq 0} = \sum_{k\in\Lambda_i} \mathbf{F}_{ki}^{\tilde{\mathbf{S}}}\mathbb{1}_{\mathbf{Z}_{ki}\neq 0} \tag{9}$$

where Λ_i is defined as the set comprising all the possible indices k other than i such that $\mathbf{Z}_{ik} = 0$:

$$\Lambda_i \triangleq \{k\colon 1 \leq k \leq n, k \neq i, \mathbf{Z}_{ik} = 0\} \tag{10}$$

Since each indicator function $\mathbb{I}_{\mathbf{Z}_{ki}\neq 0}$ is lower semicontinuous, it can be verified that $\mathbf{R}_{\tilde{\mathbf{S}}}$ is lower semicontinuous if $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for each $k \in \Lambda_i$. In the following text, we let $Q(\mathbf{Z}^i) = \frac{1}{2}\|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^i\|_2^2$. The superscript with bracket indicates the iteration number of PGD or the iteration number of the coordinate descent without confusion. The PGD-style iterative method for optimizing (8) is as follows:

$$\tilde{\mathbf{Z}^i}^{(t)} = \mathbf{Z}^{i^{(t-1)}} - \frac{2}{\tau s}(\boldsymbol{X}^\top \boldsymbol{X}\mathbf{Z}^{i^{(t-1)}} - \boldsymbol{X}^\top \mathbf{x}_i) \tag{11}$$

where $\tau > 1$ is any constant greater than 1 and s is the Lipschitz constant for the gradient of function $Q(\cdot)$, namely

$$\|\nabla Q(\mathbf{y}) - \nabla Q(\mathbf{z})\|_2 \leq s\|\mathbf{y} - \mathbf{z}\|_2,\ \forall\, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$$

For all the other k such that $k \notin \Lambda_i$,

$$\mathbf{Z}^{(t)}_{ki} = \begin{cases} \mathbf{u}_k & : \quad k \neq i \\ 0 & : \quad k = i \end{cases} \tag{12}$$

and for all $k \in \Lambda_i$,

$$\mathbf{Z}^{(t)}_{ki} = \begin{cases} \underset{v\in\{\mathbf{u}_k, 0\}}{\arg\min} H_k(v) & : \quad \mathbf{u}_k \neq 0 \text{ or } \mathbf{u}_k = 0 \text{ and } \mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0 \\ \varepsilon & : \quad \mathbf{u}_k = 0 \text{ and } \mathbf{F}^{\tilde{\mathbf{S}}}_{ki} < 0 \end{cases} \tag{13}$$

where ε is any real number such that $\varepsilon \neq 0$ and $H_k(\varepsilon) \leq H_k(\mathbf{Z}^{(t-1)}_{ki})$. For each $k \in 1 \ldots n$ and $k \neq i$, H_k is defined below

$$H_k(v) = \frac{\tau s}{2}(v - \tilde{\mathbf{Z}}^{(t)}_{ki})^2 + \lambda|v| + \gamma \mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}\mathbb{I}_{v\neq 0} \tag{14}$$

for $v \in \mathbb{R}$. $\mathbf{u}$ is defined as

$$\mathbf{u} = \max\{|\tilde{\mathbf{Z}^i}^{(t)}| - \frac{\lambda}{\tau s}, 0\} \circ \operatorname{sign}(\tilde{\mathbf{Z}^i}^{(t)}) \tag{15}$$

where $\circ$ means element-wise multiplication.

Proposition 1 shows that the PGD-style iterative method decreases the value of the objective function in each iteration.

Proposition 1. *Let the sequence $\{\mathbf{Z}^{i^{(t)}}\}_t$ be generated by the PGD-style iterative method with (11), (12) and (13), then the sequence of the objective $\{F(\mathbf{Z}^{i^{(t)}})\}_t$ decreases, and the following inequality holds for $t \geq 1$:*

$$F(\mathbf{Z}^{i^{(t)}}) \leq F(\mathbf{Z}^{i^{(t-1)}}) - \frac{(\tau - 1)s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2 \tag{16}$$

And it follows that the sequence $\{F(\mathbf{Z}^{i^{(t)}})\}_t$ converges.

Remark 1. (11), (12) and (13) in each iteration of the proposed PGD-style iterative method are similar to the update rules of the ordinary PGD. (11) performs gradient descent on the differential part, (12) and (13) can be viewed as an approximate solution to the proximal mapping $\min_{\mathbf{v}\in\mathbb{R}^n} H(\mathbf{v}) = \frac{\tau s}{2}\|\mathbf{v} - \tilde{\mathbf{Z}^i}^{(t)}\|_2^2 + \lambda\|\mathbf{v}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{v})$. Since $\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)$ is not always lower semicontinuous, $\arg\min_{\mathbf{v}\in\mathbb{R}^n} H(\mathbf{v})$ is not guaranteed to exist. One can see a simple example wherein $\mathbf{u}_k = 0$ and $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} < 0$ for some $k \in \Lambda_i$. In this case, $\inf_{v\in\mathbb{R}} H_k(v) = \frac{\tau s}{2}(\tilde{\mathbf{Z}}^{(t)}_{ki})^2 + \gamma\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}$ but this infinitum can not be achieved.

The PGD-style iterative proximal method starts from $t = 1$ and continues until the sequence $\{F(\mathbf{Z}^{i^{(t)}})\}$ converges or maximum iteration number is achieved. When the proximal method converges or terminates for each $\mathbf{Z}^i$, the step for $\mathbf{Z}^i$ in one iteration of coordinate descent is finished and the optimization algorithm proceed to optimize other sparse codes. We initialize $\mathbf{Z}$ as $\mathbf{Z}^{(0)} = \mathbf{Z}^{(\ell^1)}$ and $\mathbf{Z}^{(\ell^1)}$ is the sparse codes generated by solving (2) with some proper weighting parameter $\lambda^{(\ell^1)}$. In all the experimental results shown in the next section, we empirically set $\lambda = 0.1$. The algorithm of learning NRℓ^1-Graph is described in Algorithm 1.

3.2 TIME COMPLEXITY

Let the maximum iteration number of coordinate descent be M_c, and maximum iteration number be M_p for the PGD-style iterative method solving (8), then the time complexity of running the coordinate descent for NRℓ^1-Graph is $\mathcal{O}(M_c M_p n^2 d)$.

3.3 THEORETICAL ANALYSIS FOR OPTIMIZATION

It can be observed that optimization by coordinate descent in Section 3.1 is essential for the overall optimization of NRℓ^1-Graph, and each step of the coordinate descent (8) is a difficult nonconvex problem and crucial for obtaining the support regularized sparse code, where the nonconvexity comes from the neighborhood regularization term $\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)$ (9). Therefore, the optimization of (8) plays an important role in the overall optimization of NRℓ^1-Graph. In the previous section, a PGD-style iterative method is proposed to decrease the value of the objective in each iteration. In this section, we provide further theoretical analysis on the optimization of problem (8) when $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$. This condition is equivalent to the condition that the neighborhood regularization function $\mathbf{R}_{\tilde{\mathbf{S}}}(\cdot)$ is lower semicontinuous. Under this condition, we prove that the sequence $\{\mathbf{Z}^{i^{(t)}}\}_t$ produced by the PGD-style iterative method converges to the suboptimal solution which is a critical point of the objective (8). By connecting the support regularized function to the capped-ℓ^1 norm and the nonconvexity analysis of the

support regularization term, we present the bound for ℓ^2-distance between the sub-optimal solution and the globally optimal solution to (8) in Theorem 1. Note that our analysis is valid for all $1 \leq i \leq n$.

Therefore, if $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$, the neighborhood regularization term $\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)$ is lower semicontinuous with respect to $\mathbf{Z}^i$ in (9). In this case, the PGD-style iterative method proposed in Section 3.1 for each iteration $t \geq 1$ becomes

$$\tilde{\mathbf{Z}}^{i^{(t)}} = \mathbf{Z}^{i^{(t-1)}} - \frac{2}{\tau s}(\boldsymbol{X}^\top \boldsymbol{X}\mathbf{Z}^{i^{(t-1)}} - \boldsymbol{X}^\top \mathbf{x}_i) \tag{17}$$

$$\mathbf{Z}_{ki}{}^{(t)} = \begin{cases} \underset{v \in \{\mathbf{u}_k, 0\}}{\arg\min} H_k(v) & : \quad k \in \Lambda_i \\ \mathbf{u}_k & : \quad k \notin \Lambda_i \end{cases} \tag{18}$$

which is equivalent to the updates rules in the ordinary proximal gradient descent method. The supplemetary explains the meaning of the condition that $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$.

