Journal of Machine Learning Research 25 (2024) 1-42 Submitted 10/21; Revised 11/23; Published 1/24

Pursuit of the Cluster Structure of Network Lasso: Recovery Condition and Non-convex Extension

Shotaro Yagishita a15.fjng@g.chuo-u.ac.jp
Department of Industrial and Systems Engineering
Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan

Jun-ya Gotoh jgoto@kc.chuo-u.ac.jp
Department of Data Science for Business Innovation
Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan

Editor: Mladen Kolar

Abstract

Network lasso (NL for short) is a technique for estimating models by simultaneously clustering data samples and fitting the models to them. It often succeeds in forming clusters thanks to the geometry of the sum of ℓ_2 norm employed therein, but there may be limitations due to the convexity of the regularizer. This paper focuses on clustering generated by NL and strengthens it by creating a non-convex extension, called network trimmed lasso (NTL for short). Specifically, we initially investigate a sufficient condition that guarantees the recovery of the latent cluster structure of NL on the basis of the result of Sun et al. (2021) for convex clustering, which is a special case of NL for ordinary clustering. Second, we extend NL to NTL to incorporate a cardinality (or, ℓ_0-)constraint and rewrite the constrained optimization problem defined with the ℓ_0 norm, a discontinuous function, into an equivalent unconstrained continuous optimization problem. We develop ADMM algorithms to solve NTL and show their convergence results. Numerical illustrations indicate that the non-convex extension provides a more clear-cut cluster structure when NL fails to form clusters without incorporating prior knowledge of the associated parameters.

Keywords: sparse modeling, clustering, network lasso, network trimmed lasso, alternating direction method of multipliers (ADMM)

1. Introduction

In data analysis, fundamental methodologies such as regression and clustering can be enhanced by coupling with side information concerning the underlying structure of the data set. For instance, consider a scenario where a batch of data samples comprises outcomes from multiple sources, and their relationship is (partially or fully) understood. In such a context, multiple models can be estimated to fit the data set and identify sample clusters. To tackle such tasks, network lasso (NL for short) has recently been proposed by Hallac et al. (2015) and is considered effective.

Let $a_i \in \mathbb{R}^p$ and $b_i \in \mathbb{R}$ denote p inputs and a real-valued output, respectively, of the i-th sample, $i \in [n] := \{1, ..., n\}$, and let us suppose that certain samples are known to be similar. If such similarity for $i, j \in [n]$ is given by non-negative weights, $\tilde{w}_{\{i,j\}}$, the NL

©2024 Shotaro Yagishita and Jun-ya Gotoh.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/21-1190.html.

version of the ordinary least squares regression can be formulated as the following convex optimization:

$$\underset{x_1,\ldots,x_n\in\mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\sum_{i\in[n]}(b_i - a_i^\top x_i)^2 + \gamma \sum_{i,j\in[n]:i<j} \tilde{w}_{\{i,j\}}\|x_i - x_j\|_2, \tag{1}$$

where $\|z\|_2 := \sqrt{\sum_{j=1}^p z_j^2}$ denotes the ℓ_2 norm of a vector $z \in \mathbb{R}^p$, and $\gamma > 0$ is a parameter to be tuned so as to strike a balance between the first and second terms of (1). Intuitively, minimizing the first term encourages the model ($b = a^\top x_i$) to adjust to each sample (a_i, b_i), while minimizing the second term accelerates the merging of alike samples, as weight $\tilde{w}_{\{i,j\}}$ is substantial when samples i and j have similarities. Especially, the second term is the sum of ℓ_2 norms and, for large γ, an optimal solution $(x_1^\star, ..., x_n^\star)$ is expected to satisfy $\|x_i^\star - x_j^\star\|_2 = 0$ for many pairs $\{i, j\} \in \mathcal{E}$, which implies samples satisfying the equation collapse into one point $\hat{x}$ such that $\hat{x} = x_i^\star = x_j^\star$. The ℓ_2 norm plays a similar role in this contraction to that in the group lasso (Yuan and Lin, 2006).

In general, we introduce a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$, where the node set $\mathcal{V} = [n]$ denotes the index set of samples, the edge set $\mathcal{E} \subset \{\{i, j\} : i, j \in [n];\, i \neq j\}$ indicates the pairwise adjacency or similarity, and $W = (w_{\{i,j\}})_{\{i,j\}\in\mathcal{E}} \in \mathbb{R}_{\geq 0}^{|\mathcal{E}|}$ denotes non-negative weights on all the edges to represent the pairwise similarity. (The higher $w_{\{i,j\}}$, the closer the vertices i and j.) Let $f_i : \mathbb{R}^p \to \mathbb{R}$ be a loss function for sample $i \in [n]$. NL (Hallac et al., 2015) is then formulated as the following optimization problem:

$$\underset{x_1,\ldots,x_n\in\mathbb{R}^p}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x_i) + \gamma \sum_{\{i,j\}\in\mathcal{E}} w_{\{i,j\}}\|x_i - x_j\|_2. \tag{2}$$

Obviously, (1) is an example of NL (2), where the sum of squared residuals, $f_i(x_i) = \frac{1}{2}(b_i - a_i^\top x_i)^2$, $i \in [n]$, are employed as the loss functions. NL includes other methods as special cases. If only the input vectors $a_i \in \mathbb{R}^p, i \in [n]$, are given and we employ

$$f_i(x_i) = \frac{1}{2}\|x_i - a_i\|_2^2, \tag{3}$$

and set $\mathcal{E} = \{\{i, j\} : i, j \in [n];\, i \neq j\}$ and $w_{\{i,j\}} = 1$, NL (2) is reduced to *convex clustering* (Pelckmans et al., 2005). Lindsten et al. (2011) and Hocking et al. (2011) considered extensions where $w_{\{i,j\}}$ were not necessarily equal to 1. With an optimal solution $(x_1^*, ..., x_n^*)$, nodes i and j are assigned to the same cluster if and only if $x_i^* = x_j^*$. We call x_i^* the *centroid* of node i. Namely, samples that share the same centroid form a cluster. The exact contraction property of the second term of (2) enables obtaining a clustering result of the data set $a_1, ..., a_n$ for sufficiently large γ. Recent studies show that NL numerically performs well in various tasks when information about node similarity (i.e., W) is appropriately given in advance (e.g., Hallac et al., 2015; Jung et al., 2018; Hocking et al., 2011; Chi and Lange, 2015; Sun et al., 2021).

In this paper, we further investigate and extend NL with a focus on its clustering capabilities. First, we study whether NL can recover true latent clusters. For convex clustering, Zhu et al. (2014), Panahi et al. (2017), and Sun et al. (2021) have established sufficient conditions for recovering the set of the latent clusters. Sun et al. (2021) have led to

the result including the results of Zhu et al. (2014) and Panahi et al. (2017) as special cases. For NL (not limited to convex clustering), Jung et al. (2018) and Jung and Tran (2019) have analyzed the gap between the optimal solution of NL and true parameter values. However, their conditions do not guarantee the recovery of the true clusters. In contrast, we provide sufficient conditions to recover the latent clusters for NL. The ease of the recovery will be given by ranges of the parameter γ, for which NL (2) recovers the (unseen) true clusters if they exist.

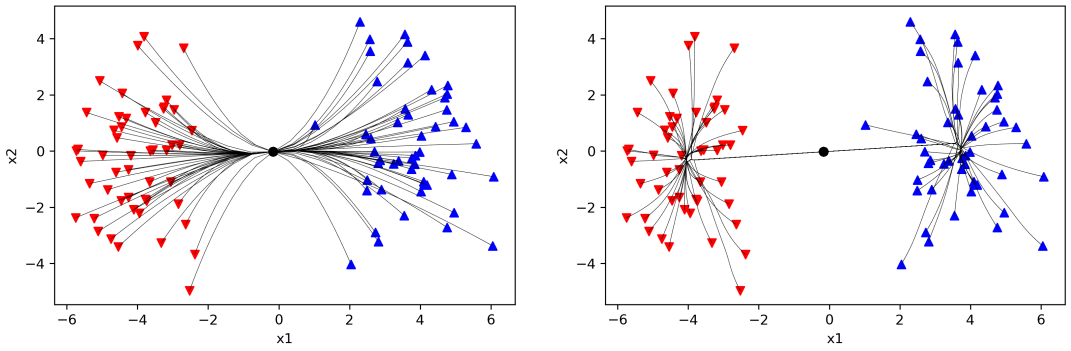
Besides, since (under a suitable assumption) the optimal solution to convex clustering is unique and independent of initial solutions whereas the k-means approach depends on initial solutions, it is almost certain that there is a gap between convex clustering and the usual k-means approach (or its non-convex optimization version). The second part of this paper aims to establish a bridge between the two realms: convex vs. non-convex. It is important to note that the performance of (2) highly depends on how the prior information is given by the weights W (and/or $\mathcal{E}$). The left panel of Figure 1 shows the regularization paths, i.e., the loci of the centroids obtained by convex clustering without prior information (i.e., $w_{\{i,j\}} = 1$ for all $\{i,j\} \in \mathcal{E}$ and $\mathcal{E} = \{\{i,j\} \mid i \neq j,\ i,j \in \mathcal{V}\}$). While there are two latent clusters (red and blue), all the centroids shrink to the middle point in an equal manner and we cannot obtain the two clusters even with a large γ. For a practical use of convex clustering, it is often suggested to set the weights $w_{\{i,j\}}$, as $w_{\{i,j\}} = \exp(-\alpha\|a_i - a_j\|_2^2)$, where $\alpha > 0$ is a parameter. The right panel of Figure 1 shows that convex clustering with this technique resulted in a clear-cut clustering. We should note that how to provide such prior information depends on the task at hand, and there are no general tips for NL (2) (e.g., for regression).

To address this issue, we consider introducing a cardinality constraint instead of the group l_2-penalty in NL (2). We show that the cardinality-constrained problem can be equivalently rewritten by a non-convex but continuous unconstrained optimization problem, which we call *network trimmed lasso* (NTL for short). This reformulation is parallel to that of the *trimmed lasso*, which is studied by, for example, Gotoh et al. (2018), Bertsimas et al. (2017), and Amir et al. (2021). We also propose algorithms based on the alternating direction method of multipliers (ADMM) to solve NTL. For a non-convex subproblem in the proposed algorithms, a closed-form solution is derived. Additionally, we show the convergence of proximal ADMM, an extension of ADMM, to a locally optimal solution of NTL, which is a non-convex optimization problem. Advantages of NTL over ordinary NL or clustered federated learning algorithms (Ghosh et al., 2020; Sattler et al., 2020) are demonstrated through numerical experiments.

Contributions of the paper are summarized as follows:

- We provide sufficient conditions under which NL can recover a latent cluster structure (Theorem 2). While a similar guarantee is known for convex clustering (Sun et al., 2021), our result is the first for a general framework of NL applicable beyond convex clustering (e.g., Equation 1).

- We propose NTL to improve the cluster structure detection ability of NL (Section 3). NTL is a continuous unconstrained optimization reformulation (12) where the objective function includes a nonconvex continuous penalty term called NTL penalty, and we show that any local optimum of NTL satisfies the cardinality constraint on



(a) Without prior information (b) With prior information

Figure 1: Failed cluster path of centroids via convex clustering without prior information (left panel) and successful cluster path with prior information, where $w_{\{i,j\}} = \mathrm{e}^{-0.1\|a_i - a_j\|_2^2}$ (right panel).

As the regularization parameter γ grows, all the centroids are reduced to the mean of n points, i.e., $\frac{1}{n}\sum_{i\in\mathcal{V}} a_i$, which is located near the origin in each picture. The left-hand side panel shows the result of convex clustering with $w_{\{i,j\}} = 1$, failing to form clusters even for large γ's. On the other hand, the right-hand side panel shows the case where the distance of points is used and succeeded in providing a clear-cut cluster structure even with small γ's.

the number of unmerged node pairs in a given network (Theorem 6 for general case; Corollary 8 for clustering; Corollary 10 for general convex quadratic case).

- To obtain a local optimum of NTL (12), we introduce ADMM-based algorithms to solve a problem (15) that involves a generalized version of NTL penalty. We show that, under mild conditions, any sequence of points generated by (proximal) ADMM converges to a local optimum (Proposition 13 and Theorem 14). Combining this with the propositions stated in the preceding bullet point ensures that NTL can form clear-cut clusters.

The rest of this paper is organized as follows. The next section is devoted to showing sufficient conditions that NL recovers a latent cluster structure. Section 3 presents a cardinality-constrained version of NL to get a more distinguishing cluster structure than NL does, and shows the equivalence between the cardinality-constrained problem and NTL. In Section 4, we develop algorithms for NTL and provide its convergence results. Section 5 reports numerical examples, demonstrating the effectiveness of NTL. The paper concludes with Section 6. All proofs of propositions are included in Appendix A.

1.1 Notation and Preliminaries

A continuous differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is said to be *L-smooth* (or have a *Lipschitz continuous gradient with modulus* L) if there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

for any $x, y \in \mathbb{R}^p$. If f is L-smooth, then the inequality

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|x - y\|_2^2 \tag{4}$$

holds for all $x, y \in \mathbb{R}^p$, which implies $\frac{L}{2}\|\cdot\|_2^2 - f$ is convex (see, e.g., Beck, 2017, Lemma 5.7). For a differentiable convex function $f : \mathbb{R}^p \to \mathbb{R}$, the following statements are equivalent (Beck, 2017, Theorem 5.8):

- f is L-smooth.
- The inequality (4) holds for all $x, y \in \mathbb{R}^p$.
- $\frac{L}{2}\|\cdot\|_2^2 - f$ is convex.

For a convex function $f : \mathbb{R}^p \to \mathbb{R}$, the *subdifferential* of f at $x \in \mathbb{R}^p$ is defined by

$$\partial f(x) := \{z \in \mathbb{R}^p \mid f(y) \geq f(x) + z^\top (y - x) \quad \forall y \in \mathbb{R}^p\}.$$

Note that $\partial f(x) \neq \emptyset$ (Beck, 2017, Theorem 3.14). We call $f : \mathbb{R}^p \to \mathbb{R}$ *strongly convex* with a positive constant α (or simply, α-*strongly convex*) if $f - \frac{\alpha}{2}\|\cdot\|_2^2$ is a convex function. If f is α-strongly convex, by using Theorems 3.63 and 5.24 of Beck (2017) and the Cauchy-Schwarz inequality, we obtain

$$\|z\|_2 \geq \alpha \|x - \overline{x}\|_2, \tag{5}$$

for all $x \in \mathbb{R}^p, z \in \partial f(x)$, where $\overline{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} f(x)$ and its existence and uniqueness are guaranteed by strong convexity of f (Beck, 2017, Theorem 5.25). It is also known (e.g., Beck, 2017, Theorem 5.25) that if f is α-strongly convex and $\overline{x} \in \operatorname*{argmin}_{x \in \mathbb{R}^p} f(x)$, the inequality

$$\frac{\alpha}{2}\|x - \overline{x}\|_2^2 \leq f(x) - f(\overline{x}), \tag{6}$$

holds for all $x \in \mathbb{R}^p$. The *directional derivative* of $f : \mathbb{R}^p \to \mathbb{R}$ at a point $x \in \mathbb{R}^p$ in the direction $v \in \mathbb{R}^p$ is defined by

$$\mathrm{d}f(x; v) := \lim_{\eta \searrow 0} \frac{f(x + \eta v) - f(x)}{\eta}.$$

A point $x^* \in \mathbb{R}^p$ is called a *(directional-)stationary point* of an optimization problem $\min_x f(x)$ if the directional derivative $\mathrm{d}f(x^*; v)$ exists and is non-negative for any $v \in \mathbb{R}^p$. The maximum and minimum eigenvalues of a symmetric matrix A are denoted by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively.

2. Recovery Conditions for Network Lasso

In this section, we show recovery conditions for NL (2) to identify a latent cluster structure on the basis of Sun et al. (2021), which develops sufficient conditions for the recovery of a latent cluster structure for convex clustering.

Let $C_1, ..., C_N$ denote (unseen) N clusters, which satisfy that $C_i \cap C_j = \emptyset$ if $i \neq j$, and $C_1 \cup \cdots \cup C_N = \mathcal{V}$. We assume that each sample $i \in \mathcal{V}$ belongs to one of $C_1, ..., C_N$.

To define the recovery of the cluster structure, we introduce a couple of notions, as below, following Sun et al. (2021).

Definition 1 *Let $\mathcal{P} := \{C_1, \ldots, C_N\}$ and $\overline{\mathcal{P}} := \{\overline{C}_1, \ldots, \overline{C}_M\}$ be partitionings of $\mathcal{V}$.*

1. *When $\overline{\mathcal{P}} = \mathcal{P}$, we say that $\overline{\mathcal{P}}$ perfectly recovers $\mathcal{P}$.*
2. *We call $\overline{\mathcal{P}}$ a coarsening of $\mathcal{P}$ if for any $\overline{C} \in \overline{\mathcal{P}}$ there exists $I \subset \{1, \ldots, N\}$ such that $\overline{C} = \cup_{l \in I} C_l$. Moreover, $\overline{\mathcal{P}}$ is called the trivial coarsening if $\overline{\mathcal{P}} = \{\mathcal{V}\}$. Otherwise, it is called a non-trivial coarsening.*

A partitioning represents a cluster structure of the data set $\mathcal{V}$. In the remainder of this paper, we use $\mathcal{P}$ to refer to the partitioning corresponding to the true (usually, unseen) cluster structure. For the partitioning $\mathcal{P} = \{C_1, \ldots, C_N\}$, we introduce the following notation:

$$
\begin{aligned}
n_k &:= |C_k|, && k \in [N], && \text{(size of Cluster } k\text{)} \\
w_i^{(k)} &:= \sum_{j \in C_k} w_{\{i,j\}}, && i \in \mathcal{V}, k \in [N], && \text{(sum of weights of Sample } i \text{ adjacent to Cluster } k\text{)} \\
w^{(k,k')} &:= \sum_{i \in C_k} \sum_{j \in C_{k'}} w_{\{i,j\}}, && k, k' \in [N]. && \text{(sum of weights between Clusters } k \text{ and } k'\text{)}
\end{aligned}
$$

For the sake of simplicity, we set $w_{\{i,j\}} = 0$ for $\{i,j\} \notin \mathcal{E}$ in this section. Note that $w_i^{(k)}$ can be viewed as the sample i's connectivity to the cluster C_k, and $w^{(k,k')}$ as the inter-cluster connectivity between two clusters C_k and $C_{k'}$.

