GRAPH-STRUCTURED SPARSE REGULARIZATION VIA CONVEX OPTIMIZATION

Hiroki Kuroda and Daichi Kitahara

College of Information Science and Engineering, Ritsumeikan University
E-mail: {kuroda, d-kita}@media.ritsumei.ac.jp

ABSTRACT

In this paper, we present a novel convex method for the graph-structured sparse recovery. While various structured sparsities can be represented as the graph-structured sparsity, graph-structured sparse recovery remains to be a challenging nonconvex problem. To solve this difficulty, we propose a convex penalty function which automatically identifies the relevant subgraph of an underlying graph. We design a graph-structured recovery model using the proposed penalty, and develop its first-order iterative solver which consists only of simple operations such as closed-form proximity operators and difference operator on the graph. Numerical experiments show that the proposed method has better estimation accuracy than the existing convex regularizations using fixed structures.

Index Terms— Graph-structured sparsity, convex regularization, proximal splitting method, first-order algorithm.

1. INTRODUCTION

Structural information of sparsity patterns is the key for the further enhancement of sparsity-aware signal processing and machine learning. Various structured sparsities such as block-sparsity and tree-structured sparsity can be treated as the graph-structured sparsity [1–3], where sparse nonzero entries are clustered in several connected subgraphs of an underlying graph equipped with the target signal. Graph-structured sparse model has numerous applications, which include clustering of biological network [4] and anomaly detection [5–8] of, e.g., water distribution pipe breakage [9], traffic congestion [10], and disease outbreak [11].

Despite the potential usefulness of the graph-structured sparsity, it remains a challenge to effectively exploit the graph-structured sparsity for signal processing and machine learning. The graph-structured sparsity is originally modeled in [1] as a nonconvex set. Since the projection onto this set is shown to be a NP-hard problem, the existing methods, e.g., [1–3], rely on approximate solvers for the nonconvex projection problem, and thus the computation of an optimal solution is challenging. On the other hand, existing convex regularization frameworks, e.g., [12–21], cannot flexibly incorporate the graph-structured sparsity because they need to define a relevant subgraph, characterizing the support of the target signal, in prior to the estimation.

In this paper, for the graph-structured sparse recovery, we present a novel convex method where the relevant subgraph of the underlying graph is automatically identified.[1] We exploit the observation that the graph-structured sparsity can be interpreted as the group sparsity by setting the groups to *connected components* [24] of the relevant subgraph of the underlying graph. Since the relevant subgraph is unknown a priori, we design a penalty function as the minimum of the mixed ℓ_2/ℓ_1 norm over the groups defined by the connected components of all candidate subgraphs of the underlying graph. Then, we derive the proposed convex penalty as its tight convex relaxation by utilizing the variational representation of the ℓ_2 norm [22, Lemma 1] and introducing latent variables characterized by a difference operator on the underlying graph. Based on the combination of proximal splitting techniques [25–31] and theory of perspective functions [32–34] we develop a first-order iterative algorithm which converges to an optimal solution of the proposed graph-structured recovery model. Numerical experiments on recovery of graph-structured sparse signals from noisy compressive measurements show the effectiveness of the proposed method.

2. PRELIMINARIES

2.1. Mathematical Notations

$\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ respectively denote the sets of all real numbers, all nonnegative real numbers, and all positive real numbers. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. For matrices or vectors, we denote the transpose by $(\cdot)^\top$. For $\boldsymbol{x} = (x_1, \ldots, x_N)^\top \in \mathbb{R}^N$ and an index set $\mathcal{I} \subset \{1, \ldots, N\}$, $\boldsymbol{x}_\mathcal{I} := (x_n)_{n\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ denotes the subvector of $\boldsymbol{x}$ indexed by $\mathcal{I}$. We define the support of $\boldsymbol{x} \in \mathbb{R}^N$ by $\mathrm{supp}(\boldsymbol{x}) := \{n \in \{1, \ldots, N\} \,|\, x_n \neq 0\}$. The ℓ_2 norm and the ℓ_1 norm of $\boldsymbol{x} \in \mathbb{R}^N$ are respectively defined by $\|\boldsymbol{x}\|_2 := \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ and $\|\boldsymbol{x}\|_1 := \sum_{n=1}^N |x_n|$. The weighted ℓ_0 pseudo-norm of $\boldsymbol{x}$ with the weight vector $\boldsymbol{w} \in \mathbb{R}^N_{++}$ is defined by $\|\boldsymbol{x}\|_{0,\boldsymbol{w}} := \sum_{n=1}^N w_n \mathbb{1}(x_n)$, where $\mathbb{1}(x) = 0$ if $x = 0$ and $\mathbb{1}(x) = 1$ otherwise.