In the following lemma, we show that the sequence $\{\mathbf{Z}^{i^{(t)}}\}_t$ generated by (11), (12) and (13) converges to a critical point of $F(\mathbf{Z}^i)$, denoted by $\hat{\mathbf{Z}}^i$. Denote by $\mathbf{Z}^{i^*}$ the globally optimal solution to the original optimization problem (8). The following lemma also shows that both $\hat{\mathbf{Z}}^i$ and $\mathbf{Z}^{i^*}$ are local solutions to the capped-ℓ^1 regularized problem (19). Before stating the lemma, the following definitions are introduced which are essential for our analysis.

Definition 1. *(Critical points) Given the non-convex function $f\colon \mathbb{R}^n \to R \cup \{+\infty\}$ which is a proper and lower semi-continuous function.*

- *for a given $\mathbf{x} \in \mathrm{dom} f$, its Frechet subdifferential of f at $\mathbf{x}$, denoted by $\tilde{\partial} f(x)$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^n$ which satisfy*

$$\limsup_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \to \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|} \geq 0$$

- *The limiting-subdifferential of f at $\mathbf{x} \in \mathbb{R}^n$, denoted by written $\partial f(x)$, is defined by*

$$\partial f(x) = \{\mathbf{u} \in \mathbb{R}^n \colon \exists \mathbf{x}^k \to \mathbf{x}, f(\mathbf{x}^k) \to f(\mathbf{x}), \\ \tilde{\mathbf{u}}^k \in \tilde{\partial} f(\mathbf{x}^k) \to \mathbf{u}\}$$

The point $\mathbf{x}$ is a critical point of f if $0 \in \partial f(x)$.

Also, we are considering the following capped-ℓ^1 regularized problem, which replaces the indicator function in the support regularization term $\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)$ with the continuous capped-ℓ^1 regularization term $\mathbf{T}$:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n, \boldsymbol{\beta}_i = 0} L_{\mathrm{capped}-\ell^1}(\boldsymbol{\beta}) = \|\mathbf{x}_i - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \mathbf{T}(\boldsymbol{\beta}; b) \tag{19}$$

where $\mathbf{T}(\boldsymbol{\beta}; b) = \sum_{1 \leq k \leq n, k \neq i} T_k(\boldsymbol{\beta}_k; b)$, $T_k(t; b) = \gamma \mathbb{I}_{\mathbf{Z}_{ik}=0} \mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \frac{\min\{|t|, b\}}{b}$ for some $b > 0$. It can be seen that the objective function of the capped-ℓ^1 problem approaches that of (8) when $\frac{\min\{|t|,b\}}{b}$ approaches the indicator function $\mathbb{I}_{t \neq 0}$ as $b \to 0+$. Define $\mathbf{P}(\cdot; b) = \lambda\|\cdot\|_1 + \mathbf{T}(\cdot; b)$, the location solution to the capped-ℓ^1 problem is defined as follows. In the following, $\boldsymbol{X}^{(-i)}$ is the matrix comprising all but the i-th columns of $\boldsymbol{X}$, and $\mathbf{v}_{-i} \in \mathbb{R}^{n-1}$ indicates the vector comprising of all but the i-th elements of any vector $\mathbf{v} \in \mathbb{R}^n$.

Definition 2. *(Local solution) A vector $\tilde{\boldsymbol{\beta}}$ is a local solution to the problem (19) if*

$$\|2\boldsymbol{X}^{(-i)^\top}(\boldsymbol{X}^{(-i)}\tilde{\boldsymbol{\beta}}_{-i} - \mathbf{x}_i) + \dot{\mathbf{P}}(\tilde{\boldsymbol{\beta}}; b)\|_2 = 0 \tag{20}$$

where

$$\dot{\mathbf{P}}(\tilde{\boldsymbol{\beta}}; b) = [\dot{P}_1(\tilde{\boldsymbol{\beta}}_1; b) \ldots, \dot{P}_{i-1}(\tilde{\boldsymbol{\beta}}_{i-1}; b) \\ \dot{P}_{i+1}(\tilde{\boldsymbol{\beta}}_{i+1}; b) \ldots \dot{P}_n(\tilde{\boldsymbol{\beta}}_n; b)]^\top,$$

$P_k(t; b) = \lambda|t| + T_k(t; b)$ *for* $k = 1, \ldots, n$, $k \neq i$.

Note that in the above definition and the following text, $\dot{P}_k(t; b)$ can be chosen as any value between the right differential $\frac{\partial P_k}{\partial t}(t+; b)$ (or $\dot{P}_k(t+; b)$) and left differential $\frac{\partial P_k}{\partial t}(t-; b)$ (or $\dot{P}_k(t-; b)$) for $k = 1, \ldots, n$, $k \neq i$.

Definition 3. *(Degree of Nonconvexity of a Regularizer) For $\kappa \geq 0$ and $t \in \mathbb{R}$, define*

$$\theta(t, \kappa) := \sup_s \{-\mathrm{sgn}(s-t)(\dot{P}(s; b) - \dot{P}(t; b)) - \kappa|s-t|\}$$

as the degree of nonconvexity for function P. If $\mathbf{u} = (u_1, \ldots, u_n)^\top \in \mathbb{R}^n$, $\theta(\mathbf{u}, \kappa) = [\theta(u_1, \kappa), \ldots, \theta(u_n, \kappa)]$. sgn is the sign function.

Note that $\theta(t, \kappa) = 0$ for convex function P, and it is also used in the analysis of sparse estimation with concave regularization (Zhang and Zhang, 2012).

Let $\hat{\mathbf{S}}_i = \mathrm{supp}(\hat{\mathbf{Z}}^i)$ where $\mathrm{supp}(\cdot)$ indicates the support of a vector, i.e. the indices of its nonzero elements. Denote by $\mathbf{Z}^{i^*}$ the globally optimal solution to (8), and $\mathbf{S}^*_i = \mathrm{supp}(\mathbf{Z}^{i^*})$, then we have

Lemma 1. *For any $1 \leq i \leq n$, if $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$, then the sequence $\{\mathbf{Z}^{i^{(t)}}\}_t$ generated by (17) and (18) converges to a critical point of $F(\mathbf{Z}^i)$, which is denoted by $\hat{\mathbf{Z}}^i$. Moreover, if*

$$0 < b < \min\{\min_{k \in \hat{\mathbf{S}}_i} |\hat{\mathbf{Z}}^i_k|, \max_{k \notin \hat{\mathbf{S}}_i, \mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \neq 0, k \in \Lambda_i} \frac{\gamma \mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{(\frac{\partial Q}{\partial \mathbf{Z}^i_k}|_{\mathbf{Z}^i = \hat{\mathbf{Z}}^i} - \lambda)_+}, \tag{21}$$

$$\min_{k\in\mathbf{S}_i^*}|\mathbf{Z}_k^{i\,*}|, \max_{k\notin\mathbf{S}_i^*,\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}\neq 0,k\in\Lambda_i}\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{(\frac{\partial Q}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i*}}-\lambda)_+}\}$$

(if the denominator is 0, $\frac{\cdot}{0}$ is defined to be $+\infty$ in the above inequality), then both $\hat{\mathbf{Z}}^i$ and $\mathbf{Z}^{i*}$ are local solutions to the capped-ℓ^1 regularized problem (19).

Using the degree of nonconvexity of the regularizer $\mathbf{P}$, we have the following theorem showing that the sub-optimal solution $\hat{\mathbf{Z}}^i$ obtained by our PGD-style iterative method can be close to the globally optimal solution to the original problem (8), i.e. $\mathbf{Z}^{i*}$. In the following text, $\mathbf{B_I}$ indicates a submatrix of $\mathbf{B}$ whose columns correspond to the nonzero elements of $\mathbf{I}$, and $\sigma_{\min}(\cdot)$ indicates the smallest singular value of a matrix.