Theorem 2 *Suppose that f_i is strictly convex and L_i-smooth, $i \in \mathcal{V}$. Let $\mathcal{P} = \{C_1, \ldots, C_N\}$ be the (unseen) true partitioning of $\mathcal{V}$, and let $f^{(k)}(x) := \sum_{i \in C_k} f_i(x)$, $k \in [N]$. Assume that for each $k \in [N]$, $f^{(k)}$ is α_k-strongly convex, and let $\overline{x}^{(k)} = \operatorname{argmin}_{x \in \mathbb{R}^p} f^{(k)}(x)$. Suppose that $\overline{x}^{(k)} \neq \overline{x}^{(k')}$ for $k \neq k'$. Let*

$$
\mu_{ij}^{(k)} := \sum_{l \neq k} \left| w_i^{(l)} - w_j^{(l)} \right| + \frac{L_i + L_j}{\alpha_k} \sum_{l \neq k} w^{(k,l)}, \qquad i, j \in C_k, k \in [N],
$$

and suppose that $n_k w_{\{i,j\}} > \mu_{ij}^{(k)}$ for all $i, j \in C_k, k \in [N]$ s.t. $i \neq j$. Let

$$
\begin{aligned}
\gamma_{\max} &:= \min_{k \neq k'} \left\{ \frac{\|\overline{x}^{(k)} - \overline{x}^{(k')}\|_2}{\frac{1}{\alpha_k} \sum_{l \neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}} \sum_{l \neq k'} w^{(k',l)}} \right\}, \\
\gamma_{\min} &:= \max_k \max_{\substack{i,j \in C_k \\ i \neq j}} \left\{ \frac{\left\| \nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)}) \right\|_2}{n_k w_{\{i,j\}} - \mu_{ij}^{(k)}} \right\},
\end{aligned}
$$

where we set $\frac{a}{0} = \infty$ for $a > 0$.

Let $(x_1^*, ..., x_n^*)$ be an optimal solution to (2), and $\overline{\mathcal{P}}$ be the quotient set of $\mathcal{V}$ by equivalence relation $x_i^* = x_j^*$.

1. *If $\gamma_{\min} \leq \gamma < \gamma_{\max}$, $\overline{\mathcal{P}}$ perfectly recovers $\mathcal{P}$.*

2. If $\gamma_{\min} \leq \gamma < \max_k \dfrac{\|\nabla f^{(k)}(\overline{x})\|_2}{\sum_{l \neq k} w^{(k,l)}}$, $\overline{\mathcal{P}}$ *is a non-trivial coarsening of* $\mathcal{P}$*, where*

$$\overline{x} := \operatorname*{argmin}_{x \in \mathbb{R}^p} \sum_{k \in [N]} f^{(k)}(x) = \operatorname*{argmin}_{x \in \mathbb{R}^p} \sum_{i \in \mathcal{V}} f_i(x).$$

This theorem implies that if a problem instance satisfies the condition $\gamma_{\min} < \gamma_{\max}$, NL recovers the true clusters at some point on the cluster path, $\gamma \in [\gamma_{\min}, \gamma_{\max})$, as demonstrated in Figure 1(b). Unfortunately, however, it is practically difficult to know whether an instance of NL meets the condition in advance. Therefore, let us derive implications of the condition, $\gamma_{\min} \leq \gamma < \gamma_{\max}$, to understand what kind of information contributes toward fulfilling the condition.

Note first that the condition is more likely to be met as $\gamma_{\min}$ is smaller and $\gamma_{\max}$ is larger. (Note also that we can construct an example where $\gamma_{\min} > \gamma_{\max}$ holds.) By the definition of $\gamma_{\min}$, a smaller $\mu_{ij}^{(k)}$ is preferable. The first term of $\mu_{ij}^{(k)}$ gauges a *within-cluster dissimilarity* since $\left|w_i^{(l)} - w_j^{(l)}\right|$ represents the difference of connectivity of the two samples, $i, j \in C_k$, to the other clusters $C_l, (l \neq k)$, whereas the second term gauges an *inter-cluster similarity* since $\sum_{l \neq k} w^{(k,l)}$ represents the connectivity between C_k and different clusters $C_l, (l \neq k)$, so the smaller the two terms, the more they contribute to the reduction of $\mu_{ij}^{(k)}$. Besides, the coefficient, $(L_i + L_j)/\alpha_k$, of the second term of $\mu_{ij}^{(k)}$ can be smaller when the number of samples in the same cluster is larger since, roughly speaking, the denominator α_k is the sum of the (lower bounds of) curvature of all f_i's within the cluster C_k while the numerator is the sum of (two upper bounds, L_i and L_j, of) curvature of f_i and f_j. Also, the denominator, $n_k w_{\{i,j\}} - \mu_{ij}^{(k)}$, in the definition of $\gamma_{\min}$ shows that a larger number n_k of samples in a cluster and larger weights $w_{\{i,j\}}$ between samples in the same cluster contribute to a decrease in $\gamma_{\min}$. The numerator, $\|\nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)})\|$, in the definition of $\gamma_{\min}$ reflects a within-cluster dissimilarity based on the gradient of f_i's at the centroid $\overline{x}^{(k)}$. On the other hand, the numerator, $\|\overline{x}^{(k)} - \overline{x}^{(k')}\|_2$, in the definition of $\gamma_{\max}$ denotes the inter-cluster dissimilarity of the centroids $\overline{x}^{(k)}, \overline{x}^{(k')}$, whereas the denominator, $\frac{1}{\alpha_k}\sum_{l \neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}}\sum_{l \neq k'} w^{(k',l)}$, represents the inter-cluster similarity based on the weights on the inter-cluster edges.

Overall, Theorem 2 suggests that the situation where the weights on the inter-cluster edges are smaller and those on the within-cluster edges are larger is preferable for NL. Although it is difficult to feed informative weights to NL in general, the popular choice $w_{\{i,j\}} = \exp(-\alpha\|a_i - a_j\|_2^2)$ for the convex clustering seems to be appropriate if the cluster structure is considered to be determined by the Euclidean distance of the data points $\{a_i\}_{i \in \mathcal{V}}$, as demonstrated in Figure 1. The importance of $w_{\{i,j\}}$ will be further discussed in the following remark.

Remark 3 *Theorem 2 implies that if the weights* $(w_{\{i,j\}})_{\{i,j\} \in \mathcal{E}}$ *are chosen adequately, NL is guaranteed to return the true cluster structure* $\{C_1, ..., C_N\}$ *at some point on the cluster path. To see this through an example, let us suppose that* $\mathcal{G}$ *is a complete graph* $\mathcal{E} = \{\{i, j\} \mid i \neq j,\ i, j \in \mathcal{V}\}$, $\{f_i\}_{i \in \mathcal{V}}$ *satisfies the assumption of Theorem 2, and the weights are defined*

as

$$w_{\{i,j\}} \begin{cases} \leq w, & (i \in C_k,\ j \in C_{k'},\ k \neq k'), \\ = 1, & (i, j \in C_k,\ i \neq j), \end{cases}$$

for a constant $w \in [0, 1]$*. Note that when* w *is equal to* 1*, we can say the weights have no information; as* w *gets closer to* 0*, the weights more reflect the true cluster structure. Observing that*

$$\sum_{l \neq k} w^{(k,l)} \leq \sum_{l \neq k} n_k n_l w \leq n^2 w \to 0 \quad (w \to 0)$$

for all $k \in [N]$ *and*

$$\sum_{l \neq k} \left| w_i^{(l)} - w_j^{(l)} \right| \leq \sum_{l \neq k} n_l w \leq nw \to 0 \quad (w \to 0)$$

for all $i, j \in C_k, k \in [N]$*, we have*

$$\gamma_{\max} \to \infty,$$
$$\gamma_{\min} \to \max_k \max_{\substack{i,j \in C_k \\ i \neq j}} \left\{ \frac{\left\| \nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)}) \right\|_2}{n_k} \right\},$$

as $w \to 0$*. This implies that for sufficiently small* w *the interval* $[\gamma_{\min}, \gamma_{\max})$ *becomes wider so that we can find a value of* γ *in the range in an easier manner. This example indicates that if* $(w_{\{i,j\}})_{\{i,j\} \in \mathcal{E}}$ *are given so that they reflect the true cluster structure sufficiently, NL returns the true clusters with some* $\gamma \in [\gamma_{\min}, \gamma_{\max})$*.*

Remark 4 *Since* $f_i(x_i) = \frac{1}{2}(b_i - a_i^\top x_i)^2$ *is not strictly convex for* $p \geq 2$*, Theorem 2 does not apply to optimization problem* (1)*. However, the condition is fulfilled if we modify* $f_i(x_i)$ *by adding, for example, an* ℓ_2*-regularizer* $\frac{\varepsilon}{2}\|x_i\|_2^2$ *for a small* $\varepsilon > 0$ *(that is,* $f_i(x_i) = \frac{1}{2}(b_i - a_i^\top x_i)^2 + \frac{\varepsilon}{2}\|x_i\|_2^2$*), as is often done to stabilize the estimation. In fact,* f_i *is clearly strictly convex and has a Lipschitz continuous gradient with modulus* $\|a_i\|_2^2 + \varepsilon$*, and* $f^{(k)}$ *is strongly convex with modulus* $\lambda_{\min}(\sum_{i \in C_k} a_i a_i^\top) + n_k \varepsilon$ *from* $\nabla^2 f^{(k)}(x) = \sum_{i \in C_k} a_i a_i^\top + n_k \varepsilon I_p$*, where* I_p *is the* p*-dimensional identity matrix.*

Remark 5 *While our result covers the case where* $f_i(x_i) = \frac{1}{2}\|x_i - a_i\|_2^2$ *for all* $i \in \mathcal{V}$*, i.e., convex clustering, Theorem 2 is slightly weaker than the result of Sun et al. (2021) for convex clustering because of the generalization beyond convex clustering. In their result, the thresholds corresponding to* $\gamma_{\max}$ *and* $\gamma_{\min}$*, between which recovery of true clusters is guaranteed, are given, respectively, by*

$$\gamma'_{\max} = \min_{k \neq k'} \left\{ \frac{\|a^{(k)} - a^{(k')}\|_2}{\frac{1}{n_k} \sum_{l \neq k} w^{(k,l)} + \frac{1}{n_{k'}} \sum_{l \neq k'} w^{(k',l)}} \right\},$$

$$\gamma'_{\min} = \max_k \max_{\substack{i,j \in C_k \\ i \neq j}} \left\{ \frac{\|a_i - a_j\|_2}{n_k w_{\{i,j\}} - \sum_{l \neq k} \left| w_i^{(l)} - w_j^{(l)} \right|} \right\},$$

where $a^{(k)} = \frac{1}{n_k}\sum_{i\in C_k} a_i$. *Applying Theorem 2 and from* $\sum_{l\neq k}\left|w_i^{(l)} - w_j^{(l)}\right| \leq \mu_{ij}^{(k)}$, *we have*

$$\gamma_{\max} = \min_{k\neq k'}\left\{\frac{\|a^{(k)} - a^{(k')}\|_2}{\frac{1}{n_k}\sum\limits_{l\neq k} w^{(k,l)} + \frac{1}{n_{k'}}\sum\limits_{l\neq k'} w^{(k',l)}}\right\} = \gamma'_{\max},$$

$$\gamma_{\min} = \max_{k}\max_{\substack{i,j\in C_k\\ i\neq j}}\left\{\frac{\|a_i - a_j\|_2}{n_k w_{\{i,j\}} - \mu_{ij}^{(k)}}\right\} \geq \gamma'_{\min}.$$

This shows that $[\gamma'_{\min}, \gamma'_{\max}) \supset [\gamma_{\min}, \gamma_{\max})$ *holds, namely, the result of Sun et al. (2021) admits a wider interval than Theorem 2 for convex clustering.*

3. Network Trimmed Lasso

In the previous section, we see that when the prior information $(w_{\{i,j\}})_{\{i,j\}\in\mathcal{E}}$ is given adequately, we can use NL for clustering. However, in the absence of the prior information, clustering by NL does not work well, as seen in Figure 1. Rather than not forming reasonable clusters, NL might not even form clusters, resulting in $\overline{\mathcal{P}} = \{\mathcal{V}\}$. Furthermore, it might not be easy to adequately define prior information for other tasks such as regression, as will be demonstrated in Section 5. In this section, we consider an extension of NL that forces data samples, $i \in \mathcal{V}$, to form clusters by incorporating a non-convex constraint.

3.1 Cardinality-Constrained Formulation and Its Equivalent Continuous Penalty Reformulation

In NL (2), the cluster structure is captured by the number of non-zero components of the vectors $(\|x_i - x_j\|_2)_{\{i,j\}\in\mathcal{E}}$. In light of this, we consider a minimization problem (7)–(8), where the fitting of the data set to models is optimized under a designated cardinality of non-zero components of the vector:

$$\underset{x_1,\ldots,x_n}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x_i) \tag{7}$$

$$\text{subject to} \quad \left|\left\{\{i,j\}\in\mathcal{E} : \|x_i - x_j\|_2 > 0\right\}\right| \leq K, \tag{8}$$

where K is a non-negative integer such that $K \leq |\mathcal{E}|$. As K decreases, the nodes agglomerate and form clusters. Hocking et al. (2011) treats convex clustering as a convex relaxation of problem (7)–(8).

While it is easier to interpret the hyperparameter K in (7)–(8) than γ in NL (2), the left-hand side of (8) is a discontinuous function in $(x_1, ..., x_n)$ and is known to be difficult to attain the global optimality of (7)–(8) in general. Therefore, we approach the problem by rewriting the cardinality constraint with an equivalent continuous counterpart.

Let $\xi := (\|x_i - x_j\|_2)_{\{i,j\}\in\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$, and denote the sum of the $|\mathcal{E}|-K$ smallest components of the vector ξ by

$$\tau_K(x_1, ..., x_n) := \xi_{(K+1)} + \cdots + \xi_{(|\mathcal{E}|)} \text{ with } \xi = (\|x_i - x_j\|_2)_{\{i,j\}\in\mathcal{E}}, \tag{9}$$

where $\xi_{(i)}$ denotes the i-th largest component of ξ. Note that $\tau_K(x_1, ..., x_n) \geq 0$ for any $(x_1, ..., x_n)$. It is easy to see that the problem (7)–(8) is equivalent to the following problem:

$$\underset{x_1,...,x_n}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x_i) \tag{10}$$

$$\text{subject to} \quad \tau_K(x_1, ..., x_n) = 0, \tag{11}$$

by noting that (8) and (11) are equivalent (see Gotoh et al., 2018, Subsection 5.2). Note that (10)–(11) is a continuous optimization problem if f_i are continuous, whereas (7)–(8) is not because the left-hand side of constraint (8) is a discontinuous function. Now consider the following penalty form of the constrained problem (10)–(11):

$$\underset{x_1,...,x_n}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x_i) + \gamma\tau_K(x_1, ..., x_n), \tag{12}$$

where $\gamma > 0$. The second term of the objective function of (12) plays a role of a penalty function of the cardinality constraint (8) in that (i) $\tau_K(x_1, ..., x_n) \geq 0$ for all $(x_1, ..., x_n)$, and (ii) $\tau_K(x_1, ..., x_n) > 0$ if and only if $\left|\left\{\{i, j\} \in \mathcal{E} : \xi_{\{i,j\}} > 0\right\}\right| > K$.

We call the problem (12) *network trimmed lasso* (NTL for short). If we set $K = 0$, $\tau_K(x_1, ..., x_n) = \sum_{\{i,j\}\in\mathcal{E}} \|x_i - x_j\|_2$, so the problem (12) is reduced to NL (2) with $w_{\{i,j\}} = 1$ for all $\{i, j\} \in \mathcal{E}$.

While (12) is now an unconstrained problem, another parameter γ is introduced instead. We will show below that if we take γ large enough, (12) is guaranteed to be equivalent to the constrained problem (10)–(11), and accordingly, to the cardinality-constrained problem (7)–(8).

Theorem 6 *1. Suppose that f_i is L_i-smooth for each $i \in \mathcal{V}$, and let $\boldsymbol{x}^\gamma := (x_1^\gamma, ..., x_n^\gamma)$ be an optimal solution of* (12)*. Suppose that there exists $C > 0$ such that $\|x_i^\gamma\|_2 \leq C$ for all $i \in \mathcal{V}$ and any $\gamma > 0$. Then $\boldsymbol{x}^\gamma$ is optimal to* (10)–(11) *if*

$$\gamma > \sum_{i\in\mathcal{V}} (\|\nabla f_i(0)\|_2 + 2L_i C). \tag{13}$$

2. In addition to the L_i-smoothness, suppose that f_i is convex for each $i \in \mathcal{V}$, and let $\boldsymbol{x}^\gamma := (x_1^\gamma, ..., x_n^\gamma)$ be a locally optimal solution of (12)*. Suppose that there exists $C > 0$ such that $\|x_i^\gamma\|_2 \leq C$ for all $i \in \mathcal{V}$ and any $\gamma > 0$. Then $\boldsymbol{x}^\gamma$ is locally optimal to* (10)–(11) *if the inequality* (13) *holds.*

By Statement 1. of Theorem 6, we are motivated to solve NTL (12) instead of the cardinality-constrained problem (7)–(8) since NTL (12) is an unconstrained minimization of a continuous function while (7)–(8) involves a constraint defined by a discontinuous function. Despite the continuity of the objective function, developing a global optimization algorithm for (12) is not easy especially when the number of variables is large. On the other hand, Statement 2. of Theorem 6 yields conditions under which a locally optimal solution to (7)–(8) is obtained by a locally optimal solution to NTL (12), which is attainable by, for example, proximal ADMM (Li and Pong, 2015) as shown in the next section. As a result,

we can obtain non-trivial clusters since the solution returned by the algorithm satisfies the cardinality constraint (8).

Both statements of Theorem 6 suppose that the size of the solution set is bounded by a constant C. In the following, we will see a few examples where values of C can be explicitly given.