[1] This paper can be regarded as an extension of our earlier studies [22, 23] which focus only on the block-sparsity.

We denote an undirected graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of undirected edges. Any nonempty graph $\mathcal{G}$ can be uniquely decomposed into *connected components* [24], that is, *maximal connected subgraphs* which form a *partition* of $\mathcal{V}$. We denote the connected components of $\mathcal{G}$ by

$$c_k(\mathcal{G}) \subset \mathcal{V} \quad \text{for } k = 1, \dots, \kappa(\mathcal{G}),$$

where $\kappa(\mathcal{G})$ is the number of connected components of $\mathcal{G}$.

2.2. Problem Formulation

We consider the estimation of $\boldsymbol{x}^\star \in \mathbb{R}^N$ supposed to be graph-structured sparse with an underlying graph $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ where $\mathcal{V}_0 := \{1, \dots, N\}$. Namely, nonzero entries of $\boldsymbol{x}^\star$ are clustered in several connected subgraphs of $\mathcal{G}_0$. In other words, $\boldsymbol{x}^\star$ is group-sparse by setting the groups to the connected components of $\mathcal{G}^\star = (\mathcal{V}_0, \mathcal{E}^\star)$, where $\mathcal{E}^\star \subset \mathcal{E}_0$ is defined by eliminating irrelevant edges from $\mathcal{E}_0$. More precisely, we suppose that

$$\boldsymbol{x}^\star_{c_k(\mathcal{G}^\star)} \approx \boldsymbol{0} \text{ for many } k \in \{1, \dots, \kappa(\mathcal{G}^\star)\}.$$

To evaluate the cost of edges eliminated from $\mathcal{E}_0$, we also consider weights of edges, which are represented by a function $w_0 \colon \mathcal{E}_0 \to \mathbb{R}_{++}$. Note that, if such weights are not available, we can simply set $w_0(e) = 1$ for every $e \in \mathcal{E}_0$.

3. PROPOSED GRAPH-STRUCTURED SPARSE RECOVERY METHOD

3.1. Design of Penalty Function

We focus on the observation that the graph-structured sparsity can be interpreted as the group-sparsity if we set the groups to connected components of an appropriate subgraph of the underlying graph $\mathcal{G}_0$. The group-sparsity can be effectively exploited with the mixed ℓ_2/ℓ_1 norm, i.e., the sum of the ℓ_2 norms of the grouped variables. However, since the appropriate subgraph is unknown a priori, it is not suitable to directly apply the mixed ℓ_2/ℓ_1 norm to the graph-structured sparsity. Thus, we newly introduce a penalty function as the minimum of the mixed ℓ_2/ℓ_1 norm over all candidate subgraphs of $\mathcal{G}_0$. Although the introduced penalty function is nonconvex, we derive its tight convex relaxation by using the variational representation of the ℓ_2 norm and introducing latent variables.

Concretely, we begin by defining the mixed ℓ_2/ℓ_1 norm induced by a graph $\mathcal{G}$:

$$\|\boldsymbol{x}\|^{\mathcal{G}}_{2,1} := \sum_{k=1}^{\kappa(\mathcal{G})} \sqrt{|c_k(\mathcal{G})|}\, \|\boldsymbol{x}_{c_k(\mathcal{G})}\|_2, \tag{1}$$

where $\boldsymbol{x}_{c_k(\mathcal{G})}$ is the subvector indexed by the connected component $c_k(\mathcal{G})$, and $\sqrt{|c_k(\mathcal{G})|}$ is the weight based on the suggestions in, e.g., [16–19], for group-sparse recovery. To automatically identify the appropriate subgraph, we design the penalty function $\psi^{\mathcal{G}_0}_d(\boldsymbol{x})$ by