Theorem 1. (Sub-optimal solution is close to the globally optimal solution) For any $1\leq i\leq n$, let $\tilde{\mathbf{E}}_i=\hat{\mathbf{S}}_i\cup\mathbf{S}_i^*$. Suppose $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}\geq 0$ for all $k\in\Lambda_i$, $\boldsymbol{X}_{\mathbf{E}_i}$ is not singular with $\kappa_0\triangleq\sigma_{\min}(\boldsymbol{X}_{\mathbf{E}_i})>0$, $2\kappa_0^2>\kappa>0$, and b is chosen according to (28) as in Lemma 1. Let $\mathbf{U}_i=(\hat{\mathbf{S}}_i\setminus\mathbf{S}_i^*)\cup(\mathbf{S}_i^*\setminus\hat{\mathbf{S}}_i)$ be the symmetric difference between $\hat{\mathbf{S}}_i$ and $\mathbf{S}_i^*$, then

$$\begin{aligned}&\|\mathbf{Z}^{i*}-\hat{\mathbf{Z}}^i\|_2\\&\leq\frac{1}{2\kappa_0^2-\kappa}\Bigg(\Big(\sum_{k\in\mathbf{U}_i\cap\hat{\mathbf{S}}_i}(\max\{0,\frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b}-\kappa|\hat{\mathbf{Z}}_{ki}-b|\})^2\\&+\sum_{k\in\mathbf{U}_i\setminus\hat{\mathbf{S}}_i}(\max\{0,\frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b}-\kappa b\})^2\Big)^{\frac{1}{2}}+\|\mathbf{t}\|_2\Bigg)\end{aligned}\tag{22}$$

where $\mathbf{t}\in\mathbb{R}^n$, $\mathbf{t}_m=2\lambda\mathbb{I}_{\mathbf{Z}_m^{i\,*}\hat{\mathbf{Z}}_m^i<0}+0\mathbb{I}_{\mathbf{Z}_m^{i\,*}\hat{\mathbf{Z}}_m^i>0}$ for $m\in\hat{\mathbf{S}}_i\cap\mathbf{S}_i^*$, and $\mathbf{t}_m=0$ for all other m.

Remark 2. Note that the bound for distance between the sub-optimal solution and the globally optimal solution presented in Theorem 1 does not require typical Restricted Isometry Property (RIP) conditions as those in (Candès, 2008). Also, when $\frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b}-\kappa|\hat{\mathbf{Z}}_{ki}-b|$ and $\frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b}-\kappa b$ are no greater than 0 and $\mathbf{Z}^{i*}$ and $\hat{\mathbf{Z}}^i$ has the same sign in the intersection of their support, the sub-optimal solution $\hat{\mathbf{Z}}^i$ is equal to the globally optimal solution. When $\frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b}-\kappa|\hat{\mathbf{Z}}_{ki}-b|$ and $\frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b}-\kappa b$ are small positive numbers and $\mathbf{Z}^{i*}$ and $\hat{\mathbf{Z}}^i$ has similar sign in the intersection of their support, $\hat{\mathbf{Z}}^i$ is close to the globally optimal solution.

4 ACCELERATION BY RANDOM PROJECTION

The gradient descent step (11) in the proposed PGD-style iterative method has time complexity of $\mathcal{O}(nd)$, leading to the time complexity of $\mathcal{O}(M_cM_pn^2d)$ for the overall optimization of NRℓ^1-Graph. The literature has extensively employed randomized algorithms for accelerating the numeral computation of different kinds of matrix optimization problems including low rank approximation and matrix decomposition (Frieze et al., 2004; Drineas et al., 2004; Mahoney and Drineas, 2009; Drineas et al., 2011; Halko et al., 2011). In order to accelerate the numerical computation involved in the proposed PGD-style iterative method, we adopt the randomized low rank approximation by random projection (Halko et al., 2011) to obtain a low rank approximation of the data matrix $\boldsymbol{X}$ so as to accelerate the computation of gradient for PGD. Low rank approximation by random projection has also employed in the most recent sparse linear regression method (Zhang et al., 2016). (Zhang et al., 2016) focuses on the convex ℓ^1-regularized sparse linear model, and it remains interesting to explore the provable efficient PGD-style optimization by low rank approximation via random projection on the nonconvex and nonsmooth optimization problem of NRℓ^1-Graph.

Formally, a random matrix $\Omega\in\mathbb{R}^{n\times k}$ is computed such that each element Ω_{ij} is sampled independently from the Gaussian distribution $\mathcal{N}(0,1)$. With the QR decomposition of $\boldsymbol{X}\Omega$, i.e. $\boldsymbol{X}\Omega=\mathbf{QR}$ where $\mathbf{Q}\in\mathbb{R}^{d\times k}$ is an orthogonal matrix of rank k and $\mathbf{R}\in\mathbb{R}^{k\times k}$ is an upper triangle matrix. The columns of $\mathbf{Q}$ form the orthogonal basis for the sample matrix $\boldsymbol{X}\Omega$. Then $\boldsymbol{X}$ is approximated by projecting $\boldsymbol{X}$ onto the range of $\boldsymbol{X}\Omega$: $\mathbf{QQ}^\top\boldsymbol{X}=\mathbf{QW}=\tilde{\boldsymbol{X}}$ where $\mathbf{W}=\mathbf{Q}^\top\boldsymbol{X}\in\mathbb{R}^{k\times n}$. Replacing $\boldsymbol{X}$ with its low rank approximation $\tilde{\boldsymbol{X}}$, we resort to solve the following reduced NRℓ^1-Graph problem termed NRℓ^1-Graph by Random Projection (NRℓ^1-Graph-RP):

$$\begin{aligned}\min_{\mathbf{Z}\in\mathbb{R}^{n\times n},\mathrm{diag}(\mathbf{Z})=\mathbf{0}}\tilde{L}(\mathbf{Z})&=\sum_{i=1}^n\|\mathbf{x}_i-\tilde{\boldsymbol{X}}\mathbf{Z}^i\|_2^2+\lambda\|\mathbf{Z}^i\|_1\\&+\gamma\mathbf{R}_{\mathbf{S}}(\mathbf{Z})\end{aligned}\tag{23}$$

Similar to NRℓ^1-Graph, the PGD-style iterative method is used in each step of coordinate descent for the optimization of NRℓ^1-Graph-RP which solves the following reduced problem for $\mathbf{Z}^i$:

$$\min_{\mathbf{Z}^i\in\mathbb{R}^n,\mathbf{Z}_i^i=0}\tilde{F}(\mathbf{Z}^i)=\|\mathbf{x}_i-\tilde{\boldsymbol{X}}\mathbf{Z}^i\|_2^2+\lambda\|\mathbf{Z}^i\|_1+\gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)\tag{24}$$

And the gradient descent step of the PGD-style iterative method (11) is reduced to

$$\begin{aligned}\tilde{\mathbf{Z}}^{i(t)}&=\mathbf{Z}^{i(t-1)}-\frac{2}{\tau s}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\mathbf{Z}^{i(t-1)}-\tilde{\boldsymbol{X}}^\top\mathbf{x}_i)\\&=\mathbf{Z}^{i(t-1)}-\frac{2}{\tau s}(\mathbf{W}^\top\mathbf{Q}^\top\mathbf{QW}\mathbf{Z}^{i(t-1)}-\mathbf{W}^\top\mathbf{Q}^\top\mathbf{x}_i)\end{aligned}\tag{25}$$

The complexity of this step is reduced from $\mathcal{O}(nd)$ to $\mathcal{O}(nk + dk)$ wherein $k \ll \min\{d, n\}$ and significant efficiency improvement is achieved. Note that the computational cost of QR decomposition for $\boldsymbol{X}\Omega$ is less than $2dk^2$, which is acceptable with a small k.

The subsequent steps of the PGD-style iterative method (12), (13) remain unchanged for NRℓ^1-Graph-RP, therefore, the overall complexity of the optimization of NRℓ^1-Graph-RP by the PGD-style iterative method is reduced from $\mathcal{O}(M_c M_p n^2 d)$ to $\mathcal{O}(M_c M_p n(nk+dk))$ using such randomized low rank decomposition of the data matrix $\boldsymbol{X}$.

(Halko et al., 2011) proved that the low rank approximation $\hat{\boldsymbol{X}}$ is close to $\boldsymbol{X}$ in terms of the spectral norm:

Lemma 2. *(Corollary 10.9 by (Halko et al., 2011)) Let $k_0 \geq 2$ and $p = k - k_0 \geq 4$, then with probability at least $1 - 6e^{-p}$, then the spectral norm of $\boldsymbol{X} - \hat{\boldsymbol{X}}$ is bounded by*

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2 \leq C_{k,k_0} \tag{26}$$

where

$$C_{k,k_0} = \big(1 + 17\sqrt{1 + \frac{k_0}{p}}\big)\sigma_{k_0+1} + \frac{8\sqrt{k}}{p+1}\big(\sum_{j>k_0}\sigma_j^2\big)^{\frac{1}{2}}$$

and $\sigma_1 \geq \sigma_2 \geq \ldots$ are the singular values of $\boldsymbol{X}$.

Let $\tilde{\mathbf{Z}}^i$ be the globally optimal solution to (24), and $\tilde{\mathbf{S}}_i = \mathrm{supp}(\tilde{\mathbf{Z}}^i)$. We have the following theorem establishing the upper bound for the gap between $\tilde{\mathbf{Z}}^i$ and $\mathbf{Z}^{i*}$.