Example 1 (Network trimmed lasso for ordinary clustering) *Consider the clustering problem of a data set $a_i \in \mathbb{R}^p, i \in \mathcal{V}$ with $f_i(x_i) = \frac{1}{2}\|x_i - a_i\|_2^2, i \in \mathcal{V}$. (Note that we do not limit to the case where $\mathcal{E} = \{\{i,j\} \mid i \neq j,\ i,j \in \mathcal{V}\}$.) NTL then becomes*

$$\underset{x_1,\ldots,x_n \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i - a_i\|_2^2 + \gamma \tau_K(x_1,\ldots,x_n). \tag{14}$$

For this clustering problem, we can find a threshold value of γ of Theorem 6 explicitly in a simple manner. To see this, first observe the following lemma, which shows the boundedness of locally optimal solutions to (14).

Lemma 7 *Let $C = \max_{i\in\mathcal{V}} \|a_i\|_2$. For any $\gamma > 0$, any locally optimal solution $\boldsymbol{x}^*$ of (14) satisfies $\|x_i^*\|_2 \leq C$ for all $i \in \mathcal{V}$.*

From Theorem 6 and Lemma 7, we obtain the following result, which dictates an explicit threshold value of the penalty parameter γ for ordinary clustering.

Corollary 8 *Let $C = \max_{i\in\mathcal{V}} \|a_i\|_2$. If $\gamma > 3nC$, then any optimal solution (resp. locally optimal solution) of (14) is also optimal (resp. locally optimal) to the cardinality-constrained clustering problem (i.e., Problem (7)–(8) with $f_i(x_i) = \frac{1}{2}\|x_i - a_i\|_2^2$).*

Besides the clustering problem (14), there are further examples where the threshold for γ can be derived explicitly. Consider a general case where f_i is α_i-strongly convex for all $i \in \mathcal{V}$. The following lemma claims that any locally optimal solution to NTL (12) is then bounded.

Lemma 9 *Assume that f_i is α_i-strongly convex for all $i \in \mathcal{V}$. Denote an unique optimizer of $\min f_i(x)$ by $\overline{x}_i$. Let $C = \left(\frac{2}{\alpha}\sum_{j\in\mathcal{V}} (f_j(0) - f_j(\overline{x}_j))\right)^{\frac{1}{2}} + \max_{i\in\mathcal{V}} \|\overline{x}_i\|_2$, where $\alpha = \min_{i\in\mathcal{V}} \alpha_i$. Then for any $\gamma > 0$, any locally optimal solution $\boldsymbol{x}^*$ of (12) satisfies $\|x_i^*\|_2 \leq C$ for all $i \in \mathcal{V}$.*

In the case where each f_i is a strictly convex quadratic function, by using Lemma 9 a threshold value of γ can be specified as follows.

Corollary 10 *Suppose that for a positive definite matrix $A_i \in \mathbb{R}^{p\times p}$ and a p-vector $B_i \in \mathbb{R}^p$, f_i is given by $f_i(x_i) = \frac{1}{2}x_i^\top A_i x_i - B_i^\top x_i$ for $i \in \mathcal{V}$. Let $C = \left(\frac{1}{\alpha}\sum_{i\in\mathcal{V}} B_i^\top A_i^{-1} B_i\right)^{\frac{1}{2}} + \max_{i\in\mathcal{V}} \|A_i^{-1}B_i\|_2$, where we set $\alpha = \min_{i\in\mathcal{V}} \lambda_{\min}(A_i)$. If $\gamma > \sum_{i\in\mathcal{V}}(\|B_i\|_2 + 2\lambda_{\max}(A_i)C)$, then any optimal solution (resp. locally optimal solution) to (12) is also optimal (resp. locally optimal) to (7)–(8).*

4. Algorithm

In this section, we develop a proximal ADMM that finds a locally optimal solution for (a generalized version of) NTL (12) to ensure that the cardinality constraint (8) is satisfied. Based on the ordinary ADMM (Section 4.1) and the efficient solution to a subproblem of ADMM (Section 4.2), we describe a proximal ADMM in Section 4.3 and show its convergence results in Section 4.4. We close the section by presenting the procedure of generating cluster paths with respect to the cardinality parameter K.

4.1 ADMM

As the first algorithm, we consider the alternating direction method of multipliers (ADMM) (e.g., Boyd et al., 2011). For NL (including convex clustering), Chi and Lange (2015) and Hallac et al. (2015) propose a method based on ADMM.

In this subsection, we deal with a more general problem, which includes NTL (12) as a special case. Similar to the trimmed lasso function (9), let us define the function T_K on $\mathbb{R}^{pm}$ by

$$T_K((z_k)_{k\in[m]}) = \|z_{(K+1)}\|_2 + \cdots + \|z_{(m)}\|_2,$$

where $K \in \{0, 1, \ldots, m\}$, $z_k \in \mathbb{R}^p$, and $\|z_{(k)}\|_2$ denotes the k-th largest component of $(\|z_1\|, ..., \|z_m\|) \in \mathbb{R}^m$. Note that T_K is a continuous function. With this function, our target optimization problem is formulated as

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}) + \gamma T_K(D\boldsymbol{x}), \tag{15}$$

where $\gamma > 0$, $f : \mathbb{R}^N \to \mathbb{R}$, and D is a $pm \times N$ matrix. Note that if we set $f(\boldsymbol{x}) = \sum_{i\in\mathcal{V}} f_i(x_i)$ and D is a matrix such that $\boldsymbol{z} = D\boldsymbol{x}$ with $z_{\{i,j\}} = x_i - x_j$ for all $\{i, j\} \in \mathcal{E}$, then the problem (15) is reduced to NTL (12).

To apply ADMM, we rewrite the problem (15) as the following equality-constrained formulation:

$$\underset{\boldsymbol{x},\boldsymbol{z}}{\text{minimize}} \quad f(\boldsymbol{x}) + \gamma T_K(\boldsymbol{z}) \tag{16}$$

$$\text{subject to} \quad \boldsymbol{z} = D\boldsymbol{x}. \tag{17}$$

By introducing the dual variables $\boldsymbol{y} \in \mathbb{R}^{pm}$ for the equality constraints (17), the augmented Lagrangian function of (16)–(17) is defined as

$$L_\rho(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) = f(\boldsymbol{x}) + \gamma T_K(\boldsymbol{z}) + \boldsymbol{y}^\top(\boldsymbol{z} - D\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{z} - D\boldsymbol{x}\|_2^2,$$

with a positive constant ρ. ADMM is then described as Algorithm 1.

4.2 Closed-Form Solution of Subproblem (18)

We can derive a closed-form solution of Subproblem (18). First, it is easy to see that (18) is reduced to

$$\boldsymbol{z}^{t+1} \in \text{prox}_{\frac{\gamma}{\rho}T_K}(D\boldsymbol{x}^t - \frac{1}{\rho}\boldsymbol{y}^t) = \underset{\boldsymbol{z}}{\text{argmin}}\Big\{\frac{\gamma}{\rho}T_K(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z} - (D\boldsymbol{x}^t - \frac{1}{\rho}\boldsymbol{y}^t)\|_2^2\Big\}, \tag{21}$$

Algorithm 1 ADMM for (15)

Input: $\boldsymbol{x}^0, \boldsymbol{y}^0, \rho > 0$ and $t = 0$.
repeat

$$\boldsymbol{z}^{t+1} \in \operatorname*{argmin}_{\boldsymbol{z}} L_\rho(\boldsymbol{x}^t, \boldsymbol{z}, \boldsymbol{y}^t), \tag{18}$$

$$\boldsymbol{x}^{t+1} \in \operatorname*{argmin}_{\boldsymbol{x}} L_\rho(\boldsymbol{x}, \boldsymbol{z}^{t+1}, \boldsymbol{y}^t), \tag{19}$$

$$\boldsymbol{y}^{t+1} = \boldsymbol{y}^t + \rho(\boldsymbol{z}^{t+1} - D\boldsymbol{x}^{t+1}). \tag{20}$$

$t = t + 1$
until Stopping criterion satisfied.
Output: $\boldsymbol{x}^t$.

where

$$\operatorname{prox}_g(\boldsymbol{x}) := \operatorname*{argmin}_{\boldsymbol{z}} \left\{ g(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 \right\}$$

is the proximal mapping of $\boldsymbol{x}$ with respect to a function g. Note that (21) may not be a singleton since T_K is non-convex.

Though the minimization in (21) is a non-convex optimization, we can derive a closed-form solution, $\boldsymbol{z}^{t+1}$, in a similar manner to Lu and Li (2018) and Bertsimas et al. (2017). For simplicity of notation, let $\boldsymbol{a} = D\boldsymbol{x}^t - \frac{1}{\rho}\boldsymbol{y}^t$. With this, the minimization in (21) can be equivalently rewritten as follows.

$$\begin{aligned}
\min_{\boldsymbol{z}} \quad \gamma T_K(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z} - \boldsymbol{a}\|_2^2 = \quad & \min_{\boldsymbol{z}} \ \gamma \sum_{k=K+1}^{m} \|z_{(k)}\|_2 + \frac{\rho}{2}\sum_{k=1}^{m}\|z_k - a_k\|_2^2 \\
= \quad & \min_{\boldsymbol{z}} \quad \left\{ \gamma \min_{\substack{I_k \in \{0,1\} \\ \sum_{k=1}^{m} I_k = m-K}} \left\{ \sum_{k=1}^{m} \|z_k\|_2 I_k \right\} + \frac{\rho}{2}\sum_{k=1}^{m}\|z_k - a_k\|_2^2 \right\} \\
= \quad & \min_{\substack{I_k \in \{0,1\} \\ \sum_{k=1}^{m} I_k = m-K}} \left\{ \min_{\boldsymbol{z}} \left\{ \gamma \sum_{k=1}^{m} \|z_k\|_2 I_k + \frac{\rho}{2}\sum_{k=1}^{m}\|z_k - a_k\|_2^2 \right\} \right\} \qquad (22) \\
= \quad & \min_{\substack{I_k \in \{0,1\} \\ \sum_{k=1}^{m} I_k = m-K}} \left\{ \sum_{k=1}^{m} \underbrace{\min_{z_k} \left\{ \gamma \|z_k\|_2 I_k + \frac{\rho}{2}\|z_k - a_k\|_2^2 \right\}}_{P_{(k)}} \right\},
\end{aligned}$$

where the second equality is obtained by introducing integer variables I_k, which play a role as an indicator of the smallest $m - K$ components, and the third and fourth equalities are established by interchanging "min" and "min," or "min" and "summation," which is possible because of the separability with respect to $\boldsymbol{z} = (z_k)_{k\in[m]}$. For fixed I_k, we next evaluate the term

$$P_{(k)} := \min_{z_k} \left\{ \gamma \|z_k\|_2 I_k + \frac{\rho}{2}\|z_k - a_k\|_2^2 \right\}.$$

To this end, let

$$P := \min_z \left\{ \pi(z) := \gamma \|z\|_2 \iota + \frac{\rho}{2} \|z - a\|_2^2 \right\}$$

for simplicity. Observe that when $\iota = 0$, we have $\operatorname*{argmin}_z \pi(z) = \{a\}$ and $P = 0$; when $\iota = 1$, we have

$$\operatorname*{argmin}_z \pi(z) = \operatorname{prox}_{\frac{\gamma}{\rho}\|\cdot\|_2}(a) = \begin{cases} 0, & \|a\|_2 \leq \frac{\gamma}{\rho}, \\ \left(1 - \frac{\gamma}{\rho\|a\|_2}\right) a, & \|a\|_2 > \frac{\gamma}{\rho}, \end{cases}$$

and $P = \phi(\|a\|_2)$, where

$$\phi(t) := \begin{cases} \frac{1}{2}t^2, & 0 \leq t \leq \frac{\gamma}{\rho}, \\ \frac{\gamma}{\rho}t - \frac{1}{2}\left(\frac{\gamma}{\rho}\right)^2, & t > \frac{\gamma}{\rho}. \end{cases}$$

Accordingly, with $a = a_k$, the problem (22) can be reduced to

$$\min_{\substack{I_k \in \{0,1\} \\ \sum_{k=1}^m I_k = m-K}} \sum_{k=1}^m P_{(k)} = \min_{\substack{I_k \in \{0,1\} \\ \sum_{k=1}^m I_k = m-K}} \sum_{k=1}^m I_k \phi(\|a_k\|_2).$$

Since $\phi(t)$ is increasing on $(0, \infty)$, an optimal solution of (21) is given by

$$z_k^{t+1} = \begin{cases} a_k, & \text{if } \|a_k\|_2 \text{ is in the largest } K \text{ components of } (\|a_k\|_2)_{k\in[m]}, \\ \operatorname{prox}_{\frac{\gamma}{\rho}\|\cdot\|_2}(a_k), & \text{if } \|a_k\|_2 \text{ is in the smallest } m-K \text{ components of } (\|a_k\|_2)_{k\in[m]}. \end{cases} \tag{23}$$

4.3 Proximal ADMM

As for subproblem (19), it is possible to derive a closed-form solution under restrictive assumptions (e.g., that of f being a strictly convex quadratic function). However, it can be hard to obtain a closed-form solution to (19) for some f.

To make the $\boldsymbol{x}$-update (19) at each iteration efficient, we consider *proximal ADMM* (Li and Pong, 2015). Suppose that f is L-smooth, so that the objective function of (19) is bounded above as

$$\begin{aligned} &L_\rho(\boldsymbol{x}, \boldsymbol{z}^{t+1}, \boldsymbol{y}^t) \\ &\leq f(\boldsymbol{x}^t) + \nabla f(\boldsymbol{x}^t)^\top (\boldsymbol{x} - \boldsymbol{x}^t) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}^t\|_2^2 + (\boldsymbol{y}^t)^\top (\boldsymbol{z}^{t+1} - D\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{z}^{t+1} - D\boldsymbol{x}\|_2^2 \end{aligned}$$

by the inequality (4). The minimizer of the right-hand side is given by

$$\boldsymbol{x}^{t+1} = \left(I_N + \frac{\rho}{L} D^\top D\right)^{-1} \left(\boldsymbol{x}^t - \frac{1}{L}\nabla f(\boldsymbol{x}^t) + \frac{1}{L} D^\top (\boldsymbol{y}^t + \rho \boldsymbol{z}^{t+1})\right), \tag{24}$$

where I_N is the N-dimensional identity matrix. Note that the formula (24) can be efficiently computed by a matrix-vector multiplication once the inverse on the right-hand side is fixed at the beginning of the algorithm.

Proximal ADMM can be defined with a more general update rule that would include (24) as a special case. For a continuously differentiable function ϕ on $\mathbb{R}^N$, we define the Bregman distance of $\boldsymbol{x}$ and $\boldsymbol{x}'$ by

$$B_\phi(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}') - \nabla\phi(\boldsymbol{x}')^\top(\boldsymbol{x} - \boldsymbol{x}').$$

In proximal ADMM, $\boldsymbol{x}^{t+1}$ is updated by

$$\boldsymbol{x}^{t+1} \in \operatorname*{argmin}_{\boldsymbol{x}} \left\{ L_\rho(\boldsymbol{x}, \boldsymbol{z}^{t+1}, \boldsymbol{y}^t) + B_\phi(\boldsymbol{x}, \boldsymbol{x}^t) \right\} \tag{25}$$

in place of (19). If we employ $\phi(\boldsymbol{x}) = \frac{L}{2}\|\boldsymbol{x}\|_2^2 - f(\boldsymbol{x})$, (25) can be reduced to (24). Algorithm 2 is the description of proximal ADMM, where the subroutine (25) is employed for x-update as well as the proximal mapping (23) of T_K for z-update.

Algorithm 2 Proximal ADMM for (15)

Input: $\boldsymbol{x}^0, \boldsymbol{y}^0, \rho > 0$, and $t = 0$.
repeat
  Let $\boldsymbol{a}^t := D\boldsymbol{x}^t - \frac{1}{\rho}\boldsymbol{y}^t$, then $\boldsymbol{z}^{t+1}$ is determined by (23) (i.e., Equation 18).
  $\boldsymbol{x}^{t+1}$ is determined by (25).
  $\boldsymbol{y}^{t+1}$ is determined by (20).
  $t = t + 1$
until Stopping criterion satisfied.
Output: $\boldsymbol{x}^t$.

Note that when we set $\phi(\boldsymbol{x}) = 0$, proximal ADMM is reduced to the ordinary ADMM (Algorithm 1).

4.4 Convergence of Proximal ADMM

The main goal of this subsection is to show that under practical assumptions proximal ADMM converges to a local minimum of (15) where the matrix D is assumed to be arbitrary $pm \times N$ real matrix. To show the convergence, we first give a formula of the directional derivative of T_K, which is a generalization of the result for the case where $p = 1$, given by Amir et al. (2021).

Lemma 11 *Let* $\Lambda_1 = \{k \mid \|z_k\|_2 < \|z_{(K)}\|_2\}$ *and* $\Lambda_2 = \{k \mid \|z_k\|_2 = \|z_{(K)}\|_2\}$. *The directional derivative of* T_K *at* $\boldsymbol{z} \in \mathbb{R}^{pm}$ *in the direction* $\boldsymbol{v} \in \mathbb{R}^{pm}$ *is given by*

$$\mathrm{d}T_K(\boldsymbol{z}; \boldsymbol{v}) = \sum_{k \in \Lambda_1} \delta(z_k, v_k)^\top v_k + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m - K - |\Lambda_1|}} \sum_{k \in \Lambda} \delta(z_k, v_k)^\top v_k,$$

where

$$\delta(z, v) := \begin{cases} \frac{z}{\|z\|_2}, & z \neq 0, \\ \frac{v}{\|v\|_2}, & z = 0, v \neq 0, \\ 0, & z = 0, v = 0 \end{cases}$$

and $\|z_{(0)}\|_2 = \infty$.

The following result claims that stationary points and local minima are equivalent in (15) when f is differentiable convex.

Proposition 12 *Suppose that f is a differentiable convex function. If $\boldsymbol{x}^*$ is a directional-stationary point of* (15)*, then it is locally optimal to* (15)*.*

The rest of this subsection is devoted to convergence results of proximal ADMM, for which proofs are based on ideas of Li and Pong (2015). The differences between their results and ours are summarized as follows:

- To apply Proposition 12, we will prove the convergence to a directional-stationary point. On the other hand, they prove convergence to a *limiting-stationary point*, which is a weaker stationary point than a directional-stationary point (see e.g., Cui et al., 2018, pp.3350–3351).
- They assume the second-order differentiability of f, while we only assume the first-order differentiability of f.

Let us start with results under the slightly stronger assumption that $(\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2, \|\boldsymbol{z}^{t+1}-\boldsymbol{z}^t\|_2, \|\boldsymbol{y}^{t+1}-\boldsymbol{y}^t\|_2)$ converges to $(0,0,0)$.

Proposition 13 *Suppose that f is convex, and f and ϕ are continuously differentiable. If the sequence $\{(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t)\}$ generated from proximal ADMM has a partial limit $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$ and $(\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2, \|\boldsymbol{z}^{t+1}-\boldsymbol{z}^t\|_2, \|\boldsymbol{y}^{t+1}-\boldsymbol{y}^t\|_2)$ converges to $(0,0,0)$, then $\boldsymbol{x}^*$ is a local minimum of* (15)*.*

Note that Proposition 13 ensures that if the whole sequence $\{(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t)\}$ converges to a point $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$, then $\boldsymbol{x}^*$ is a local minimum of (15).

By modifying the assumptions, we have a stronger convergence result than Proposition 13.

Theorem 14 *Suppose that the following assumptions hold:*

(A1) D is surjective, that is, $\sigma := \lambda_{\min}(DD^\top) > 0$;

(A2) f is convex;

(A3) $f+\phi$ is L_1-smooth;

(A4) $f+\phi+\frac{\rho}{2}\|D\cdot\|_2^2$ is α_1-strongly convex;

(A5) ϕ is L_2-smooth and α_2-strongly convex;

(A6) There exists $0<r<1$ such that $\rho > \frac{2}{\sigma(\alpha_1+\alpha_2)}\left(\frac{L_1^2}{r}+\frac{L_2^2}{1-r}\right)$,

where we allow $L_1, L_2, \alpha_1, \alpha_2$ to be 0, but we must have $\alpha_1+\alpha_2>0$. If the sequence $\{(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t)\}$ generated from proximal ADMM has a partial limit $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$, then $\boldsymbol{x}^*$ is a local minimum of (15)*.*

In the concrete examples of NTL considered in this paper, the assumption (A1) holds only for the piecewise constant fitting problem in Subsection 5.4. However, Proposition 13 is still valid for all the problems, although it is a weaker convergence result than Theorem 14. At first glance, the assumption (A4) may appear redundant because the assumptions (A2) and (A5) seem to imply (A4) with $\alpha_1 \geq \alpha_2$. However, this is not the case, as the assumption (A4) admits the condition $\alpha_1 > \alpha_2 = 0$, which will be satisfied in Examples 2 and 3 below.

We will note below how to choose ϕ and ρ based on Theorem 14.

Example 2 *Consider the case where D is surjective, f is L-smooth and α-strongly convex, and $\phi = 0$. The assumptions (A3)–(A5) are then fulfilled with $L_1 = L$, $L_2 = 0$, $\alpha_1 = \alpha$, and $\alpha_2 = 0$. By choosing $\rho > \frac{2L^2}{\sigma\alpha r}$ for some $0 < r < 1$, the assumption (A6) is also fulfilled.*

Example 3 *Consider the case where D is surjective and f is L-smooth and convex, and $\phi = \frac{L}{2}\|\cdot\|_2^2 - f$. In this case, ϕ is L-smooth since ϕ is differentiable convex and $\frac{L}{2}\|\cdot\|_2^2 - \phi = f$ is convex. Accordingly, the assumptions (A3)–(A5) are then fulfilled with $L_1 = L_2 = \alpha_1 = L$ and $\alpha_2 = 0$. By choosing $\rho > \frac{2L}{\sigma}\left(\frac{1}{r} + \frac{1}{1-r}\right)$ for some $0 < r < 1$, the assumption (A6) also holds.*

Example 4 *When D is not surjective, we cannot apply Theorem 14 because of the equation $\sigma = \lambda_{\min}(DD^\top) = 0$. In this case, we interpret $\rho > \frac{2}{\sigma(\alpha_1+\alpha_2)}\left(\frac{L_1^2}{r} + \frac{L_2^2}{1-r}\right) = \infty$ as a formality. In the computational examples of Section 5, we choose ϕ such that the assumptions (A3)–(A5) hold and take a large ρ to mitigate the inconsistency.*

While Theorem 14 assumes that proximal ADMM has a partial limit, the existence of a partial limit is guaranteed by the following theorem.

Theorem 15 *In addition to the assumptions (A1), (A3)–(A6), suppose that f is coercive, i.e., $\lim_{\|\boldsymbol{x}\|_2\to\infty} f(\boldsymbol{x}) = \infty$, and that there exists $0 < \zeta < \sigma\rho r$ such that*

$$f_{\inf} := \inf_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) - \frac{1}{2\zeta}\|\nabla f(\boldsymbol{x})\|_2^2 \right\} > -\infty.$$

Then the sequence $\{(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t)\}$ generated from proximal ADMM is bounded.

Example 5 *If f is L-smooth and bounded below, then the inequality*

$$\inf_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2 \right\} > -\infty$$

holds (see Li and Pong, 2015, Remark 3). Note that a continuous and coercive function is bounded below. If f is L-smooth and coercive, we choose ρ so that it satisfies not only the inequality in Example 2 or 3, but also the condition $\rho > \frac{L}{\sigma r}$.

Combining the above convergence results with Theorem 6 (or its corollaries) ensures that we can obtain a local minimum of (7)–(8) by applying proximal ADMM to (12) for a sufficiently large γ. Since the obtained solution satisfies the cardinality constraint (8), it is guaranteed that NTL can always provide a distinct cluster structure, even when no prior information is available, unlike NL.

4.5 Computation of Cluster Path

To get a cluster path of NTL (12) on the basis of proximal ADMM, we use a warm start. Let $\{K_t\}_{t=1}^T \subset \{0, 1, \ldots, |\mathcal{E}|\}$ be a decreasing sequence of the cardinality parameter K.

Algorithm 3 Cluster Path

Input: $\boldsymbol{x}^0, \boldsymbol{y}^0 = 0, \rho > 0, \{K_t\}_{t=1}^T$.
for $t = 1$ to T **do**
  Get $\boldsymbol{x}^t$ by using proximal ADMM to solve (12) with K_t, $\boldsymbol{x}^{t-1}$, and $\boldsymbol{y}^0$.
end for
Output: $\{\boldsymbol{x}^t\}_{t=0}^T$.

Note that when f_i is convex for all $i \in \mathcal{V}$, NL is a convex optimization problem and a global optimum is attained by any local search method, but NTL is a non-convex optimization, and the output of proximal ADMM is expected to be very sensitive to the initial point $(\boldsymbol{x}^0, \boldsymbol{y}^0)$ (and K). The choice of the initial point of cluster path is discussed through numerical experiments in Section 5.

5. Numerical Examples

This section presents several numerical examples to demonstrate how NTL behaves in comparison with NL (Subsections 5.1, 5.3, 5.4) or clustered federated learning algorithms (Subsection 5.2). We used ADMM (Algorithm 1) to solve NTL and Algorithm 3 to generate a cluster path. For ADMM to solve NTL, we used the following termination condition: $\|\boldsymbol{z}^{t+1} - D\boldsymbol{x}^{t+1}\|_2 \leq \sqrt{p|\mathcal{E}|}\varepsilon^{\mathrm{abs}} + \varepsilon^{\mathrm{rel}} \max\{\|\boldsymbol{z}^{t+1}\|_2, \|D\boldsymbol{x}^{t+1}\|_2\}$ and $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2 \leq \sqrt{pn}\varepsilon^{\mathrm{abs}} + \varepsilon^{\mathrm{rel}}\|\boldsymbol{x}^{t+1}\|_2$ are satisfied with $\varepsilon^{\mathrm{abs}} = \varepsilon^{\mathrm{rel}} = 10^{-5}$, or the number of iterations reaches 1000. For NL, we also used ADMM and increased γ when generating a cluster path. ADMM for NL was terminated either when $\|\boldsymbol{z}^{t+1} - D\boldsymbol{x}^{t+1}\|_2 \leq \sqrt{p|\mathcal{E}|}\varepsilon^{\mathrm{abs}} + \varepsilon^{\mathrm{rel}} \max\{\|\boldsymbol{z}^{t+1}\|_2, \|D\boldsymbol{x}^{t+1}\|_2\}$ and $\rho\|D(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)\|_2 \leq \sqrt{p|\mathcal{E}|}\varepsilon^{\mathrm{abs}} + \varepsilon^{\mathrm{rel}}\|\boldsymbol{y}^{t+1}\|_2$ were satisfied (as appeared in Boyd et al., 2011) for $\varepsilon^{\mathrm{abs}} = \varepsilon^{\mathrm{rel}} = 10^{-5}$, or when the number of iterations reached 1000.

5.1 Ridge Regression under Two Latent Clusters

We first address a regression problem. Unlike convex clustering, regression often lacks hints on how to provide prior information, so we consider the complete graph $\mathcal{E}$ and $w_{\{i,j\}} = 1$ for all $\{i, j\} \in \mathcal{E}$. We solved NL (2) and NTL (12), respectively, for simple regression models using two data sets, each consisting of $n = 100$ data points, $(a_1, b_1), ..., (a_{100}, b_{100})$, which are plotted in the top row of Figure 2, where the red and blue points correspond to two latent clusters. Obviously, the left-hand side panel is the case where the regression lines have different slopes and the data set has a clear cluster. In contrast, in the right-hand side panel, the two regression lines have similar slopes while keeping the linear separability of the two clouds. For each data point $(a_i, b_i) \in \mathbb{R}^2$, we consider the loss function of the form:

$$f_i(x_{i,1}, x_{i,2}) = \frac{1}{2}\|b_i - x_{i,1} - a_i x_{i,2}\|_2^2 + \frac{\varepsilon}{2}x_{i,2}^2, \quad i = 1, ..., 100,$$

where $x_{i,1}$ and $x_{i,2}$ are the intercept and the slope, respectively, of the model corresponding to data point i, and $\varepsilon > 0$ is a parameter to trade-off between the squared residual and

the ℓ_2-regularizer. In this experiment we set $\varepsilon = 10^{-2}$ and consider a complete graph, i.e., $\mathcal{E} = \{\{i,j\} \mid i \neq j,\ i,j \in \mathcal{V}\}$, and uniform weights $w_{\{i,j\}} = 1$ for all $i,j \in \mathcal{V}$. Initial points of cluster path are defined by $\boldsymbol{y}^0 = 0, x_i^0 = \underset{x \in \mathbb{R}^2}{\operatorname{argmin}} f_i(x)$.

For NL, let $(x_{i,1}(\gamma), x_{i,2}(\gamma))$ denote the centroid of data point i, obtained by ADMM under parameter γ. The second row of Figure 2 shows the cluster paths of centroids, $\{(x_{i,1}(\gamma), x_{i,2}(\gamma)) : \gamma = 10^{-3} \times (1.2)^{t-1}, t = 1, ..., 50\}$, generated by NL (via ADMM) with γ increasing. Starting with the initial points, $(x_{i,1}(0), x_{i,2}(0)) = (b_i, 0) = x_i^0$, which are highlighted in red or blue, they converge to a single black point in the middle as γ grows. We can see, however, from these two panels that NL failed to capture the cluster structure well for either data set in that the loci of centroids kept separated until only one cluster was formed at the center point with a sufficiently large γ.

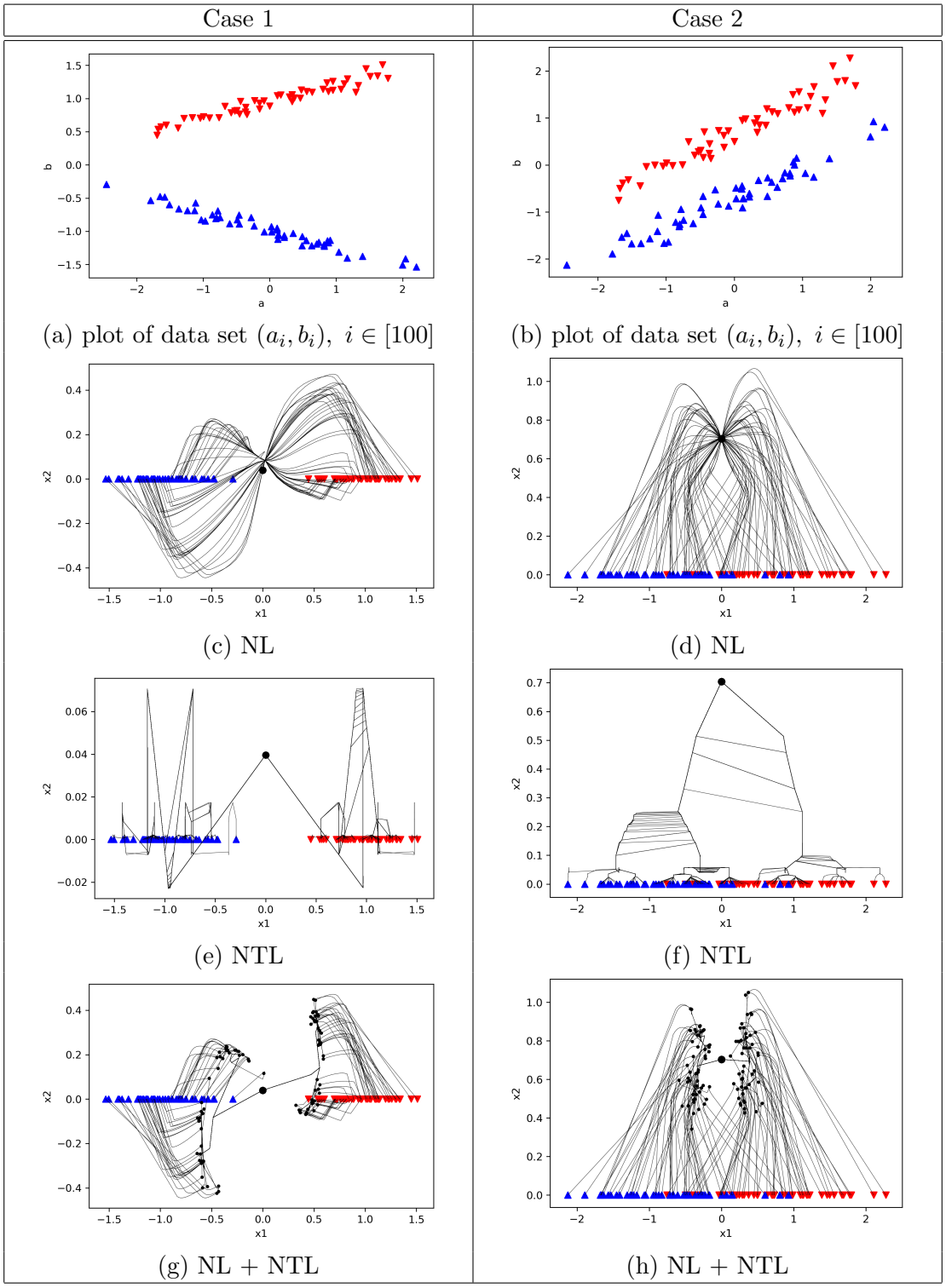
The third row of Figure 2 shows cluster paths generated by Algorithm 3. We employed the same initial points as in NL. From Example 4, we set $\rho = 10^4$ and set γ to be larger than the threshold presented in Corollary 10. We generated the cluster path with $K = 4500, 4450, ..., 50, 0$ in decreasing order. We can see from the third row of Figure 2 that NTL recovers true clusters for data set 1, but not fully for data set 2, in that we can see that some points joined in the opposite clusters for several small K's.

Finally, we consider using NL to improve NTL. The bottom row of Figure 2 shows cluster paths generated by NTL starting with the initial point generated by NL. The midpoint in the cluster path of NL, denoted by small black points in the bottom row of Figure 2, was employed as the initial point for NTL.[1] The choice of this initial point is motivated by the fact that samples belonging to the same true cluster are still likely to be closer to each other even if NL does not work well, as shown in the second row of Figure 2. In contrast with the case where NTL is only applied, we can see that it is better classified for both data sets. These results support the use of NTL when no prior information is available.

5.2 Comparison with Clustered Federated Learning Algorithms

In this section, we compare NTL with two clustered federated learning (CFL for short) algorithms proposed by Ghosh et al. (2020) and Sattler et al. (2020) in terms of clustering performance. CFL is a form of federated learning (FL for short) that aims to cluster distributed nodes called clients, each having data samples that are not shared with the other clients. The main purpose of FL is to better estimate a single model that is to be shared by clients without sharing data each client owns. Different from FL, CFL allows clusters of clients to have different models. Although their original motivations are different, NTL and CFL have the same task of finding clusters (of nodes and clients, respectively), and we here compare the quality of recovery of latent clusters obtained by NTL with that obtained by the two existing CFL algorithms. In order to adapt NTL to CFL, let each node $i \in [n]$ correspond to a client of CFL, and suppose ν_i data samples are assigned to estimate a model at each client (or node). For our numerical comparison, we consider $n = 100$ clients and four cases where each client equally has $\nu = 1, 10, 100$, or 1000 data samples, $\left((a_{hj}^{(i)})_{j \in [p]}, b_h^{(i)}\right), h \in [\nu]$. To test the recovering ability, we also assumed that each client was implicitly driven by $N = 4$ or $N = 8$ linear models, each corresponding to a latent cluster.

1. More precisely, the midpoint was defined among the points of centroids $\{(x_1(\gamma), ..., x_n(\gamma)) : \gamma = 10^{-3} \times (1.2)^{t-1}, t = 1, ..., 50\}$ where at least one centroid was different from one of the others.



(a) plot of data set $(a_i, b_i),\ i \in [100]$

(b) plot of data set $(a_i, b_i),\ i \in [100]$

(c) NL

(d) NL

(e) NTL

(f) NTL

(g) NL + NTL

(h) NL + NTL

Figure 2: Two data sets (a),(b) for simple regression and cluster paths of centroids (c)–(h)

Each linear model $k \in [N]$ is defined by $p = 10$ coefficients $(x_j^{(k)})_{j\in[10]}$, and each $x_i^{(k)}$ was drawn from the Bernoulli distribution with probability 0.5. The index of the latent cluster of client i, denoted by $k_i \in [N]$, was uniformly randomly determined. Each element $a_{hj}^{(i)}$ was independently drawn from the uniform distribution on $(0, 1)$ and $b_h^{(i)}$ was determined by $b_h^{(i)} = \sum_{j\in[10]} a_{hj}^{(i)} x_j^{(k_i)} + e_h^{(i)}$ where $e_h^{(i)}$ was independently drawn from $\mathrm{N}(0, 1)$. As for the loss function, we used the squared loss for the two CFL algorithms and the ℓ_2-regularized squared loss

$$f_i(x_i) = \frac{1}{2}\|b_i - A_i x_i\|_2^2 + \frac{\varepsilon}{2}\|x_i\|_2^2$$

for NTL to make f strictly convex, where $A_i := (a_{hj}^{(i)})_{h,j} \in \mathbb{R}^{\nu\times 10}$ and $\varepsilon = 10^{-2}$. The complete graph was applied to NTL for the same reason as the previous subsection and γ was set to be larger than the threshold presented in Corollary 10.