Theorem 2. (Optimal solution to the reduced problem (24) is close to the that to the original problem) For any $1 \leq i \leq n$, let $\mathbf{G}_i = \tilde{\mathbf{S}}_i \cup \mathbf{S}_i^*$. Suppose $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$, $\boldsymbol{X}_{\mathbf{G}_i}$ is not singular with $\tau_0 \triangleq \sigma_{\min}(\boldsymbol{X}_{\mathbf{G}_i}) > 0$, $2\tau_0^2 > \tau > 0$. Then under the conditions of Lemma 2, with probability at least $1 - 6e^{-p}$,

$$\begin{aligned}&\|\mathbf{Z}^{i*} - \tilde{\mathbf{Z}}^i\|_2 \\ &\leq \frac{\big(A_1 + A_2\big)^{\frac{1}{2}} + \|\mathbf{t}\|_2 + C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})}{2\tau_0^2 - \tau}\end{aligned} \tag{27}$$

where $A_1 = \sum_{k\in\mathbf{G}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma \mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\tilde{\mathbf{Z}}_{ki} - b|\})^2$, $A_2 = \sum_{k\in\mathbf{G}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma \mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2$, $\mathbf{t} \in \mathbb{R}^n$, $\mathbf{t}_m = 2\lambda\mathbb{I}_{\mathbf{Z}^{i}_m{}^*\tilde{\mathbf{Z}}^i_m<0} + 0\mathbb{I}_{\mathbf{Z}^{i}_m{}^*\tilde{\mathbf{Z}}^i_m>0}$ for $m \in \tilde{\mathbf{S}}_i \cap \mathbf{S}_i^*$, and $\mathbf{t}_m = 0$ for all other m. Moreover, $A = \frac{(\|\mathbf{x}_i\|_2 + \sqrt{\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{0}) + \|\mathbf{x}_i\|_2^2})}{\sigma_{\min}(\boldsymbol{X}_{\tilde{\mathbf{S}}_i}) - C_{k,k_0}}$, and b satisfies

$$0 < b < \min\{\min_{k\in\tilde{\mathbf{S}}_i}|\tilde{\mathbf{Z}}^i_k|, \max_{k\notin\tilde{\mathbf{S}}_i, \mathbf{F}^{\tilde{\mathbf{S}}}_{ki}\neq 0, k\in\Lambda_i}\frac{\gamma\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{(\frac{\partial Q}{\partial \mathbf{Z}^i_k}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} - \lambda)_+}, \tag{28}$$

$$\min_{k\in\mathbf{S}_i^*}|\mathbf{Z}^{i}_k{}^*|, \max_{k\notin\mathbf{S}_i^*, \mathbf{F}^{\tilde{\mathbf{S}}}_{ki}\neq 0, k\in\Lambda_i}\frac{\gamma\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{(\frac{\partial Q}{\partial \mathbf{Z}^i_k}|_{\mathbf{Z}^i=\mathbf{Z}^{i*}} - \lambda)_+}\}$$

Combining Theorem 1 and Theorem 2, we have the bounded gap between the sub-optimal solution obtained by the PGD-style iterative method for the reduced problem (24) and the globally optimal solution to the original problem (8) for each step of coordinate descent. To the best of our knowledge, our result is among the very few results on the provable efficient optimization by randomized matrix decomposition on nonconvex and nonsmooth optimization problems.

5 EXPERIMENTAL RESULTS

We apply NRℓ^1-Graph to data clustering by performing spectral clustering on the weighted adjacency matrix $\mathbf{W}^*$ of the sparse graph by Algorithm 1. The superior clustering performance of NRℓ^1-Graph is demonstrated by extensive experimental results on various data sets. NRℓ^1-Graph is compared to K-means (KM), Spectral Clustering (SC), ℓ^1-Graph, Sparse Manifold Clustering and Embedding (SMCE) (Elhamifar and Vidal, 2011), and ℓ^2-Rℓ^1-Graph introduced in Section 2. Two measures are used to evaluate the performance of different clustering methods, i.e. the Accuracy (AC) and the Normalized Mutual Information (NMI) (Zheng et al., 2004). Note that we empirically replace $\mathbf{W}$ with the sparse code matrix $\boldsymbol{\alpha}$ in the definition of neighborhood distance (7), which still encourages locally consistent neighborhood while leading to better results, and all the theoretical results hold with $\mathbf{F}^{\tilde{\mathbf{S}}}$ changed accordingly.

5.1 CLUSTERING ON COIL-20, COIL-100, FACE DATA SETS AND MNIST DATA

COIL-20 Database has 1440 images of resolution 32×32 for 20 objects with background removed in all images. The dimensionality of this data is 1024. The extended version of COIL-20, COIL-100 Database, contains 100 objects with 72 images of resolution 32×32 for each object. The images of each object were taken 5 degrees apart when each object was rotated on a turntable. we also demonstrate the clustering result on popular face data sets including Yale-B, CMU PIE, CMU Multi-PIE, UMIST Face Data. The Extended Yale Face Database B contains 64 frontal face images for each of the 38 subjects, and the face images are taken under different illuminations for each subject. CMU PIE face data contains 11554 cropped face images of size 32×32 for 68 persons, and there are around 170 facial images for each person under different illumination and expressions. CMU Multi-PIE (MPIE) data (Gross et al., 2010) contains the

Table 1: Clustering Results on Various Data Sets, the top two results for each measure and data set are in bold.

Data Set	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph	NRℓ^1-Graph-RP
COIL-20	AC	0.6504	0.4271	0.7854	0.7549	0.7854	**0.9174**	**0.9257**
	NMI	0.7616	0.6202	0.9148	0.8754	0.9148	**0.9671**	**0.9716**
COIL-100	AC	0.4928	0.2833	0.5310	0.5625	0.5625	**0.7846**	**0.7972**
	NMI	0.7522	0.5913	0.8015	0.8057	0.8059	**0.9238**	**0.9284**
Yale-B	AC	0.0948	0.1060	0.7850	0.3409	**0.7850**	**0.8111**	0.7092
	NMI	0.1254	0.1524	0.7760	0.3909	0.7760	**0.8095**	**0.7216**
CMU PIE	AC	0.0829	0.0718	0.2318	0.1603	**0.3012**	**0.3190**	0.2965
	NMI	0.1865	0.1760	0.3378	0.3406	**0.5121**	0.4993	**0.5103**
MPIE S1	AC	0.1167	0.1309	0.5892	0.1721	**0.5892**	**0.6582**	0.5729
	NMI	0.5021	0.5289	0.7653	0.5514	0.7653	**0.8540**	**0.8052**
MPIE S2	AC	0.1330	0.1437	0.6994	0.1898	**0.6994**	**0.7226**	0.6584
	NMI	0.4847	0.5145	0.8149	0.5293	0.8149	**0.8826**	**0.8505**
MPIE S3	AC	0.1322	0.1441	0.6316	0.1856	**0.6316**	**0.6753**	0.6194
	NMI	0.4837	0.5150	0.7858	0.5155	0.7858	**0.8657**	**0.8134**
MPIE S4	AC	0.1313	0.1469	0.6803	0.1823	**0.6803**	**0.7260**	0.6563
	NMI	0.4876	0.5251	0.8063	0.5294	0.8066	**0.8926**	**0.8413**
UMIST	AC	0.4216	0.4174	0.4417	0.4452	0.4991	**0.6765**	**0.7078**
	NMI	0.6377	0.6095	0.6489	0.6641	0.6893	**0.7982**	**0.8084**
MNIST	AC	0.5236 ± 0.0092	0.3504 ± 0.0069	0.5714 ± 0.0160	**0.6542** ± 0.0301	0.5561 ± 0.0187	0.6259 ± 0.0249	**0.6296** ± 0.0252
	NMI	0.4770 ± 0.0053	0.3607 ± 0.0087	0.6091 ± 0.0116	**0.6796** ± 0.0185	0.5986 ± 0.0171	**0.6501** ± 0.0196	0.6442 ± 0.0259

facial images captured in four sessions. The UMIST Face Database consists of 575 images of size 112×92 for 20 people. Each person is shown in a range of poses from profile to frontal views. MNIST is comprised of 60000 training images and 10000 test images of ten digits from 0 to 9, and each image is of size 28×28 and represented as a 784-dimensional vector.