The quality of clusters obtained by the three algorithms is evaluated based on adjusted Rand index (Hubert and Arabie, 1985) (ARI for short; see, e.g., Vinh et al., 2010, Section 2 for the details). ARI takes a value between 0 and 1, and when it is closer to 1, the clustering performance is considered to be higher.

As for NTL, a cluster path was generated by Algorithm 3 with $\rho = \nu \times 10^4$, $K = 4900, 4880, \ldots, 20, 0$, where, motivated by the results of the previous subsection, the initial solution $\boldsymbol{x}^0$ was computed by NL. The performance of NTL was measured by the best ARI value along the cluster path. As for the algorithm of Ghosh et al. (2020), the number of clusters must be given in advance, and we considered three values, $N-1, N$, and $N+1$ for the number of clusters. Note that the true number of latent clusters is N, so this setting seems to be favorable for this algorithm. As for the algorithm of Sattler et al. (2020), it requires a parameter ($\gamma_{\max}$ in their notation) that would affect the number of clusters of outputs. We set it as $0.1, 0.2, \ldots,$ or 0.9, and show the best results in terms of ARI among the nine cases.

Tables 1 and 2 show the mean values and standard deviations of the maximum values of ARI for each method when repeated 50 times for each pair of N and ν with the above settings. We see from Tables 1 and 2 that as ν increases, NTL outperforms the other methods in both cases. In particular, a significant margin is found for the case where more latent clusters exist. These numerical results indicate that NTL is competitive with the novel clustered federated learning algorithms.

5.3 Ordinary Clustering Problem

This subsection compares (14) in Example 1 with convex clustering (CC for short). Namely, we set $f_i(x_i) = \frac{1}{2}\|x_i - a_i\|_2^2$ and $\mathcal{E} = \{\{i, j\} \mid i \neq j,\ i, j \in \mathcal{V}\}$ in NL (2) and NTL (12). As mentioned in Section 1, in this case, we have access to prior information.

Firstly, we consider the half moons data set ($n = 200,\ p = 2$). In Figure 3, the given (true) cluster labels of the data points are indicated by different colors (red versus blue). As for the weights for CC, we consider two cases: (i) uniform weights, $w_{\{i,j\}} = 1$ for all $\{i, j\} \in \mathcal{E}$, and (ii) non-uniform weights. In order to define non-uniform weights for case (ii), let us denote the k-nearest neighbors of a point $i \in \mathcal{V}$ by

$$\mathrm{NN}(i, k) := \{j \in \mathcal{V} \mid a_j \text{ is one of } k \text{ nearest neighbors of } a_i\}.$$

	$\nu = 1$		$\nu = 10$		$\nu = 100$		$\nu = 1000$	
method	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
CFLA(G)	**0.0827**	(0.0561)	**0.4682**	(0.1687)	0.8320	(0.1353)	0.7734	(0.1626)
CFLA(S)	0.0494	(0.0414)	0.0821	(**0.1372**)	0.3913	(0.3252)	0.9488	(0.0966)
NTL	0.0573	(**0.0365**)	0.2392	(0.1549)	**0.9289**	(**0.0808**)	**0.9773**	(**0.0467**)

Table 1: Mean and standard deviation of maximum adjusted Rand index ($N = 4$).

	$\nu = 1$		$\nu = 10$		$\nu = 100$		$\nu = 1000$	
method	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
CFLA(G)	**0.0401**	(0.0276)	**0.2816**	(0.0711)	0.5568	(0.1286)	0.4784	(0.0932)
CFLA(S)	0.0330	(0.0232)	0.0430	(**0.0591**)	0.2658	(0.1435)	0.8676	(0.1016)
NTL	0.0298	(**0.0169**)	0.1330	(0.0816)	**0.8189**	(**0.1167**)	**0.9698**	(**0.0165**)

Table 2: Mean and standard deviation of maximum adjusted Rand index ($N = 8$).

The algorithms proposed by Ghosh et al. (2020) and Sattler et al. (2020) are denoted by CFLA(G) and CFLA(S), respectively. The best value for each setting is shown in boldface.

With this, we define

$$w_{\{i,j\}} = \begin{cases} \exp(-0.5\|a_i - a_j\|_2^2), & \text{if } i \in \mathrm{NN}(j, 20) \text{ or } j \in \mathrm{NN}(i, 20), \\ 0, & \text{otherwise,} \end{cases}$$

for $\{i, j\} \in \mathcal{E}$.

For CC, a cluster path of centroids for $\gamma \in \{10^{-3} \times 2^{t-1}\}_{t=1}^{50}$ is computed with initial points $\boldsymbol{y}^0 = 0, x_i^0 = a_i$ $(i \in \mathcal{V})$.[2] As for NTL, we used $\rho = 10^4$ and started from the same initial points, computing a cluster path of centroids for $K \in \{19900, 19800, ..., 100, 0\}$. From Corollary 8, the penalty parameter γ is set to be $\gamma = 3n \max_i \|a_i\|_2 \times 1.001$. For k-means, the number of clusters is set to 2.

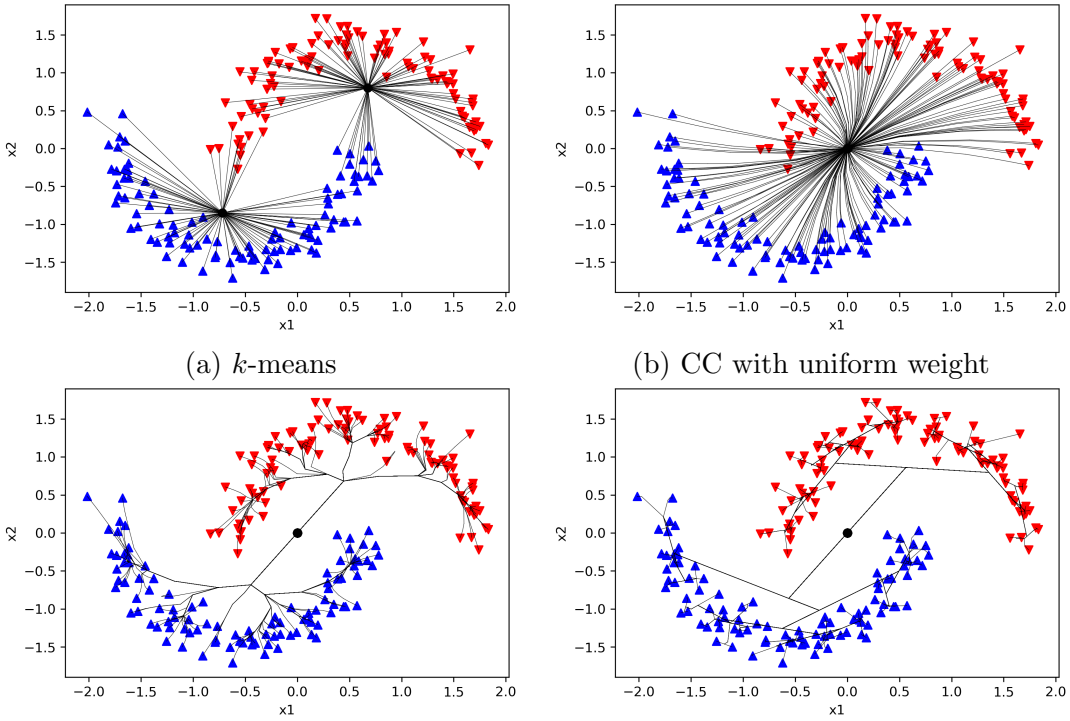
We can see from Figure 3 that the CC with uniform weight failed to form clusters until it degenerated to a single point. Although the k-means formed two clusters, it failed to recover two halfmoons. On the other hand, the weighted CC and NTL succeeded in recovering them along the cluster paths. Comparing with the two methods, NTL generates small clusters at the beginning of the cluster path, which is more informative than CC about the closeness of data points.

Next, using several real data sets,[3] we quantitatively compared the quality of clustering on the basis of ARI.

For weights for CC, we consider the following two cases:

$$w^1_{\{i,j\}} = \begin{cases} \exp(-0.5\|a_i - a_j\|_2^2), & \text{if } i \in \mathrm{NN}(j, \lceil \frac{n}{2} \rceil) \text{ or } j \in \mathrm{NN}(i, \lceil \frac{n}{2} \rceil), \\ 0, & \text{otherwise} \end{cases}$$

2. When all centroids degenerate at a single point, the computation of the path was stopped.
3. Data sets from scikit-learn https://scikit-learn.org/stable/datasets/index.html. The digit data set was resampled so that $n = 500, 100, 50$.



(a) k-means (b) CC with uniform weight

(c) CC with Gaussian weight (d) NTL

Figure 3: Data sets and cluster path of centroids for half moons.

method	iris	wine	digits ($n = 500$)	digits ($n = 100$)	digits ($n = 50$)
CC (uniform)	0.0015	0.0000	0.0082	0.0014	0.0013
CC (w^1)	0.5681	0.7577	0.5302	0.4577	0.3812
CC (w^2)	0.5681	0.7994	**0.5346**	**0.5101**	**0.4443**
NTL	**0.5778**	**0.8260**	0.3967	0.4134	0.4207

Table 3: Maximum adjusted Rand index through the cluster path.

The best value for each data set is shown in boldface. The range of the cardinality parameter K for NTL is set to $\{11000, 10900, ..., 100, 0\}$, $\{15500, 15400, ..., 100, 0\}$, $\{124000, 123900, ..., 100, 0\}$, $\{4900, 4880, ..., 20, 0\}$, and $\{1220, 1210, ..., 10, 0\}$ for iris, wine, digits ($n = 500$), digits ($n = 100$), and digits ($n = 50$), respectively.

and

$$w^2_{\{i,j\}} = \begin{cases} \exp(-0.5\|a_i - a_j\|_2^2), & \text{if } i \in \mathrm{NN}(j, \lceil \frac{n}{10} \rceil) \text{ or } j \in \mathrm{NN}(i, \lceil \frac{n}{10} \rceil), \\ 0, & \text{otherwise,} \end{cases}$$

where $\lceil l \rceil$ denotes the smallest integer no less than l. Note that $(w^2_{\{i,j\}})_{\{i,j\}\in\mathcal{E}}$ put more zeros on edges than $(w^1_{\{i,j\}})_{\{i,j\}\in\mathcal{E}}$. The range of the NTL cardinality parameter K is set as shown under Table 3. The other settings are the same as in the previous (half-moon) example.

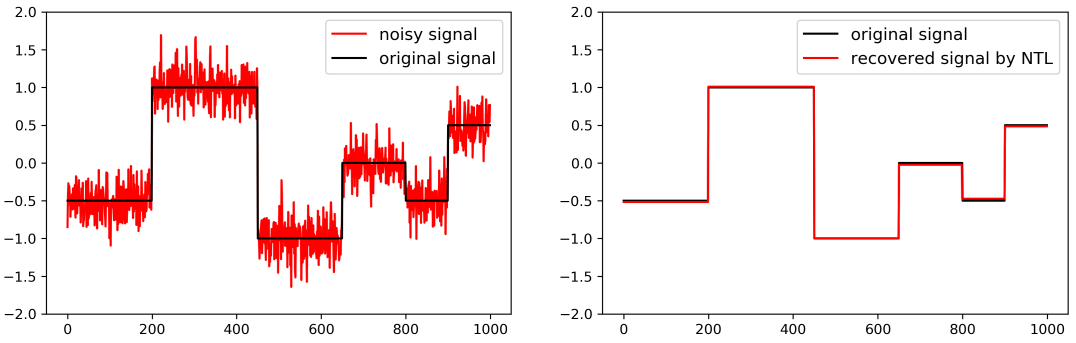
Table 3 summarizes the largest values of ARI along the cluster paths. We see from the table that the weighted CC with w^2 performed best for three data sets, as Theorem 2 implies. On the other hand, NTL recorded the best performance with the two data sets. We cannot say which one is better, but from this experiment, CC performs poorly in the absence of prior knowledge. In contrast, it is worth noting that NTL performed as well as weighted CC even without prior information.

5.4 Piecewise Constant Fitting

As the final example, we consider the problem of recovering a piecewise constant signal from a noisy signal (Calafiore and El Ghaoui, 2014, Example 9.16) by using NL and NTL. Specifically, we consider a situation where $n = 1000$ noisy signals $\hat{x}_1, ..., \hat{x}_{1000} \in \mathbb{R}$ are generated as $\hat{x}_i = x^{\mathrm{o}}_i + e_i$ with x^{o} being given as the original signal given as the black stepwise function in Figure 4 and e_i being independently drawn from a normal distribution $\mathrm{N}(0, 0.2^2)$. Given the time series structure, we set $\mathcal{V} = \{1, ..., 1000\}$, $\mathcal{E} = \{\{i, i+1\} \mid i, i+1 \in \mathcal{V}\}$, $f_i(x_i) = \frac{1}{2}(x_i - \hat{x}_i)^2$, and

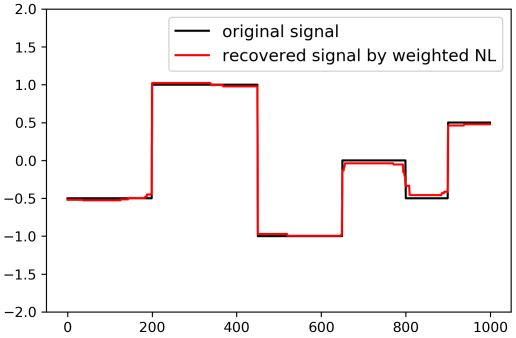
$$D := \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

It is known that $\sigma := \lambda_{\min}(DD^\top) = 2(1 - \cos(\frac{\pi}{1000})) \approx 9.87 \times 10^{-6}$ (Kulkarni et al., 1999, Theorem 2.2). In this example, we consider not only the perspective of the cluster recovery

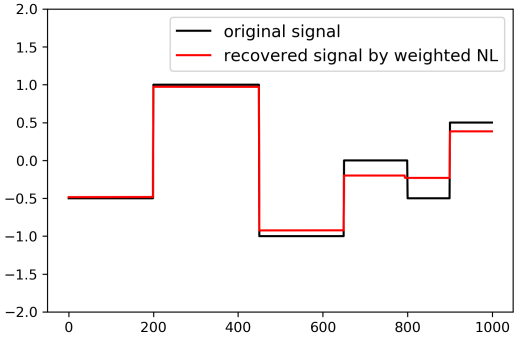


(a) plot of data set

(b) NTL



(c) NL (best in terms of quality)



(d) NL (best in terms of cluster recovery)

Figure 4: Noisy signal and recovered signals.

but also the quality of the solution. The quality of the solution here means the closeness of the recovered signal and the original signal, measured by $\|\boldsymbol{x}^* - \boldsymbol{x}^{\mathrm{o}}\|_2$.

For NL, we computed the cluster path of centroids for $\gamma \in \{10^{-3} \times (1.2)^{t-1}\}_{t=1}^{100}$ with the initial points $\boldsymbol{y}^0 = 0, x_i^0 = \hat{x}_i$ and the prior information $w_{\{i,i+1\}} = \exp(-0.5\|\hat{x}_i - \hat{x}_{i+1}\|_2^2)$. As for NTL, we applied ADMM from the same initial point, using $K = 5$ and $\gamma = 3n \max_i \|\hat{x}_i\|_2 \times 1.001$. As for the parameter ρ for ADMM, we here employ a heuristics similar to Li and Pong (2015) so as not to excessively restrict the movement of $\boldsymbol{x}^t$ and $\boldsymbol{z}^t$ at early iterations of ADMM. Specifically, starting with the initial value $\rho \leftarrow 1$, it was updated by the formula $\rho \leftarrow \min\{10\rho, \frac{2}{0.99\sigma}\}$ every 100 iterations. Noting that Examples 2 and 5 suggest $\rho > \max\{\frac{2}{\sigma}, \frac{1}{\sigma}\} = \frac{2}{\sigma} > 2 \times 10^5$ to fulfill the assumption of Theorem 14, the employed heuristics aims to increase the value gradually until the assumption is fulfilled.

Figure 4 shows how well NL and NTL recover the original signal, which is denoted by the black solid line, from the noisy signal, which is shown by the red solid line in the upper left panel. Since there is a degree of freedom in the evaluation criteria, two best-case results are given for NL. The panel (c) is the best in solution quality in the sense that the smallest value of $\|\boldsymbol{x}^* - \boldsymbol{x}^{\mathrm{o}}\|_2$ was attained out of 100 values of γ. On the other hand, the panel (d) is the best in the cardinality in the sense that the employed γ is the smallest out of the 100 values such that the number of jumped points is less than 5, which is the number of jumps in the original signal. We see from the panel (d) of Figure 4 that NL detected the jump points almost exactly as Theorem 2 implies, but the levels of the piecewise constants

method	iris	wine	digits ($n = 500$)	digits ($n = 100$)	digits ($n = 50$)
CC (uniform)	7.9790	15.2130	216.0687	5.4037	0.6222
CC (w^1)	5.8021	16.5441	735.3687	9.9490	1.2041
CC (w^2)	1.3671	1.4319	59.0054	0.7677	0.1367
NTL	3.8742	10.0048	81.8792	4.0477	1.1689

Table 4: Average CPU time (sec.) of one ADMM run for the experiment in Subsection 5.3.

The algorithms were implemented in MATLAB2021a and all the computations were conducted on a PC with OS: Windows 10 Pro for Workstations, CPU: Intel Xeon W-10885M 2.80 GHz, and 16.0 GB memory. The table shows that the computation time of the ADMM run increased at a faster rate than linearly in n for all problems.

are far from the original signal. We think this is due to the fact that the degree of each node is at most 2, so prior information was not given enough to recover the signal by NL. Employing more zeros as the edge weights worked better in the experiment of the previous subsection, but this example indicates that that is not always true. This indicates that it is not easy to give weights adequately for NL in advance. On the other hand, NTL not only detects the jumps/drops accurately but also estimates the levels of the piecewise constants more accurately than the best case of NL (lower left panel).

6. Concluding Remarks

This paper investigates the cluster structure of network lasso (NL) from multiple perspectives. Firstly, we establish a condition under which NL can recover (unseen) true clusters. Secondly, to obtain clusters that might not be attained by NL, we consider a cardinality constraint on the number of unmerged pairs of centroids and present an equivalent unconstrained reformulation called network trimmed lasso (NTL). We additionally show the convergence of ADMM to a locally optimal solution of NTL and the cardinality-constrained problem. Consequently, NTL can form distinct clusters even in the absence of prior information. Numerical examples illustrate how NTL outperforms the ordinary NL, particularly when no prior information is available. Our findings indicate that NL should be employed when provided with ample prior information and NTL otherwise. While the convergence of the algorithm is guaranteed, using the ADMM approach in situations where the underlying graph is both dense and large would result in impractical solution times, as can also be seen from Table 4. For example, when the graph is a complete graph, i.e., $\mathcal{E} = \{\{i, j\} \mid i \neq j,\ i, j \in \mathcal{V}\}$, ADMM would have to handle $\frac{n(n-1)}{2}p$-dimension vectors, which would be prohibitively large, even for a moderately sized n (e.g. $n = 1000$). The development of an efficient algorithm to handle such big data sets has been left for future research.

Acknowledgments

Jun-ya Gotoh was supported in part by JSPS KAKENHI Grant 19H02379, 19H00808, and 20H00285. The authors would like to thank the action editor and three anonymous reviewers for their comments and suggestions that helped improve the quality of the manuscript.

Appendix A. Proofs

In this section, we prove the propositions in this paper.

A.1 Proof of Theorem 2

Proof *Proof of Statement 1.* Let $(x^{(1)*}, \ldots, x^{(N)*})$ be an optimal solution of the following problem,

$$\underset{x^{(1)},\ldots,x^{(N)}}{\text{minimize}} \quad \sum_{k=1}^{N} f^{(k)}(x^{(k)}) + \gamma \sum_{k<l} w^{(k,l)} \|x^{(k)} - x^{(l)}\|_2, \tag{26}$$

which is equivalent to NL (3) with the symbols introduced in the statement of the theorem.

We first show that $\gamma < \gamma_{\max}$ implies $x^{(k)*} \neq x^{(k')*}$ for all $k \neq k'$. The optimality condition of (26) is then given by

$$\nabla f^{(k)}(x^{(k)*}) + \gamma \sum_{l \neq k} w^{(k,l)} z^{(k,l)} = 0, \qquad k \in [N], \tag{27}$$

where $z^{(k,k')} \in \partial \|x^{(k)*} - x^{(k')*}\|_2$ and $z^{(k,k')} = -z^{(k',k)}$, for any $k, k' \in [N]$ such that $k \neq k'$. Here, $\partial \|x^{(k)*} - x^{(k')*}\|_2$ denotes the subdifferential of $\|\cdot\|_2$ at $x^{(k)*} - x^{(k')*}$, and the subdifferential of $\|\cdot\|_2$ at x is given by

$$\partial \|x\|_2 = \begin{cases} \{\frac{x}{\|x\|_2}\}, & x \neq 0, \\ \{z \in \mathbb{R}^p \mid \|z\|_2 \leq 1\}, & x = 0. \end{cases}$$

By noting that $\|z^{(k,k')}\|_2 \leq 1$, and combining it with the triangle inequality and the equation (27), we obtain

$$\begin{aligned} \|\nabla f^{(k)}(x^{(k)*})\|_2 &\leq \gamma \sum_{l \neq k} w^{(k,l)} \|z^{(k,l)}\|_2 \\ &\leq \gamma \sum_{l \neq k} w^{(k,l)} \end{aligned} \tag{28}$$

for all $k \in [N]$. Since $f^{(k)}$ is α_k-strongly convex, we have for arbitrary $k, k' \in [N]$ such that $k \neq k'$,

$$\begin{aligned} \|\overline{x}^{(k)} - \overline{x}^{(k')}\|_2 &\leq \|\overline{x}^{(k)} - x^{(k)*}\|_2 + \|x^{(k)*} - x^{(k')*}\|_2 + \|x^{(k')*} - \overline{x}^{(k')}\|_2 \\ &\leq \|x^{(k)*} - x^{(k')*}\|_2 + \frac{1}{\alpha_k}\|\nabla f^{(k)}(x^{(k)*})\|_2 + \frac{1}{\alpha_{k'}}\|\nabla f^{(k')}(x^{(k')*})\|_2 \\ &\leq \|x^{(k)*} - x^{(k')*}\|_2 + \gamma\Big(\frac{1}{\alpha_k}\sum_{l \neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}}\sum_{l \neq k'} w^{(k',l)}\Big), \end{aligned} \tag{29}$$

where the first inequality is due to the triangle inequality, the second one follows from (5), and the final one from (28). If the term to the right of γ on the right-hand side of (29) is equal to zero, then $\|x^{(k)*} - x^{(k')*}\|_2 > 0$ holds by the assumption that $\overline{x}^{(k)} \neq \overline{x}^{(k')}$ for $k \neq k'$. Otherwise, combining the inequality (29) and the definition of $\gamma_{\max}$, we obtain

$$\begin{aligned}
&\|x^{(k)*} - x^{(k')*}\|_2 \\
&\geq \|\overline{x}^{(k)} - \overline{x}^{(k')}\|_2 - \gamma\Big(\frac{1}{\alpha_k}\sum_{l\neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}}\sum_{l\neq k'} w^{(k',l)}\Big) \\
&= \left(\frac{\|\overline{x}^{(k)} - \overline{x}^{(k')}\|_2}{\frac{1}{\alpha_k}\sum_{l\neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}}\sum_{l\neq k'} w^{(k',l)}} - \gamma\right)\Big(\frac{1}{\alpha_k}\sum_{l\neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}}\sum_{l\neq k'} w^{(k',l)}\Big) \\
&\geq (\gamma_{\max} - \gamma)\Big(\frac{1}{\alpha_k}\sum_{l\neq k} w^{(k,l)} + \frac{1}{\alpha_{k'}}\sum_{l\neq k'} w^{(k',l)}\Big) \\
&> 0.
\end{aligned}$$

Thus, $x^{(k)*} \neq x^{(k')*}$ for all $k \neq k'$.

Next, we show that $\gamma_{\min} \leq \gamma$ implies $x_i^* = x^{(k)*}$ for all $i \in C_k$, $k \in [N]$. To this end, we now prove that the optimality condition of (2) is satisfied, that is, there exists $(z_{ij})_{i\neq j}$ such that $z_{ij} \in \partial\|x_i^* - x_j^*\|_2$ and $z_{ij} = -z_{ji}$ for all $i \neq j,\ i, j \in \mathcal{V}$, and

$$\nabla f_i(x_i^*) + \gamma \sum_{j\neq i} w_{\{i,j\}} z_{ij} = 0$$

for all $i \in \mathcal{V}$. Let

$$z_{ij}^* := \begin{cases} z^{(k,k')}, & (i \in C_k,\ j \in C_{k'},\ k \neq k'), \\ \frac{1}{n_k w_{\{i,j\}}}\left\{\frac{1}{\gamma}\left(\nabla f_j(x^{(k)*}) - \nabla f_i(x^{(k)*})\right) + p_j^{(k)} - p_i^{(k)}\right\}, & (i, j \in C_k,\ i \neq j), \end{cases}$$

where

$$p_i^{(k)} := \sum_{l\neq k}\left(w_i^{(l)} - \frac{1}{n_k} w^{(k,l)}\right) z^{(k,l)}.$$

Obviously, it holds that $z_{ij}^* = -z_{ji}^*$ for any $i \neq j,\ i, j \in \mathcal{V}$. For all $i \in C_k,\ j \in C_{k'},\ k \neq k'$, it is valid $z_{ij}^* = z^{(k,k')} \in \partial\|x^{(k)*} - x^{(k')*}\|_2 = \partial\|x_i^* - x_j^*\|_2$. For arbitrary $i, j \in C_k$, $k \in [N]$, we have

$$\begin{aligned}
&\|\nabla f_j(x^{(k)*}) - \nabla f_i(x^{(k)*})\|_2 \\
&\leq \|\nabla f_j(x^{(k)*}) - \nabla f_j(\overline{x}^{(k)})\|_2 + \|\nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)})\|_2 + \|\nabla f_i(\overline{x}^{(k)}) - \nabla f_i(x^{(k)*})\|_2 \\
&\leq \|\nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)})\|_2 + (L_i + L_j)\|\overline{x}^{(k)} - x^{(k)*}\|_2 \\
&\leq \|\nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)})\|_2 + \frac{\gamma(L_i + L_j)}{\alpha_k}\sum_{l\neq k} w^{(k,l)},
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second one from the L_i-smoothness of f_i, and the third one from (5) and (28). Accordingly, we obtain

$$\begin{aligned}
&\left\|z_{ij}^*\right\|_2 \\
&= \frac{1}{n_k w_{\{i,j\}}} \left\| \frac{1}{\gamma} \left(\nabla f_j(x^{(k)*}) - \nabla f_i(x^{(k)*}) \right) + p_j^{(k)} - p_i^{(k)} \right\|_2 \\
&\leq \frac{1}{n_k w_{\{i,j\}} \gamma} \left\| \nabla f_j(x^{(k)*}) - \nabla f_i(x^{(k)*}) \right\|_2 + \frac{1}{n_k w_{\{i,j\}}} \left\| p_j^{(k)} - p_i^{(k)} \right\|_2 \\
&\leq \frac{1}{n_k w_{\{i,j\}} \gamma} \left\| \nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)}) \right\|_2 + \frac{1}{n_k w_{\{i,j\}}} \Big(\sum_{l \neq k} \left| w_i^{(l)} - w_j^{(l)} \right| + \frac{L_i + L_j}{\alpha_k} \sum_{l \neq k} w^{(k,l)} \Big) \\
&\leq \frac{1}{n_k w_{\{i,j\}} \gamma_{\min}} \left\| \nabla f_j(\overline{x}^{(k)}) - \nabla f_i(\overline{x}^{(k)}) \right\|_2 + \frac{\mu_{ij}^{(k)}}{n_k w_{\{i,j\}}} \\
&\leq \frac{n_k w_{\{i,j\}} - \mu_{ij}^{(k)}}{n_k w_{\{i,j\}}} + \frac{\mu_{ij}^{(k)}}{n_k w_{\{i,j\}}} \\
&= 1,
\end{aligned}$$

where the first inequality follows from the triangle inequality, the second one from the definition of $p^{(k)} := (p_i^{(k)})_{i \in C_k}$ and the previous inequality, the third and fourth ones from the definitions of $\mu_{ij}^{(k)}$ and $\gamma_{\min}$, respectively. This implies $z_{ij}^* \in \partial \|0\|_2 = \partial \|x_i^* - x_j^*\|_2$ for all $i, j \in C_k$, $k \in [N]$. On the other hand, we have

$$\begin{aligned}
&\nabla f_i(x_i^*) + \gamma \sum_{j \neq i} w_{\{i,j\}} z_{ij}^* \\
&= \nabla f_i(x^{(k)*}) + \gamma \sum_{l \neq k} w_i^{(l)} z^{(k,l)} \\
&\qquad\qquad + \gamma \sum_{\substack{j \neq i \\ j \in C_k}} w_{\{i,j\}} \frac{1}{n_k w_{\{i,j\}}} \left\{ \frac{1}{\gamma} \left(\nabla f_j(x^{(k)*}) - \nabla f_i(x^{(k)*}) \right) + p_j^{(k)} - p_i^{(k)} \right\} \\
&= \frac{1}{n_k} \sum_{j \in C_k} \nabla f_j(x^{(k)*}) + \gamma \sum_{l \neq k} w_i^{(l)} z^{(k,l)} + \frac{1}{n_k} \gamma \sum_{\substack{j \neq i \\ j \in C_k}} \sum_{l \neq k} \left(w_j^{(l)} - w_i^{(l)} \right) z^{(k,l)} \\
&= \frac{1}{n_k} \sum_{j \in C_k} \nabla f_j(x^{(k)*}) + \frac{1}{n_k} \gamma \sum_{l \neq k} \sum_{j \in C_k} w_j^{(l)} z^{(k,l)} \\
&= \frac{1}{n_k} \left(\nabla f^{(k)}(x^{(k)*}) + \gamma \sum_{l \neq k} w^{(k,l)} z^{(k,l)} \right) \\
&= 0,
\end{aligned}$$

where the first equality is by the definition of z_{ij}^*, the second one is by the definitions of $p^{(k)}$ and $w^{(l)}$, and the final equality follows from (27). These results show that $(x_1^*, ..., x_n^*)$ is the unique optimal solution of (2) because of the strict convexity of the objective function of (2). Thus we conclude that $\overline{\mathcal{P}}$ perfectly recovers $\mathcal{P}$.

Proof of Statement 2. Observe that if $x^{(1)*} = \cdots = x^{(N)*}$, then $x^{(k)*} = \overline{x}$ holds for $k \in [N]$ because of the definition of $\overline{x}$ and the strict convexity of $\sum_{i \in \mathcal{V}} f_i(x)$. From the inequality (28), we have

$$\max_k \frac{\|\nabla f^{(k)}(\overline{x})\|_2}{\sum_{l \neq k} w^{(k,l)}} \leq \gamma.$$

Therefore, if $\gamma < \max_k \frac{\|\nabla f^{(k)}(\overline{x})\|_2}{\sum_{l \neq k} w^{(k,l)}}$, then $\overline{x}^{(1)} = \cdots = \overline{x}^{(N)}$ does not hold. In addition, if we take $\gamma \geq \gamma_{\min}$, then $x_i^* = \overline{x}^{(k)} (i \in C_k)$ is the optimal solution of (2), as in the proof of Statement 1. Thus $\overline{\mathcal{P}}$ is a non-trivial coarsening of $\mathcal{P}$. ■

A.2 Proof of Theorem 6

Proof *Proof of Statement 1.* Note that if $\tau_K(x_1^\gamma, ..., x_n^\gamma) = 0$ holds, $\boldsymbol{x}^\gamma$ is a minimizer of (10)–(11). Assume that $\tau_K(x_1^\gamma, ..., x_n^\gamma) > 0$. In this case, let $\mathcal{E}' \subset \mathcal{E}$ be a set of edges $\{i, j\} \in \mathcal{E}$ whose $\|x_i^\gamma - x_j^\gamma\|_2$ is in the smallest $|\mathcal{E}| - K$ components and divide $\mathcal{V}$ into connected components $\mathcal{C}_1, \ldots, \mathcal{C}_m$ of the graph $(\mathcal{V}, \mathcal{E}')$, then we set

$$x_i' := \sum_{j \in \mathcal{C}_k} \frac{x_j^\gamma}{|\mathcal{C}_k|},$$

for $i \in \mathcal{C}_k$, $k \in [m]$. Obviously, $\tau_K(x_1', ..., x_n') = 0$ and $\|x_i'\|_2 \leq C$ are fulfilled. If $i, j \in \mathcal{C}_k, k \in [m]$ and $i \neq j$, then there exists a simple path between i and j on $(\mathcal{V}, \mathcal{E}')$, so

$$\begin{aligned} \|x_i^\gamma - x_j^\gamma\|_2 &\leq \sum_{\{i', j'\} \in \mathcal{E}'} \|x_{i'}^\gamma - x_{j'}^\gamma\|_2 \\ &\leq \tau_K(x_1^\gamma, ..., x_n^\gamma). \end{aligned}$$

Thus we obtain

$$\begin{aligned} \|x_i' - x_i^\gamma\|_2 &\leq \sum_{j \in \mathcal{C}_k} \frac{\|x_i^\gamma - x_j^\gamma\|_2}{|\mathcal{C}_k|} \\ &\leq \sum_{j \in \mathcal{C}_k} \frac{\tau_K(x_1^\gamma, ..., x_n^\gamma)}{|\mathcal{C}_k|} \\ &\leq \tau_K(x_1^\gamma, ..., x_n^\gamma), \end{aligned} \tag{30}$$

for all $i \in \mathcal{C}_k, k \in [m]$. From $\|x_i^\gamma\|_2 \leq C$ and f_i's L_i-smoothness, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{V}} f(x_i^\gamma) + \gamma\tau_K(\boldsymbol{x}^\gamma) - \left(\sum_{i\in\mathcal{V}} f(x_i') + \gamma\tau_K(\boldsymbol{x}')\right) \\
&\geq \gamma\tau_K(\boldsymbol{x}^\gamma) + \sum_{i\in\mathcal{V}} \left(\nabla f_i(x_i^\gamma)^\top (x_i^\gamma - x_i') - \frac{L_i}{2}\|x_i^\gamma - x_i'\|_2^2\right) \\
&\geq \gamma\tau_K(\boldsymbol{x}^\gamma) - \sum_{i\in\mathcal{V}} \|x_i^\gamma - x_i'\|_2 \left(\|\nabla f_i(x_i^\gamma)\|_2 + \frac{L_i}{2}\|x_i^\gamma - x_i'\|_2\right) \\
&\geq \gamma\tau_K(\boldsymbol{x}^\gamma) - \sum_{i\in\mathcal{V}} \|x_i^\gamma - x_i'\|_2 \left(\|\nabla f_i(0)\|_2 + \|\nabla f_i(x_i^\gamma) - \nabla f_i(0)\|_2 + \frac{L_i}{2}\left(\|x_i^\gamma\|_2 + \|x_i'\|_2\right)\right) \\
&\geq \gamma\tau_K(\boldsymbol{x}^\gamma) - \sum_{i\in\mathcal{V}} \|x_i^\gamma - x_i'\|_2 \left(\|\nabla f_i(0)\|_2 + L_i\|x_i^\gamma\|_2 + \frac{L_i}{2}\left(\|x_i^\gamma\|_2 + \|x_i'\|_2\right)\right) \\
&\geq \gamma\tau_K(\boldsymbol{x}^\gamma) - \sum_{i\in\mathcal{V}} \|x_i^\gamma - x_i'\|_2 \left(\|\nabla f_i(0)\|_2 + 2L_iC\right) \\
&\geq \gamma\tau_K(\boldsymbol{x}^\gamma) - \sum_{i\in\mathcal{V}} \tau_K(\boldsymbol{x}^\gamma) \left(\|\nabla f_i(0)\|_2 + 2L_iC\right) \\
&= \tau_K(\boldsymbol{x}^\gamma) \left(\gamma - \sum_{i\in\mathcal{V}} \left(\|\nabla f_i(0)\|_2 + 2L_iC\right)\right) \\
&> 0,
\end{aligned} \tag{31}$$

where the first and fourth inequalities follow from the L_i-smoothness of f_i, where we apply the inequality (4) to the first one, the second one from the Cauchy-Schwarz inequality, the third one from the triangle inequality, the fifth one from the boundedness of $\boldsymbol{x}^\gamma$ and $\boldsymbol{x}'$, the sixth one from the inequality (30). The above inequality (31) contradicts the optimality of $\boldsymbol{x}^\gamma$.