All the clustering results on various data sets are shown in Table 1. For MNIST data, we randomly sample 1000 images from each class to constitute a total number of 10000 images on which clustering is performed. We conduct such random sampling by 10 times and record the average performance. It can be observed from Table 1 that ℓ^2-Rℓ^1-Graph produces better clustering accuracy than ℓ^1-Graph on COIL-100, since graph regularization produces locally smooth sparse codes aligned to the local manifold structure of the data. Using the neighborhood regularization term to render locally consistent neighborhood in the sparse graph so that the local structure of the sparse graph is aligned to the manifold structure of the data, NRℓ^1-Graph and NRℓ^1-Graph-RP always perform better than all the other clustering methods. Randomized rank-k decomposition of the data matrix is employed in NRℓ^1-Graph-RP to accelerate NRℓ^1-Graph with $k = \frac{\min\{d,n\}}{10}$. Compared to NR$\ell^1$-Graph, we observe that NRℓ^1-Graph-RP usually exhibits competitive or even better results, revealing its compelling performance with reduced computational complexity. For example, NRℓ^1-Graph-RP is around 8.7 times faster than NRℓ^1-Graph for each iteration of the PGD-style iterative method on the COIL-100 data, which is consistent to our complexity analysis in Section 4.

5.2 PARAMETER SETTING

There are two essential parameters for NRℓ^1-Graph, i.e. γ for the neighborhood induced graph regularization term and K for building the KNN graph. We use the sparse codes generated by ℓ^1-Graph with weighting parameter $\lambda^{(\ell^1)} = 0.1$ in (2) to initialize both NRℓ^1-Graph and ℓ^2-Rℓ^1-Graph, and set $\lambda = \gamma = 0.1$, $K = 5$ for NRℓ^1-Graph (6) empirically throughout all the experiments. The maximum iteration number M is 100 and the stopping threshold ε is 10^{-5}. The weighting parameter for the ℓ^1-norm in ℓ^1-Graph, $\lambda^{(\ell^1)}$, and that in ℓ^2-Rℓ^1-Graph, $\lambda^{(\ell^2)}$, and the regularization weight for ℓ^2-Rℓ^1-Graph, $\gamma^{(\ell^2)}$, are chosen from $[0.1, 1]$ for the best performance. Moreover, results on the parameter sensitivity are included in the supplementary of this paper.

6 CONCLUSION

We propose a novel NRℓ^1-Graph which align the sparse graph to the local manifold structure of the data by learning locally consistent neighborhood in the sparse graph. In contrast to most existing methods that use ℓ^2-norm to measure the distance between sparse codes in graph regularization, NRℓ^1-Graph employs the novel neighborhood distance to measure the distance between data points so as to impose the local smoothness on the neighborhood. The optimization algorithm using the proposed PGD-style iterative method and its theoretical properties are analyzed. The optimization of NRℓ^1-Graph is further accelerated by randomized low rank decomposition of the data matrix with theoretical guarantee, leading to the more efficient NRℓ^1-Graph by Random Projection. The effectiveness of NRℓ^1-Graph and its accelerated version for data clustering is demonstrated by experiments on various data sets.

References

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, August 2014. ISSN 0025-5610.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(910):589 – 592, 2008. ISSN 1631-073X.

Bin Cheng, Jianchao Yang, Shuicheng Yan, Yun Fu, and Thomas S. Huang. Learning with l1-graph for image analysis. *IEEE Transactions on Image Processing*, 19(4):858–866, 2010.

F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1):9–33, 2004. ISSN 1573-0565.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011. ISSN 0945-3245.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *CVPR*, pages 2790–2797, 2009.

Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *NIPS*, pages 55–63, 2011.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

Chris Fraley and Adrian E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97(458):611–631, June 2002. ISSN 0162-1459.

Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315:2007, 2007.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6):1025–1041, November 2004. ISSN 0004-5411.

Shenghua Gao, Ivor Wai-Hung Tsang, and Liang-Tien Chia. Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):92–104, 2013.

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, May 2010. ISSN 0262-8856.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011. ISSN 0036-1445.

Xiaofei He, Deng Cai, Yuanlong Shao, Hujun Bao, and Jiawei Han. Laplacian regularized gaussian mixture model for data clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 23(9):1406–1418, Sept 2011. ISSN 1041-4347.

Jialu Liu, Deng Cai, and Xiaofei He. Gaussian mixture model with local consistency. In *AAAI*, 2010.

Michael W. Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. Deep subspace clustering with sparsity prior. In *Proceedings of the 25 International Joint Conference on Artificial Intelligence*, pages 1925–1931, New York, NY, USA, 9-15 July 2016.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290: 2323–2326, 2000.

Mahdi Soltanolkotabi, Ehsan Elhamifar, and Emmanuel J. Candès. Robust subspace clustering. *Ann. Statist.*, 42(2): 669–699, 04 2014.

Yu-Xiang Wang and Huan Xu. Noisy sparse subspace clustering. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 89–97, 2013.

Shuicheng Yan and Huan Wang. Semi-supervised learning by sparse representation. In *SDM*, pages 792–801, 2009.

Yingzhen Yang, Zhangyang Wang, Jianchao Yang, Jiangping Wang, Shiyu Chang, and Thomas S. Huang. Data clustering by laplacian regularized l1-graph. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 3148–3149, 2014.

Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2005.

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593, 11 2012.

Weizhong Zhang, Lijun Zhang, Rong Jin, Deng Cai, and Xiaofei He. Accelerated sparse linear regression via random projection. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2337–2343, 2016.

Miao Zheng, Jiajun Bu, Chun Chen, Can Wang, Lijun Zhang, Guang Qiu, and Deng Cai. Graph regularized sparse coding for image representation. *IEEE Transactions on Image Processing*, 20(5):1327–1336, 2011.

Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality preserving clustering for image database. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pages 885–891, New York, NY, USA, 2004. ACM.

7 SUPPLEMENTARY

7.1 PROOFS

Proof of Proposition 1. Note that $\mathbf{u}$ is the optimal solution to the lasso problem: $\mathbf{u} = \arg\min_{\mathbf{v}\in R^n} \frac{\tau s}{2}\|\mathbf{v} - \tilde{\mathbf{Z}}_{ki}^{(t)}\|_2^2 + \lambda\|\mathbf{v}\|_1$.

Suppose $k \in \Lambda_i$. Define $T_k(v) = \frac{\tau s}{2}(v - \tilde{\mathbf{Z}}_{ki}^{(t)})^2 + \lambda|v|$ for $v \in \mathbb{R}$, then $\mathbf{u}_k = \arg\min_{v\in\mathbb{R}} T_k(v)$. Since the two functions $H_k(v)$ and $T_k(v)$ only differ at $v = 0$, $\arg\min_{v\in\{\mathbf{u}_k,0\}} H_k(v)$ is the optimal solution to $\min_{v\in R} H_k(v)$ when $\mathbf{u}_k \neq 0$ or $\mathbf{u}_k = 0$ and $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \geq 0$.

When $\mathbf{u}_k = 0$ and $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} < 0$, when $\varepsilon \to 0$ and $\varepsilon \neq 0$, $H_k(\varepsilon) \to \frac{\tau s}{2}(\tilde{\mathbf{Z}}_{ki}^{(t)})^2 + \gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}$ and $\inf_{v\in\mathbb{R}} H_k(v) = \frac{\tau s}{2}(\tilde{\mathbf{Z}}_{ki}^{(t)})^2 + \gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}$. Note that the infimum can never be achieved. Since $\inf_{v\in\mathbb{R}} H_k(v) < H_k(\mathbf{Z}_{ki}^{(t-1)})$, we can always find $\varepsilon \neq 0$ such that $H_k(\varepsilon) \leq H_k(\mathbf{Z}_{ki}^{(t-1)})$.

Suppose $k \notin \Lambda_i$, then $\mathbf{u}_k$ is the optimal solution to $\min_{v\in R} H_k(v)$ for $k \neq i$.

Define $H(\mathbf{v}) = \frac{\tau s}{2}\|\mathbf{v} - \tilde{\mathbf{Z}}^{i^{(t)}}\|_2^2 + \lambda\|\mathbf{v}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{v})$. Based on the above argument, $H(\mathbf{Z}^{i^{(t)}}) \leq H(\mathbf{Z}^{i^{(t-1)}})$ which indicates that

$$\frac{\tau s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2 + \langle \mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}, \nabla Q(\mathbf{Z}^{i^{(t-1)}})\rangle \tag{29}$$

$$+ \lambda\|\mathbf{Z}^{i^{(t)}}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^{i^{(t)}}) \leq \lambda\|\mathbf{Z}^{i^{(t-1)}}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^{i^{(t-1)}}) \tag{30}$$

Also, since s is the Lipschitz constant for the gradient of function $Q(\cdot)$, we have

$$Q(\mathbf{Z}^{i^{(t)}}) \leq Q(\mathbf{Z}^{i^{(t-1)}}) + \langle \mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}, \nabla Q(\mathbf{Z}^{i^{(t-1)}})\rangle \tag{31}$$

$$+ \frac{s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2$$

Combining (29) and (31),

$$F(\mathbf{Z}^{i^{(t)}}) \leq F(\mathbf{Z}^{i^{(t-1)}}) - \frac{(\tau-1)s}{2}\|\mathbf{Z}^{i^{(t)}} - \mathbf{Z}^{i^{(t-1)}}\|_2^2$$