Proof of Statement 2. Note that if $\tau_K(x_1^\gamma, ..., x_n^\gamma) = 0$ is fulfilled, $\boldsymbol{x}^\gamma$ is a local minimizer of (10)–(11). Assume $\tau_K(x_1^\gamma, ..., x_n^\gamma) > 0$. Let us define $\mathcal{E}'$ as in the proof of the statement 1., let

$$v_{\{i,j\}} := \begin{cases} 1, & \{i,j\} \in \mathcal{E}', \\ 0, & otherwise, \end{cases}$$

and consider the following problem:

$$\underset{x_1,...,x_n}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x_i) + \gamma \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}}\|x_i - x_j\|_2. \tag{32}$$

Note that $\sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}}\|x_i^\gamma - x_j^\gamma\|_2 = \tau_K(x_1^\gamma, ..., x_n^\gamma)$. We have $\sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}}\|x_i - x_j\|_2 \geq \tau_K(x_1, ..., x_n)$ for any $(x_1, ..., x_n)$ by the definition of τ_K. Since $\boldsymbol{x}^\gamma$ is locally optimal to (12), $\boldsymbol{x}^\gamma$ is a local minimizer of (32). Because of the convexity of (32), $\boldsymbol{x}^\gamma$ is optimal to (32). Determining $\boldsymbol{x}'$ in the same way as in the proof for the statement 1., we have $\sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}}\|x_i' - x_j'\|_2 = 0$, $\|x_i'\|_2 \leq C$, and the inequality (30). By the same calculation as in (31), we reach the contradiction to the fact that $\boldsymbol{x}^\gamma$ is optimal to (32). ■

A.3 Proof of Lemma 7

Proof Let $\mathcal{E}' \subset \mathcal{E}$ be a set of edges $\{i,j\} \in \mathcal{E}$ whose $\|x_i^* - x_j^*\|_2$ is in the smallest $|\mathcal{E}| - K$ components out of all the $|\mathcal{E}|$ components, then we define

$$v_{\{i,j\}} := \begin{cases} 1, & \{i,j\} \in \mathcal{E}', \\ 0, & otherwise, \end{cases}$$

and consider the following problem:

$$\underset{x_1,\ldots,x_n}{\text{minimize}} \quad \frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i - a_i\|_2^2 + \gamma \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i - x_j\|_2. \tag{33}$$

From the convexity of $\frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i - a_i\|_2^2$, as in the proof of the second statement of Theorem 6, $\boldsymbol{x}^*$ is an optimal solution of (33). Assume that there exists an $i \in \mathcal{V}$ such that $\|x_i^*\|_2 > C$. Let $O = \{i \mid \|x_i^*\|_2 > C\}$, and define

$$x_i' := \begin{cases} \frac{R}{\|x_i^*\|_2} x_i^*, & i \in O, \\ x_i^*, & i \notin O. \end{cases}$$

Obviously, it is valid that

$$\|x_i^* - a_i\|_2^2 = \|x_i' - a_i\|_2^2,$$

for $i \notin O$, and

$$\|x_i^* - x_j^*\|_2 = \|x_i' - x_j'\|_2,$$

for $i, j \notin O$. Because x_i' is the projection of x_i^* onto the closed convex set $\{x \in \mathbb{R}^p : \|x\|_2 \leq R\}$, we obtain

$$\begin{aligned} \|x_i^* - x_j^*\|_2^2 &= \|x_i^* - x_j'\|_2^2 \\ &= \|x_i^* - x_i' + x_i' - x_j'\|_2^2 \\ &= \|x_i^* - x_i'\|_2^2 + 2(x_i^* - x_i')^\top (x_i' - x_j') + \|x_i' - x_j'\|_2^2 \\ &\geq \left\| \left(1 - \frac{R}{\|x_i^*\|_2}\right) x_i^* \right\|_2^2 + \|x_i' - x_j'\|_2^2 \\ &> \|x_i' - x_j'\|_2^2, \end{aligned}$$

for $i \in O, j \notin O$. In the same way, we have

$$\begin{aligned} &\|x_i^* - x_j^*\|_2^2 \\ &= \|x_i^* - x_i' + x_i' - x_j' + x_j' - x_j^*\|_2^2 \\ &= \|x_i^* - x_i' + x_j' - x_j^*\|_2^2 + 2(x_i^* - x_i')^\top (x_i' - x_j') + 2(x_j' - x_j^*)^\top (x_i' - x_j') + \|x_i' - x_j'\|_2^2 \\ &\geq \|x_i' - x_j'\|_2^2, \end{aligned}$$

for $i, j \in O$, and

$$\begin{aligned}\|x_i^* - a_i\|_2^2 &= \|x_i^* - x_i' + x_i' - a_i\|_2^2 \\ &= \|x_i^* - x_i'\|_2^2 + 2(x_i^* - x_i')^\top (x_i' - a_i) + \|x_i' - a_i\|_2^2 \\ &\geq \left\| \left(1 - \frac{R}{\|x_i^*\|_2}\right) x_i^* \right\|_2^2 + \|x_i' - a_i\|_2^2 \\ &\geq \left(\|x_i^*\|_2 - R\right)^2 + \|x_i' - a_i\|_2^2 \\ &> \|x_i' - a_i\|_2^2,\end{aligned}$$

for $i \in O$. This implies that

$$\begin{aligned}\frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i^* - a_i\|_2^2 &> \frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i' - a_i\|_2^2, \\ \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i^* - x_j^*\|_2 &\geq \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i' - x_j'\|_2.\end{aligned}$$

Thus we have

$$\frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i^* - a_i\|_2^2 + \gamma \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i^* - x_j^*\|_2 > \frac{1}{2}\sum_{i\in\mathcal{V}} \|x_i' - a_i\|_2^2 + \gamma \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i' - x_j'\|_2,$$

which contradicts the fact that $\boldsymbol{x}^*$ is optimal to (33). Consequently, we have $\|x_i^*\|_2 \leq C$ for all $i \in \mathcal{V}$. ■

A.4 Proof of Corollary 8

Proof Since $f_i(x_i) = \frac{1}{2}\|x_i - a_i\|_2^2$ is 1-smooth and $\|\nabla f_i(0)\|_2 = \|-a_i\|_2 \leq C$, we have

$$\sum_{i\in\mathcal{V}} \left(\|\nabla f_i(0)\|_2 + 2L_i C\right) \leq \sum_{i\in\mathcal{V}} (C + 2C) = 3nC.$$

This completes the proof. ■

A.5 Proof of Lemma 9

Proof From the convexity of $\sum_{i\in\mathcal{V}} f_i(x_i)$, as in the proof of the second statement of Theorem 6, $\boldsymbol{x}^*$ is optimal to

$$\underset{x_1,\ldots,x_n}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x_i) + \gamma \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i - x_j\|_2, \tag{34}$$

where $v_{\{i,j\}}$ is defined in the same way. Since $\boldsymbol{x}^*$ is optimal to (34), we have

$$\sum_{i\in\mathcal{V}} f_i(x_i^*) \leq \sum_{i\in\mathcal{V}} f_i(x_i) + \gamma \sum_{\{i,j\}\in\mathcal{E}} v_{\{i,j\}} \|x_i - x_j\|_2 \leq \sum_{i\in\mathcal{V}} f_i(0). \tag{35}$$

From the strong convexity of f_i, combining (35) and (6) yields

$$\begin{aligned}\frac{\alpha}{2}\|x_i^* - \overline{x}_i\|_2^2 &\leq \sum_{j\in\mathcal{V}} \frac{\alpha_j}{2}\|x_j^* - \overline{x}_j\|_2^2 \\ &\leq \sum_{j\in\mathcal{V}} \left(f_j(x_j^*) - f_j(\overline{x}_j)\right) \\ &\leq \sum_{j\in\mathcal{V}} \left(f_j(0) - f_j(\overline{x}_j)\right)\end{aligned}$$

for all $i \in \mathcal{V}$. Applying the triangle inequality to this, we get

$$\begin{aligned}\|x_i^*\|_2 &\leq \|x_i^* - \overline{x}_i\|_2 + \|\overline{x}_i\|_2 \\ &\leq \left(\frac{2}{\alpha}\sum_{j\in\mathcal{V}} \left(f_j(0) - f_j(\overline{x}_j)\right)\right)^{\frac{1}{2}} + \|\overline{x}_i\|_2 \\ &\leq C.\end{aligned}$$

This completes the proof. ■

A.6 Proof of Corollary 10

Proof Note that for any $i \in \mathcal{V}$, f_i is $\lambda_{\min}(A_i)$-strongly convex and $\lambda_{\max}(A_i)$-smooth, and the gradient and minimizer of f_i are given by $\nabla f_i(x_i) = A_i x_i - B_i$ and $A_i^{-1}B_i$, respectively. By applying Theorem 6 and Lemma 9, we have the desired result. ■

A.7 Proof of Lemma 11

Proof First, note that the equation

$$T_K(\boldsymbol{z}) = \sum_{k\in\Lambda_1} \|z_k\|_2 + \sum_{k\in\Lambda} \|z_k\|_2 \tag{36}$$

holds for any $\Lambda \subset \Lambda_2$ such that $|\Lambda| = m - K - |\Lambda_1|$. Let

$$\begin{aligned}\Lambda_1^\eta &:= \{k \mid \|z_k + \eta v_k\|_2 < \|(z+\eta v)_{(K)}\|_2\}, \\ \Lambda_2^\eta &:= \{k \mid \|z_k + \eta v_k\|_2 = \|(z+\eta v)_{(K)}\|_2\}.\end{aligned}$$

Observe that there exists a positive number ε such that $\|z_k + \eta v_k\|_2 < \|(z+\eta v)_{(K)}\|_2$ for all $k \in \Lambda_1$ and $\|z_k + \eta v_k\|_2 > \|(z+\eta v)_{(K)}\|_2$ for all $k \in (\Lambda_1 \cup \Lambda_2)^c$ whenever $0 < \eta < \varepsilon$ because of the continuity of ℓ_2 norm. Hence $\Lambda_1 \subset \Lambda_1^\eta$ and $\Lambda_1^\eta \cup \Lambda_2^\eta \subset \Lambda_1 \cup \Lambda_2$ hold whenever $0 < \eta < \varepsilon$. From this, we obtain

$$T_K(\boldsymbol{z} + \eta\boldsymbol{v}) = \sum_{k\in\Lambda_1} \|z_k + \eta v_k\|_2 + \min_{\substack{\Lambda\subset\Lambda_2 \\ |\Lambda| = m-K-|\Lambda_1|}} \sum_{k\in\Lambda} \|z_k + \eta v_k\|_2, \tag{37}$$

for $\eta \in (0, \varepsilon)$. Combining (36) and (37) yields

$$T_K(\boldsymbol{z} + \eta \boldsymbol{v}) = \sum_{k \in \Lambda_1} (\|z_k + \eta v_k\|_2 - \|z_k\|_2) + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m - K - |\Lambda_1|}} \sum_{k \in \Lambda} (\|z_k + \eta v_k\|_2 - \|z_k\|_2).$$

Furthermore, taking the limit $\eta \searrow 0$, for any $k \in [m]$, we have

$$\frac{\|z_k + \eta v_k\|_2 - \|z_k\|_2}{\eta} \to \begin{cases} \frac{z_k}{\|z_k\|_2}^\top v_k, & z_k \neq 0, \\ \|v_k\|_2, & z_k = 0, v_k \neq 0, \\ 0, & z_k = 0, v_k = 0, \end{cases}$$

that is, $\frac{\|z_k + \eta v_k\|_2 - \|z_k\|_2}{\eta} \to \delta(z_k, v_k)^\top v_k$. Thus, we obtain

$$\begin{aligned} \mathrm{d}T_K(\boldsymbol{z}; \boldsymbol{v}) &= \lim_{\eta \searrow 0} \frac{T_K(\boldsymbol{z} + \eta \boldsymbol{v}) - T_K(\boldsymbol{z})}{\eta} \\ &= \lim_{\eta \searrow 0} \frac{\sum_{k \in \Lambda_1} (\|z_k + \eta v_k\|_2 - \|z_k\|_2)}{\eta} + \lim_{\eta \searrow 0} \frac{\min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m - K - |\Lambda_1|}} \sum_{k \in \Lambda} (\|z_k + \eta v_k\|_2 - \|z_k\|_2)}{\eta} \\ &= \sum_{k \in \Lambda_1} \lim_{\eta \searrow 0} \frac{(\|z_k + \eta v_k\|_2 - \|z_k\|_2)}{\eta} + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m - K - |\Lambda_1|}} \sum_{k \in \Lambda} \lim_{\eta \searrow 0} \frac{(\|z_k + \eta v_k\|_2 - \|z_k\|_2)}{\eta} \\ &= \sum_{k \in \Lambda_1} \delta(z_k, v_k)^\top v_k + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m - K - |\Lambda_1|}} \sum_{k \in \Lambda} \delta(z_k, v_k)^\top v_k, \end{aligned}$$

where the third equality is established by interchanging "min" and "limit," which is possible because $\{\Lambda \subset \Lambda_2 \mid |\Lambda| = m - K - |\Lambda_1|\}$ is a finite set. ■

A.8 Proof of Proposition 12

Proof To prove the proposition by contradiction, suppose that $\boldsymbol{x}^*$ is not a locally optimal solution of (15). Then there exists a sequence $\{\boldsymbol{x}^t\}$ such that $\boldsymbol{x}^t \to \boldsymbol{x}^*$ and $f(\boldsymbol{x}^*) + \gamma T_K(D\boldsymbol{x}^*) > f(\boldsymbol{x}^t) + \gamma T_K(D\boldsymbol{x}^t)$ for all t. Setting

$$\begin{aligned} \Lambda_1 &:= \{k \mid \|(D\boldsymbol{x}^*)_k\|_2 < \|(D\boldsymbol{x}^*)_{(K)}\|_2\}, \\ \Lambda_2 &:= \{k \mid \|(D\boldsymbol{x}^*)_k\|_2 = \|(D\boldsymbol{x}^*)_{(K)}\|_2\}, \\ \Lambda_1^t &:= \{k \mid \|(D\boldsymbol{x}^t)_k\|_2 < \|(D\boldsymbol{x}^t)_{(K)}\|_2\}, \\ \Lambda_2^t &:= \{k \mid \|(D\boldsymbol{x}^t)_k\|_2 = \|(D\boldsymbol{x}^t)_{(K)}\|_2\}, \end{aligned}$$

we have

$$\begin{aligned} &T_K(D\boldsymbol{x}^t) - T_K(D\boldsymbol{x}^*) \\ &= \sum_{k \in \Lambda_1} (\|(D\boldsymbol{x}^t)_k\|_2 - \|(D\boldsymbol{x}^*)_k\|_2) + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m - K - |\Lambda_1|}} \sum_{k \in \Lambda} (\|(D\boldsymbol{x}^t)_k\|_2 - \|(D\boldsymbol{x}^*)_k\|_2), \end{aligned}$$

since $\Lambda_1 \subset \Lambda_1^t$ and $\Lambda_1^t \cup \Lambda_2^t \subset \Lambda_1 \cup \Lambda_2$ hold for sufficiently large t as in the proof of Lemma 11. Noting that for any $z, z' \in \mathbb{R}^p$,

$$\|z\|_2 - \|z'\|_2 \geq \delta(z', z - z')^\top (z - z'),$$

we have

$$\begin{aligned}
&T_K(D\boldsymbol{x}^t) - T_K(D\boldsymbol{x}^*) \\
&\geq \sum_{k \in \Lambda_1} \delta((D\boldsymbol{x}^*)_k, (D\boldsymbol{v})_k)^\top (D\boldsymbol{v})_k + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m-K-|\Lambda_1|}} \sum_{k \in \Lambda} \delta((D\boldsymbol{x}^*)_k, (D\boldsymbol{v})_k)^\top (D\boldsymbol{v})_k,
\end{aligned}$$

where $\boldsymbol{v} = \boldsymbol{x}^t - \boldsymbol{x}^*$. This as well as the convexity of f and Lemma 11 yield

$$\begin{aligned}
0 &> f(\boldsymbol{x}^t) + \gamma T_K(D\boldsymbol{x}^t) - (f(\boldsymbol{x}^*) + \gamma T_K(D\boldsymbol{x}^*)) \\
&\geq \nabla f(\boldsymbol{x}^*)^\top \boldsymbol{v} \\
&\quad + \gamma \Big[\sum_{k \in \Lambda_1} \delta((D\boldsymbol{x}^*)_k, (D\boldsymbol{v})_k)^\top (D\boldsymbol{v})_k + \min_{\substack{\Lambda \subset \Lambda_2 \\ |\Lambda| = m-K-|\Lambda_1|}} \sum_{k \in \Lambda} \delta((D\boldsymbol{x}^*)_k, (D\boldsymbol{v})_k)^\top (D\boldsymbol{v})_k \Big] \\
&= \nabla f(\boldsymbol{x}^*)^\top \boldsymbol{v} + \gamma \mathrm{d}T_K(D\boldsymbol{x}^*; D\boldsymbol{v}) \\
&= \mathrm{d}(f + \gamma T_K \circ D)(\boldsymbol{x}^*; \boldsymbol{v}),
\end{aligned}$$

which contradicts the fact that $\boldsymbol{x}^*$ is a stationary point of (15). ■

A.9 Proof of Proposition 13

Proof Let $\{(\boldsymbol{x}^{t_i}, \boldsymbol{z}^{t_i}, \boldsymbol{y}^{t_i})\}$ be a subsequence of $\{(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t)\}$ that converges to $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$. From the fact that $(\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2, \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|_2, \|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|_2)$ converges to $(0, 0, 0)$, the subsequence $\{(\boldsymbol{x}^{t_i+1}, \boldsymbol{z}^{t_i+1}, \boldsymbol{y}^{t_i+1})\}$ also converges to $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$. By the relation (20), the equation

$$\boldsymbol{y}^{t_i+1} = \boldsymbol{y}^{t_i} + \rho(\boldsymbol{z}^{t_i+1} - D\boldsymbol{x}^{t_i+1})$$

holds. Letting $i \to \infty$ yields

$$\boldsymbol{z}^* = D\boldsymbol{x}^*. \tag{38}$$

Taking the limit of the optimality condition of (25), we have

$$\nabla f(\boldsymbol{x}^{t_i+1}) + \rho D^\top \left(D\boldsymbol{x}^{t_i+1} - \boldsymbol{z}^{t_i+1} - \frac{1}{\rho} \boldsymbol{y}^{t_i} \right) + \nabla \phi(\boldsymbol{x}^{t_i+1}) - \nabla \phi(\boldsymbol{x}^{t_i}) = 0,$$

and combining it with (38) and continuity of ∇f and $\nabla \phi$, we obtain

$$\nabla f(\boldsymbol{x}^*) = D^\top \boldsymbol{y}^*. \tag{39}$$

Since $\boldsymbol{z}^{t_i+1}$ is optimal to (18), the inequality

$$\begin{aligned}&\gamma T_K(\boldsymbol{z}^{t_i+1}) + (\boldsymbol{y}^{t_i})^\top \boldsymbol{z}^{t_i+1} + \frac{\rho}{2}\|\boldsymbol{z}^{t_i+1} - D\boldsymbol{x}^{t_i}\|_2^2 \\ &\leq \gamma T_K(\boldsymbol{z}^* + \eta D\boldsymbol{v}) + (\boldsymbol{y}^{t_i})^\top(\boldsymbol{z}^* + \eta D\boldsymbol{v}) + \frac{\rho}{2}\|\boldsymbol{z}^* + \eta D\boldsymbol{v} - D\boldsymbol{x}^{t_i}\|_2^2\end{aligned}$$

holds for any $\eta > 0$ and $\boldsymbol{v} \in \mathbb{R}^N$. By the continuity of T_K and (38), letting $i \to \infty$ yields

$$\gamma T_K(D\boldsymbol{x}^*) + (\boldsymbol{y}^*)^\top D\boldsymbol{x}^* \leq \gamma T_K(D\boldsymbol{x}^* + \eta D\boldsymbol{v}) + (\boldsymbol{y}^*)^\top(D\boldsymbol{x}^* + \eta D\boldsymbol{v}) + \frac{\rho}{2}\|\eta D\boldsymbol{v}\|_2^2.$$

Combining this with (39), we see that

$$\begin{aligned}&\eta\nabla f(\boldsymbol{x}^*)^\top\boldsymbol{v} + \gamma T_K(D(\boldsymbol{x}^* + \eta\boldsymbol{v})) - \gamma T_K(D\boldsymbol{x}^*) + \eta^2\frac{\rho}{2}\|D\boldsymbol{v}\|_2^2 \\ &= \eta(D^\top\boldsymbol{y}^*)^\top\boldsymbol{v} + \gamma T_K(D\boldsymbol{x}^* + \eta D\boldsymbol{v}) - \gamma T_K(D\boldsymbol{x}^*) + \eta^2\frac{\rho}{2}\|D\boldsymbol{v}\|_2^2 \\ &= (\boldsymbol{y}^*)^\top(\eta D\boldsymbol{v}) + \gamma T_K(D\boldsymbol{x}^* + \eta D\boldsymbol{v}) - \gamma T_K(D\boldsymbol{x}^*) + \frac{\rho}{2}\|\eta D\boldsymbol{v}\|_2^2 \\ &\geq 0.\end{aligned}$$

By dividing both sides of this inequality by η and taking the limit with $\eta \searrow 0$, we obtain

$$\mathrm{d}(f + \gamma T_K \circ D)(\boldsymbol{x}^*; \boldsymbol{v}) = \nabla f(\boldsymbol{x}^*)^\top\boldsymbol{v} + \gamma\mathrm{d}(T_K \circ D)(\boldsymbol{x}^*; \boldsymbol{v}) \geq 0,$$

which implies that $\boldsymbol{x}^*$ is a stationary point of (15). Since f is a differentiable convex function, $\boldsymbol{x}^*$ is shown to be locally optimal to (15) by Proposition 12. ■

A.10 Proof of Theorem 14

Proof From the optimality condition of (25) and the equation (20), we obtain

$$\begin{aligned}&\sigma\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|_2^2 \\ &\leq \|D^\top(\boldsymbol{y}^{t+1} - \boldsymbol{y}^t)\|_2^2 \\ &= \|\nabla f(\boldsymbol{x}^{t+1}) + \nabla\phi(\boldsymbol{x}^{t+1}) - \nabla\phi(\boldsymbol{x}^t) - \nabla f(\boldsymbol{x}^t) - \nabla\phi(\boldsymbol{x}^t) + \nabla\phi(\boldsymbol{x}^{t-1})\|_2^2 \\ &\leq \frac{1}{r}\|\nabla f(\boldsymbol{x}^{t+1}) + \nabla\phi(\boldsymbol{x}^{t+1}) - \nabla f(\boldsymbol{x}^t) - \nabla\phi(\boldsymbol{x}^t)\|_2^2 + \frac{1}{1-r}\|\nabla\phi(\boldsymbol{x}^t) - \nabla\phi(\boldsymbol{x}^{t-1})\|_2^2 \\ &\leq \frac{L_1^2}{r}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2 + \frac{L_2^2}{1-r}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2,\end{aligned} \tag{40}$$

where the number r satisfies the assumption *(A6)* and the first inequality follows from the assumption *(A1)*, the second one from the inequality $\|a+b\|_2^2 \leq \frac{\|a\|_2^2}{r} + \frac{\|b\|_2^2}{1-r}$, the third one from the assumptions *(A3)* and *(A5)*. On the other hand, combining the equation (20) with the triangle inequality yields

$$\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|_2 \leq \|D(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)\|_2 + \frac{1}{\rho}\|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|_2 + \frac{1}{\rho}\|\boldsymbol{y}^t - \boldsymbol{y}^{t-1}\|_2.$$

The above two inequalities imply that if the sequence $\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2$ converges to 0, then both $\|\boldsymbol{z}^{t+1}-\boldsymbol{z}^t\|_2$ and $\|\boldsymbol{y}^{t+1}-\boldsymbol{y}^t\|_2$ also converge to 0. Thus we next show that $\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2$ converges to 0.

From (20) and (40), we obtain

$$\begin{aligned}L_\rho(\boldsymbol{x}^{t+1},\boldsymbol{z}^{t+1},\boldsymbol{y}^{t+1})-L_\rho(\boldsymbol{x}^{t+1},\boldsymbol{z}^{t+1},\boldsymbol{y}^t)&=(\boldsymbol{y}^{t+1}-\boldsymbol{y}^t)^\top(\boldsymbol{z}^{t+1}-D\boldsymbol{x}^{t+1})\\&=\frac{1}{\rho}\|\boldsymbol{y}^{t+1}-\boldsymbol{y}^t\|_2^2\\&\leq\frac{L_1^2}{\sigma\rho r}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2+\frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t-\boldsymbol{x}^{t-1}\|_2^2.\end{aligned}$$

Since $L_\rho(\boldsymbol{x},\boldsymbol{z}^{t+1},\boldsymbol{y}^t)+B_\phi(\boldsymbol{x},\boldsymbol{x}^t)$ is α_1-strongly convex by the assumption *(A4)*, using the inequality (6), we have

$$\begin{aligned}L_\rho(\boldsymbol{x}^{t+1},\boldsymbol{z}^{t+1},\boldsymbol{y}^t)-L_\rho(\boldsymbol{x}^t,\boldsymbol{z}^{t+1},\boldsymbol{y}^t)&\leq-\frac{\alpha_1}{2}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2-D_\phi(\boldsymbol{x}^{t+1},\boldsymbol{x}^t)\\&\leq-\frac{\alpha_1}{2}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2-\frac{\alpha_2}{2}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2\\&=-\frac{\alpha_1+\alpha_2}{2}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2,\end{aligned}$$

where we use the α_2-strong convexity of ϕ (the assumption *(A5)*) in the second inequality. Furthermore, because $\boldsymbol{z}^{t+1}$ is a minimizer of (18), the inequality

$$L_\rho(\boldsymbol{x}^t,\boldsymbol{z}^{t+1},\boldsymbol{y}^t)-L_\rho(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t)\leq 0$$

holds. By adding the above three inequalities together, we have

$$\begin{aligned}&L_\rho(\boldsymbol{x}^{t+1},\boldsymbol{z}^{t+1},\boldsymbol{y}^{t+1})-L_\rho(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t)\\&\leq\left(\frac{L_1^2}{\sigma\rho r}-\frac{\alpha_1+\alpha_2}{2}\right)\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2+\frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t-\boldsymbol{x}^{t-1}\|_2^2.\end{aligned}\tag{41}$$

Let $\{(\boldsymbol{x}^{t_i},\boldsymbol{z}^{t_i},\boldsymbol{y}^{t_i})\}$ be a subsequence of $\{(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{y}^t)\}$ that converges to a partial limit $(\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{y}^*)$. Noting that $C:=\frac{\alpha_1+\alpha_2}{2}-\frac{1}{\sigma\rho}\left(\frac{L_1^2}{r}+\frac{L_2^2}{1-r}\right)>0$ from the assumption *(A6)*, we have

$$\begin{aligned}&L_\rho(\boldsymbol{x}^{t_i},\boldsymbol{z}^{t_i},\boldsymbol{y}^{t_i})-L_\rho(\boldsymbol{x}^1,\boldsymbol{z}^1,\boldsymbol{y}^1)\\&\leq\left\{\sum_{t=1}^{t_i-1}\left(\frac{L_1^2}{\sigma\rho r}-\frac{\alpha_1+\alpha_2}{2}\right)\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2+\frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t-\boldsymbol{x}^{t-1}\|_2^2\right\}\\&=\sum_{t=1}^{t_i-1}\left(\frac{L_1^2}{\sigma\rho r}-\frac{\alpha_1+\alpha_2}{2}\right)\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2+\sum_{t=0}^{t_i-2}\frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2\\&=\frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^1-\boldsymbol{x}^0\|_2^2-\left(\frac{\alpha_1+\alpha_2}{2}-\frac{L_1^2}{\sigma\rho r}\right)\|\boldsymbol{x}^{t_i}-\boldsymbol{x}^{t_i-1}\|_2^2-\sum_{t=1}^{t_i-1}C\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2\\&\leq\frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^1-\boldsymbol{x}^0\|_2^2-C\sum_{t=1}^{t_i-1}\|\boldsymbol{x}^{t+1}-\boldsymbol{x}^t\|_2^2.\end{aligned}\tag{42}$$

By the continuity of L_ρ, we have

$$\lim_{i\to\infty} L_\rho(\boldsymbol{x}^{t_i}, \boldsymbol{z}^{t_i}, \boldsymbol{y}^{t_i}) = L_\rho(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*) > -\infty. \tag{43}$$

Taking the limit $i \to \infty$ in (42) with (43) leads to

$$\sum_{t=1}^{\infty} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2 < \infty,$$

which implies that $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2 \to 0$. Thus, $(\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2, \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|_2, \|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|_2) \to (0, 0, 0)$ holds. Since f and ϕ are continuously differentiable and f is a convex function by the assumptions *(A2)* and *(A3)*, Proposition 13 yields the desired result. ■

A.11 Proof of Theorem 15

Proof Since the assumptions *(A1)*, *(A3)–(A6)* are fulfilled, the inequality (41) holds. By slightly transforming it, we obtain

$$\begin{aligned} L_\rho(\boldsymbol{x}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{y}^{t+1}) + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2 - L_\rho(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t) - \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 \\ \leq \left(\frac{L_1^2}{\sigma\rho r} + \frac{L_2^2}{\sigma\rho(1-r)} - \frac{\alpha_1+\alpha_2}{2}\right)\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2 \\ \leq 0, \end{aligned}$$

which implies that the sequence $L_\rho(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t) + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2$ is monotonically decreasing. Hence, we see that

$$L_\rho(\boldsymbol{x}^t, \boldsymbol{z}^t, \boldsymbol{y}^t) + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 \leq L_\rho(\boldsymbol{x}^1, \boldsymbol{z}^1, \boldsymbol{y}^1) + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^1 - \boldsymbol{x}^0\|_2^2. \tag{44}$$

On the other hand, combining (20) and the optimality condition of (25) yields

$$\nabla f(\boldsymbol{x}^t) - D^\top \boldsymbol{y}^t + \nabla\phi(\boldsymbol{x}^t) - \nabla\phi(\boldsymbol{x}^{t-1}) = 0.$$

Then, from the assumptions *(A1)* and *(A5)*, we have

$$\begin{aligned} \sigma\|\boldsymbol{y}^t\|_2^2 &\leq \|D^\top \boldsymbol{y}^t\|_2^2 \\ &\leq \|\nabla f(\boldsymbol{x}^t) + \nabla\phi(\boldsymbol{x}^t) - \nabla\phi(\boldsymbol{x}^{t-1})\|_2^2 \\ &\leq \frac{1}{r}\|\nabla f(\boldsymbol{x}^t)\|_2^2 + \frac{1}{1-r}\|\nabla\phi(\boldsymbol{x}^t) - \nabla\phi(\boldsymbol{x}^{t-1})\|_2^2 \\ &\leq \frac{1}{r}\|\nabla f(\boldsymbol{x}^t)\|_2^2 + \frac{L_2^2}{1-r}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2. \end{aligned} \tag{45}$$

Combining (44) with (45) shows that

$$
\begin{aligned}
&L_\rho(\boldsymbol{x}^1, \boldsymbol{z}^1, \boldsymbol{y}^1) + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^1 - \boldsymbol{x}^0\|_2^2 \\
&\geq f(\boldsymbol{x}^t) + \gamma T_K(\boldsymbol{z}^t) + {\boldsymbol{y}^t}^\top(\boldsymbol{z}^t - D\boldsymbol{x}^t) + \frac{\rho}{2}\|\boldsymbol{z}^t - D\boldsymbol{x}^t\|_2^2 + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 \\
&= f(\boldsymbol{x}^t) + \gamma T_K(\boldsymbol{z}^t) + \frac{\rho}{2}\left\|\boldsymbol{z}^t - D\boldsymbol{x}^t + \frac{1}{\rho}\boldsymbol{y}^t\right\|_2^2 - \frac{1}{2\rho}\|\boldsymbol{y}^t\|_2^2 + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 \\
&\geq f(\boldsymbol{x}^t) - \frac{1}{2\sigma\rho r}\|\nabla f(\boldsymbol{x}^t)\|_2^2 - \frac{L_2^2}{2\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 + \frac{L_2^2}{\sigma\rho(1-r)}\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\|_2^2 \\
&\geq \left(1 - \frac{\zeta}{\sigma\rho r}\right) f(\boldsymbol{x}^t) + \frac{\zeta}{\sigma\rho r}\left\{f(\boldsymbol{x}^t) - \frac{1}{2\zeta}\|\nabla f(\boldsymbol{x}^t)\|_2^2\right\} \\
&\geq \left(1 - \frac{\zeta}{\sigma\rho r}\right) f(\boldsymbol{x}^t) + \frac{\zeta}{\sigma\rho r} f_{\mathrm{inf}}.
\end{aligned}
$$

Since f is coercive, the above inequality implies that $\{\boldsymbol{x}^t\}$ is bounded. The boundedness of $\{\boldsymbol{y}^t\}$ and $\{\boldsymbol{z}^t\}$ follows from (45) and (20), respectively. ■

References

Tal Amir, Ronen Basri, and Boaz Nadler. The trimmed lasso: Sparse recovery guarantees and practical optimization by the generalized soft-min penalty. *SIAM Journal on Mathematics of Data Science*, 3(3):900–929, 2021.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Dimitris Bertsimas, Martin S Copenhaver, and Rahul Mazumder. The trimmed lasso: Sparsity and robustness. *arXiv preprint arXiv:1708.04527*, 2017.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Giuseppe C. Calafiore and Laurent El Ghaoui. *Optimization Models*. Cambridge University Press, 2014.

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Ying Cui, Jong-Shi Pang, and Bodhisattva Sen. Composite difference-max programs for modern statistical estimation problems. *SIAM Journal on Optimization*, 28(4):3344–3374, 2018.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, 169(1):141–176, 2018.

David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396, 2015.

Toby Dylan Hocking, Armand Joulin, Francis Bach, and Jean-Philippe Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th International Conference on Machine Learning*, pages 745–752, 2011.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2: 193–218, 1985.

Alexander Jung and Nguyen Tran. Localized linear regression in networked data. *IEEE Signal Processing Letters*, 26(7):1090–1094, 2019.

Alexander Jung, Nguyen Tran, and Alexandru Mara. When is network lasso accurate? *Frontiers in Applied Mathematics and Statistics*, 3(28):1–11, 2018.

Devadatta Kulkarni, Darrell Schmidt, and Sze-Kai Tsui. Eigenvalues of tridiagonal pseudo-toeplitz matrices. *Linear Algebra and its Applications*, 297:63–80, 1999.

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

Fredrik Lindsten, Henrik Ohlsson, and Lennart Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pages 201–204, 2011.

Zhaosong Lu and Xiaorui Li. Sparse recovery via partial regularization: Models, theory, and algorithms. *Mathematics of Operations Research*, 43(4):1290–1316, 2018.

Ashkan Panahi, Devdatt Dubhashi, Fredrik D Johansson, and Chiranjib Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2769–2777, 2017.

Kristiaan Pelckmans, Joseph De Brabanter, Johan AK Suykens, and B De Moor. Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, 2005.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

Defeng Sun, Kim-Chuan Toh, and Yancheng Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *Journal of Machine Learning Research*, 22(9):1–32, 2021.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Changbo Zhu, Huan Xu, Chenlei Leng, and Shuicheng Yan. Convex optimization procedure for clustering: Theoretical revisit. In *Advances in Neural Information Processing Systems 27*, pages 1619–1627, 2014.