□

Proof of Lemma 1. We first prove that the sequences $\{\mathbf{Z}^{i^{(t)}}\}_t$ is bounded for any $1 \leq i \leq n$. By Proposition 1, the sequence $\{F(\mathbf{Z}^{i^{(t)}})\}_t$ decreases, so we have

$$F(\mathbf{Z}^{i^{(t)}}) = \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^{i^{(t)}}\|_2^2 + \lambda\|\mathbf{Z}^{i^{(t)}}\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^{i^{(t)}}) \leq F(\mathbf{Z}^{i^{(0)}})$$

for $t \geq 1$. Therefore,

$$\|\mathbf{Z}^{i^{(t)}}\|_1 \leq \frac{1}{\lambda}F(\mathbf{Z}^{i^{(0)}})$$

It follows that $\|\mathbf{Z}^{i^{(t)}}\|_1$ is bounded, and $\|\mathbf{Z}^{i^{(t)}}\|_2$ is also bounded. Since $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \geq 0$ for all $k \in \Lambda_i$ and the indicator function $\mathbb{1}_{\cdot\neq 0}$ is semi-algebraic function, $\mathbf{R}_{\tilde{\mathbf{S}}}(\cdot)$ is also a semi-algebraic function and lower semicontinuous. Therefore, according to Theorem 1 by Bolte et al. (2014), $\{\mathbf{Z}^{i^{(t)}}\}_t$ converges to a critical point of $F(\mathbf{Z}^i)$, denoted by $\hat{\mathbf{Z}}^i$.

Let $\hat{\mathbf{v}} = 2\boldsymbol{X}^{(-i)^\top}(\boldsymbol{X}^{(-i)}\hat{\mathbf{Z}^i}_{-i} - \mathbf{x}_i) + \dot{\mathbf{P}}(\hat{\mathbf{Z}^i}_{-i}; b)$. For k such that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} = 0$ or $k \notin \Lambda_i$, since $\hat{\mathbf{Z}}^i$ is a critical point of $F(\mathbf{Z}^i)$, $\hat{\mathbf{v}}_{k_{-i}} = 0$.

Now we consider the case that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \neq 0$ and $k \in \Lambda_i$. In the following text we denote by k_{-i} the index of the element of the vector $\mathbf{v}_{-i}$ corresponding to the element $\mathbf{v}_k$ of $\mathbf{v}$, i.e. $(\mathbf{v}_{-i})_{k_{-i}} = \mathbf{v}_k$ for any $\mathbf{v} \in \mathbb{R}^n$.

For $k \in \hat{\mathbf{S}}_i$, since $\hat{\mathbf{Z}}^i$ is a critical point of $F(\mathbf{Z}^i) = \|\mathbf{x}_i - \boldsymbol{X}\mathbf{Z}^i\|_2^2 + \lambda\|\mathbf{Z}^i\|_1 + \gamma\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)$. then $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$ because $\frac{\partial\mathbf{R}_{\tilde{\mathbf{S}}}(\mathbf{Z}^i)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$. Note that $\min_{k\in\hat{\mathbf{S}}_i}|\hat{\mathbf{Z}_k^i}| > b$, so $\frac{\partial\mathbf{T}}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$. It follows that $\hat{\mathbf{v}}_{k_{-i}} = 0$.

For $k \notin \hat{\mathbf{S}}_i$, since $\frac{dP_k}{d\mathbf{Z}_k^i}(\hat{\mathbf{Z}}_k^i+; b) = \frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} + \lambda$ and $\frac{dP_k}{d\mathbf{Z}_k^i}(\hat{\mathbf{Z}}_k^i-; b) = -\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} - \lambda$, $\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} + \lambda \geq |\frac{\partial Q}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i}|$, we can choose the k_{-i}-th element of $\dot{\mathbf{P}}(\hat{\mathbf{Z}}^i; b)$ such that $\hat{\mathbf{v}}_{k_{-i}} = 0$. Therefore, $\|\hat{\mathbf{v}}\|_2 = 0$, and $\hat{\mathbf{Z}}^i$ is a local solution to the problem (19).

Now we prove that $\mathbf{Z}^{i^*}$ is also a local solution to (19). Let $\mathbf{v}^* = 2\boldsymbol{X}^{(-i)^\top}(\boldsymbol{X}^{(-i)}\mathbf{Z}^{i^*}_{-i} - \mathbf{x}_i) + \dot{\mathbf{P}}(\mathbf{Z}^{i^*}_{-i}; b)$, and Q is defined as before. For k such that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} = 0$ or or $k \notin \Lambda_i$, since $\mathbf{Z}^{i^*}$ is the globally optimal solution of $F(\mathbf{Z}^i)$, $\mathbf{v}^*_{k_{-i}} = 0$.

Again we consider the case that $\mathbf{F}_{ki}^{\tilde{\mathbf{S}}} \neq 0$ and $k \in \Lambda_i$.

For $k \in \mathbf{S}_i^*$, since $\mathbf{Z}^{i^*}$ is the globally optimal solution to problem (8), we also have $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i^*}} = 0$. If it is not the case and $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i^*}} \neq 0$, then we can change $\mathbf{Z}_k^i$ by a small amount in the direction of the gradient $\frac{\partial(Q+\lambda\|\mathbf{Z}^i\|_1)}{\partial\mathbf{Z}_k^i}$ at the point $\mathbf{Z}^i = \mathbf{Z}^{i^*}$ while $\mathbf{Z}_k^i$ is still nonzero, leading to a smaller value of the objective $F(\mathbf{Z}^i)$.

Note that $\min_{k\in\mathbf{S}_i^*}|\mathbf{Z}_k^{i^*}| > b$, so $\frac{\partial\mathbf{T}}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\hat{\mathbf{Z}}^i} = 0$, and it follows that $\mathbf{v}^*_{k_{-i}} = 0$.

For $k \notin \mathbf{S}_i^*$, since $\frac{\gamma\mathbf{F}_{ki}^{\tilde{\mathbf{S}}}}{b} + \lambda \geq \max_{k\notin\hat{\mathbf{S}}_i}|\frac{\partial Q}{\partial\mathbf{Z}_k^i}|_{\mathbf{Z}^i=\mathbf{Z}^{i*}}|$, we can choose the k_{-i}-th element of $\dot{\mathbf{P}}(\mathbf{Z}^{i^*}; b)$ such that $\mathbf{v}^*_{k_{-i}} = 0$. It follows that $\|\mathbf{v}^*\|_2 = 0$, and $\mathbf{Z}^{i^*}$ is also a

local solution to the problem (19). □

Proof of Theorem 1. According to Lemma 1, both $\hat{\mathbf{Z}}^i$ and $\mathbf{Z}^{i*}$ are local solutions to problem (19). In the following text, let $\boldsymbol{\beta}_{\mathbf{I}}$ indicates a vector whose elements are those of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$. Let $\Delta = \mathbf{Z}^{i*}_{-i} - \hat{\mathbf{Z}}^i_{-i}$, $\tilde{\Delta} = \dot{\mathbf{P}}(\mathbf{Z}^{i*}) - \dot{\mathbf{P}}(\hat{\mathbf{Z}}^i)$. By Lemma 1, we have

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \tilde{\Delta}\|_2 = 0$$

It follows that

$$\begin{aligned}&2\Delta^\top \boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \Delta^\top\tilde{\Delta}\\ &\le \|\Delta\|_2\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \tilde{\Delta}\|_2 = 0\end{aligned}$$

Also, by the proof of Lemma 1, for $k \in \hat{\mathbf{S}}_i \cap \mathbf{S}^*_i$, $(2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta)_{k_{-i}} = 2\lambda\mathbb{I}_{\mathbf{Z}^{i*}_k\hat{\mathbf{Z}}^i_k<0} + 0\mathbb{I}_{\mathbf{Z}^{i*}_k\hat{\mathbf{Z}}^i_k>0}$. We now present another property on any nonconvex function P using the degree of nonconvexity in Definition 3: $\theta(t,\kappa) := \sup_s\{-\mathrm{sgn}(s-t)(\dot{P}(s;b)-\dot{P}(t;b)) - \kappa|s-t|\}$ on the regularizer $\mathbf{P}$. For any $s,t \in \mathbb{R}$, we have

$$-\mathrm{sgn}(s-t)\big(\dot{P}(s;b) - \dot{P}(t;b)\big) - \kappa|s-t| \le \theta(t,\kappa)$$

by the definition of θ. It follows that

$$\begin{aligned}&\theta(t,\kappa)|s-t| \ge -(s-t)\big(\dot{P}(s;b) - \dot{P}(t;b)\big) - \kappa(s-t)^2\\ &-(s-t)\big(\dot{P}(s;b) - \dot{P}(t;b)\big) \le \theta(t,\kappa)|s-t| + \kappa(s-t)^2\end{aligned} \tag{32}$$

Let $\hat{\mathbf{S}}^{-i}_i = \mathrm{supp}(\hat{\mathbf{Z}}^i_{-i})$, $\mathbf{S}^{-i*}_i = \mathrm{supp}(\mathbf{Z}^{i*}_{-i})$, $\mathbf{U}^{-i}_i = (\hat{\mathbf{S}}^{-i}_i \backslash \mathbf{S}^{-i*}_i) \cup (\mathbf{S}^{-i*}_i \backslash \hat{\mathbf{S}}^{-i}_i)$. Applying (32) with $P = P_k$ for $k = 1,\dots,n$, $k \ne i$, we have

$$\begin{aligned}&2\Delta^\top\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta \le -\Delta^\top\tilde{\Delta}\\ &= -\Delta^\top_{\mathbf{U}^{-i}_i}\tilde{\Delta}_{\mathbf{U}^{-i}_i} - \Delta^\top_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\\ &\le |(\mathbf{Z}^{i*}_{-i})_{\mathbf{U}^{-i}_i} - (\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i}|^\top\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\\ &+\kappa\|(\mathbf{Z}^{i*}_{-i})_{\mathbf{U}^{-i}_i} - (\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i}\|_2^2 + \|\Delta_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\\ &\le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\|(\mathbf{Z}^{i*}_{-i})_{\mathbf{U}^{-i}_i} - (\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i}\|_2\\ &+\kappa\|\Delta\|_2^2 + \|\Delta\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\\ &\le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\|\Delta\|_2 + \kappa\|\Delta\|_2^2 + \|\Delta\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\end{aligned} \tag{33}$$

On the other hand, $\Delta^\top\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta \ge \kappa_0^2\|\Delta\|_2^2$. It follows from (33) that

$$\begin{aligned}&2\kappa_0^2\|\Delta\|_2^2 \le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\|\Delta\|_2 + \kappa\|\Delta\|_2^2\\ &+\|\Delta\|_2\|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\end{aligned}$$

When $\|\Delta\|_2 \ne 0$, we have

$$2\kappa_0^2\|\Delta\|_2 \le \|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2 + \kappa\|\Delta\|_2 + \|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i}_i}\|_2$$

$$\Rightarrow \|\Delta\|_2 \le \frac{\|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2 + \|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2}{2\kappa_0^2 - \kappa} \tag{34}$$

According to the definition of θ, it can be verified that $\theta((\hat{\mathbf{Z}}^i_{-i})_{k_{-i}},\kappa) = \max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\hat{\mathbf{Z}}_{ki} - b|\}$ for $k_{-i} \in \mathbf{U}^{-i}_i \cap \hat{\mathbf{S}}^{-i}_i$, and $\theta((\hat{\mathbf{Z}}^i_{-i})_{k_{-i}},\kappa) = \max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\}$ for $k_{-i} \in \mathbf{U}^{-i}_i \setminus \hat{\mathbf{S}}^{-i}_i$. Therefore,

$$\begin{aligned}&\|\theta((\hat{\mathbf{Z}}^i_{-i})_{\mathbf{U}^{-i}_i},\kappa)\|_2\\ &= \Big(\sum_{k\in\mathbf{U}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\hat{\mathbf{Z}}_{ki} - b|\})^2 +\\ &\sum_{k\in\mathbf{U}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2\Big)^{\frac{1}{2}}\end{aligned} \tag{35}$$

and it follows that

$$\begin{aligned}&\|\mathbf{Z}^{i*} - \hat{\mathbf{Z}}^i\|_2 = \|\Delta\|_2\\ &\le \frac{1}{2\kappa_0^2 - \kappa}\Bigg(\Big(\sum_{k\in\mathbf{U}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\hat{\mathbf{Z}}_{ki} - b|\})^2 +\\ &\sum_{k\in\mathbf{U}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2\Big)^{\frac{1}{2}} + \|\tilde{\Delta}_{\hat{\mathbf{S}}^{-i}_i\cap\mathbf{S}^{-i*}_i}\|_2\Bigg)\end{aligned} \tag{36}$$

where $\tilde{\Delta}_{m_{-i}} = -(2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta)_{m_{-i}} = -2\lambda\mathbb{I}_{\mathbf{Z}^{i*}_m\hat{\mathbf{Z}}^i_m<0} - 0\mathbb{I}_{\mathbf{Z}^{i*}_m\hat{\mathbf{Z}}^i_m>0}$ for $m \in \hat{\mathbf{S}}_i \cap \mathbf{S}^*_i$. This proves the result of this theorem. □

Proof of Theorem 2. Let $\boldsymbol{Y} = \tilde{\boldsymbol{X}}$. By the proof of Lemma 1, we have

$$\|2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2 = 0$$

It follows that

$$\begin{aligned}&\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2\\ &= \|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} - 2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\\ &+2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2\\ &\le \|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} - 2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\|_2\\ &+\|2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2\\ &= \|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} - 2\boldsymbol{Y}^{(-i)\top}\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\|_2\\ &\le \|2\boldsymbol{X}^{(-i)\top}(\boldsymbol{X}^{(-i)} - \boldsymbol{Y}^{(-i)})\tilde{\mathbf{Z}}^i_{-i}\|_2\\ &+\|2(\boldsymbol{X}^{(-i)} - \boldsymbol{Y}^{(-i)})^\top\boldsymbol{Y}^{(-i)}\tilde{\mathbf{Z}}^i_{-i}\|_2\end{aligned} \tag{37}$$

By $\tilde{F}(\mathbf{Z}^i) \le \tilde{F}(\mathbf{0})$, we have $\|\tilde{\mathbf{Z}}^i_{-i}\|_2 \le A$. Let $k_0 \ge 2$ and $p = k - k_0 \ge 4$. By Lemma 2, with probability at least $1 - 6e^{-p}$, $\|\boldsymbol{X} - \boldsymbol{Y}\|_2 \le C_{k,k_0}$. It follows from (37) that

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\tilde{\mathbf{Z}}^i_{-i} + \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^i)\|_2$$

$$\leq \sigma_{\max}(\boldsymbol{X})C_{k,k_0}A + C_{k,k_0}(\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})A$$
$$= C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})$$

Also, by Lemma 1,

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\mathbf{Z}^{i*}_{-i} + \dot{\mathbf{P}}(\mathbf{Z}^{i*}_{-i})\|_2 = 0$$

Let $\Delta = \mathbf{Z}^{i*}_{-i} - \tilde{\mathbf{Z}}^{i}_{-i}$, $\tilde{\Delta} = \dot{\mathbf{P}}(\mathbf{Z}^{i*}) - \dot{\mathbf{P}}(\tilde{\mathbf{Z}}^{i})$.

$$\|2\boldsymbol{X}^{(-i)\top}\boldsymbol{X}^{(-i)}\Delta + \tilde{\Delta}\|_2 \leq C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})$$

Now following the proof of Theorem 1, we have

$$\|\mathbf{Z}^{i*} - \tilde{\mathbf{Z}}^{i}\|_2 = \|\Delta\|_2$$
$$\leq \frac{1}{2\tau_0^2 - \tau}\Bigg(\Big(\sum_{k\in\mathbf{G}_i\cap\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa|\tilde{\mathbf{Z}}_{ki} - b|\})^2$$
$$+ \sum_{k\in\mathbf{G}_i\setminus\hat{\mathbf{S}}_i}(\max\{0, \frac{\gamma\mathbb{I}_{\mathbf{Z}_{ik}=0}\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}}{b} - \kappa b\})^2\Big)^{\frac{1}{2}} + \|\mathbf{t}\|_2$$
$$+ C_{k,k_0}A(2\sigma_{\max}(\boldsymbol{X}) + C_{k,k_0})\Bigg) \quad (38)$$

□

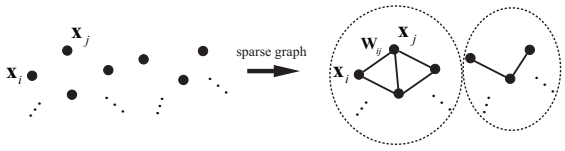


Figure 2: Diagram of building a sparse graph over the input data. The black dots represent the data points, and there is an edge between every two points $\mathbf{x}_i$ and $\mathbf{x}_j$ if their similarity $\mathbf{W}_{ij}$ calculated by the corresponding sparse codes is nonzero. Clustering based on the sparse graph leads to two clusters compassed by the two dashed ellipses.

7.2 MORE DETAILS IN THE PAPER AND MORE EXPERIMENTAL RESULTS

Figure 2 illustrates the diagram of building sparse graph over the data by the representative sparse graph based clustering methods such as ℓ^1-Graph (Yan and Wang, 2009; Cheng et al., 2010) and Sparse Subspace Clustering (SSC) (Elhamifar and Vidal, 2013). The SC baseline in this paper uses the self-tuning spectral clustering method (Zelnik-manor and Perona, 2005), and we choose this method due to its advantage of adaptively setting the kernel bandwidth for the Gaussian kernel similarity. More concretely, we construct a similarity matrix using Gaussian kernel, and the bandwidth of the Gaussian kernel similarity between two points is determined by the local statistics of the neighborhoods of these two points. We set the distance to the 7-th nearest neighbor as the local statistics, which is also the default choice suggested by the paper, then perform spectral clustering on such similarity matrix to obtain the clustering results for SC in Table 1. In addition, various sparse graph methods, including ℓ^1-Graph, ℓ^2-Rℓ^1-Graph and NRℓ^1-Graph, constructs a sparse graph upon which spectral clustering is applied to find the clusters.

It is worthwhile to mention the meaning of the condition that $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$ in (9). Let $k \in \Lambda_i$, if the number of point $\mathbf{x}_i$'s neighbors with zero k-th element of the sparse codes is larger than that with nonzero k-th element of the sparse codes, which indicates that the neighbors of $\mathbf{x}_i$ suggest that a zero k-th element of the sparse code of $\mathbf{x}_i$ is preferable, then $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki} \geq 0$ and $\mathbf{F}^{\tilde{\mathbf{S}}}_{ki}$ quantitatively measures the penalty if the sparse code element $\mathbf{Z}^i_k$ is nonzero while the neighbors of $\mathbf{x}_i$ suggest that $\mathbf{Z}^i_k = 0$ is preferable. The optimization helps point $\mathbf{x}_i$ make a sensible choice by considering the suggestion of its neighbors. We observe that $\mathbf{F}^{\hat{\mathbf{S}}}_{ki} \geq 0$ for all $k \in \Lambda_i$ happens in all the data sets used in the experiments.

We have conducted paired t-test and conclude that both NRℓ^1-Graph and NRℓ^1-Graph-RP are statistically better than other baseline methods with p-value less than 0.05 in many cases. For example, the p-value of the paired t-test between the accuracy of NRℓ^1-Graph and SMCE is less than 0.05 on the COIL-20, COIL-100 and Yale-B data.

We also present clustering results on the first c clusters for COIL-100, CMU PIE and UMIST Face Data in Table 2, 3 and 4 respectively.

In order to investigate the parameter sensitivity of our model, namely how the performance of NRℓ^1-Graph varies with parameter γ and K, we vary γ and K and illustrate the result on the UMIST Face Database in Figure 3 and Figure 4 respectively in this supplementary. The performance of NRℓ^1-Graph is noticeably better than other competing algorithms over a relatively large range of both λ and K, which demonstrates the robustness of our algorithm with respect to the parameter settings. We also observe that a too small K (near to 1) results in under regularization, and a too big K (near to 15) or too big γ (close to 0.45) risks over regularization.

Table 2: Clustering Results on COIL-100 Database. c in the left column is the cluster number, i.e. the first c clusters of the entire data are used for clustering. c has the same meaning in the following tables.

COIL-100 / # Clusters	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph
c = 20	AC	0.5875	0.4493	0.5340	0.6208	0.6681	**0.9236**
	NMI	0.7448	0.6680	0.7681	0.7993	0.7933	**0.9610**
c = 40	AC	0.5774	0.4160	0.5819	0.6028	0.5944	**0.8771**
	NMI	0.7662	0.6682	0.7911	0.7919	0.7991	**0.9504**
c = 60	AC	0.5330	0.3225	0.5824	0.5877	0.6009	**0.7808**
	NMI	0.7603	0.6254	0.8310	0.7971	0.8310	**0.8924**
c = 80	AC	0.5062	0.3135	0.5380	0.5740	0.5632	**0.8177**
	NMI	0.7458	0.6071	0.8034	0.7931	0.8036	**0.9208**
c = 100	AC	0.4928	0.2833	0.5310	0.5625	0.5625	**0.7846**
	NMI	0.7522	0.5913	0.8015	0.8057	0.8059	**0.9238**

Table 3: Clustering Results on CMU PIE Data

CMU PIE / # Clusters	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph
c = 20	AC	0.1327	0.1288	0.2435	0.2321	0.3212	**0.3606**
	NMI	0.1220	0.1342	0.2895	0.2942	0.4007	**0.4876**
c = 40	AC	0.1054	0.0867	0.2443	0.1752	0.3412	**0.3555**
	NMI	0.1534	0.1422	0.3344	0.2976	0.4789	**0.4834**
c = 68	AC	0.0829	0.0718	0.2318	0.1603	0.3012	**0.3190**
	NMI	0.1865	0.1760	0.3378	0.3406	**0.5121**	0.4993

Table 4: Clustering Results on UMIST Face Data

UMIST Face / # Clusters	Measure	KM	SC	ℓ^1-Graph	SMCE	ℓ^2-Rℓ^1-Graph	NRℓ^1-Graph
c = 8	AC	0.4330	0.4789	0.4930	0.4695	0.5399	**0.6056**
	NMI	0.5373	0.5236	0.5516	0.5744	0.5721	**0.5749**
c = 12	AC	0.4478	0.4655	0.5195	0.4955	0.5706	**0.6246**
	NMI	0.6121	0.6049	0.6086	0.6445	0.6994	**0.7244**
c = 16	AC	0.4297	0.4539	0.4539	0.4747	0.4700	**0.6982**
	NMI	0.6343	0.6453	0.6582	0.6909	0.6714	**0.7816**
c = 20	AC	0.4216	0.4174	0.4417	0.4452	0.4991	**0.6765**
	NMI	0.6377	0.6095	0.6489	0.6641	0.6893	**0.7982**

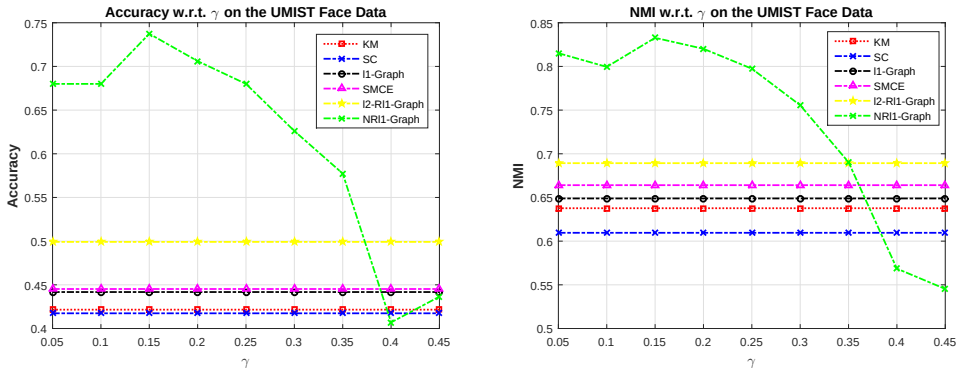


Figure 3: Clustering performance with different values of γ, i.e. the weight for the regularization term in NRℓ^1-Graph, on the UMIST Face Data. Left: Accuracy; Right: NMI

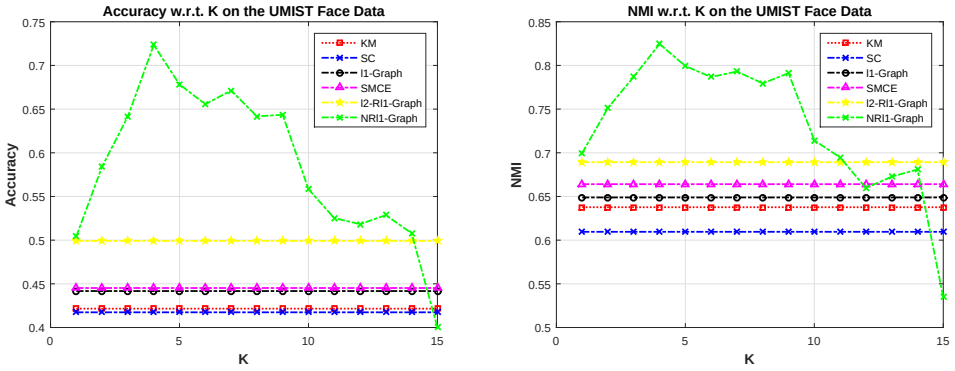


Figure 4: Clustering performance with different values of K, i.e. the number of nearest neighbors for the regularization term in NRℓ^1-Graph, on the UMIST Face Data. Left: Accuracy; Right: NMI