$$\psi^{\mathcal{G}_0}_d(\boldsymbol{x}) := \min_{\mathcal{G} \in \mathcal{S}_d} \|\boldsymbol{x}\|^{\mathcal{G}}_{2,1}, \tag{2}$$

where $\mathcal{S}_d$ consists of all candidate subgraphs of $\mathcal{G}_0$ defined by eliminating edges whose total weights are less than or equal to $d \in \mathbb{R}_+$, i.e.,

$$\mathcal{S}_d := \left\{ \mathcal{G} = (\mathcal{V}_0, \mathcal{E}) \;\middle|\; \mathcal{E} \subset \mathcal{E}_0 \text{ and } \sum_{e \in \mathcal{E}_0 \setminus \mathcal{E}} w_0(e) \le d \right\}. \tag{3}$$

To derive the convex relaxation of $\psi^{\mathcal{G}_0}_d(\boldsymbol{x})$, we exploit the following lemma, from [22, Lemma 1], which shows a variational representation of the ℓ_2 norm.

Lemma 1. *Define a coercive lower semicontinuous convex function $\phi \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+ \cup \{\infty\}$ by*

$$\phi(x, \tau) := \begin{cases} \dfrac{|x|^2}{2\tau} + \dfrac{\tau}{2}, & \text{if } \tau > 0; \\ 0, & \text{if } x = 0 \text{ and } \tau = 0; \\ \infty, & \text{otherwise.} \end{cases}$$

Then, the group-wise ℓ_2 norm is variationally represented as

$$\sqrt{|\mathcal{I}|}\, \|\boldsymbol{x}_{\mathcal{I}}\|_2 = \min_{\tau \in \mathbb{R}} \sum_{n \in \mathcal{I}} \phi(x_n, \tau),$$

for any index set $\mathcal{I} \subset \{1, \dots, N\}$.

By applying Lemma 1 for each $\sqrt{|c_k(\mathcal{G})|}\, \|\boldsymbol{x}_{c_k(\mathcal{G})}\|_2$ in the definition of $\|\boldsymbol{x}\|^{\mathcal{G}}_{2,1}$ in (1), we can rewrite $\psi^{\mathcal{G}_0}_d(\boldsymbol{x})$ in (2) as

$$\psi^{\mathcal{G}_0}_d(\boldsymbol{x}) = \min_{\mathcal{G} \in \mathcal{S}_d} \left[\min_{\boldsymbol{\tau} \in \mathbb{R}^{\kappa(\mathcal{G})}} \sum_{k=1}^{\kappa(\mathcal{G})} \sum_{n \in c_k(\mathcal{G})} \phi(x_n, \tau_k) \right]. \tag{4}$$

Let us introduce a latent vector $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)^\top \in \mathbb{R}^N$ by

$$\sigma_n = \tau_k \;\; (n \in c_k(\mathcal{G})) \quad \text{for } k = 1, \dots, \kappa(\mathcal{G}).$$

Notice that $c_k(\mathcal{G})$ $(k = 1, \dots, \kappa(\mathcal{G}))$ are the connected components of $\mathcal{G}$, and the sum of the weights of edges eliminated from $\mathcal{G}_0$ is less than or equal to d as shown in (3). Thus, $\boldsymbol{\sigma}$ is characterized by the condition that

$$\|\boldsymbol{D}\boldsymbol{\sigma}\|_{0,\boldsymbol{w}} \le d,$$

where $\boldsymbol{D} \in \{-1, 0, 1\}^{|\mathcal{E}_0| \times N}$ is the difference operator on the graph $\mathcal{G}_0$, i.e.,

$$\boldsymbol{D}\boldsymbol{\sigma} = (\sigma_n - \sigma_{n'})_{(n,n') \in \mathcal{E}_0},$$

and $\boldsymbol{w} = (w(e))_{e \in \mathcal{E}_0} \in \mathbb{R}^{|\mathcal{E}_0|}_{++}$. Moreover, since the connected components form a partition of $\mathcal{V}_0$, i.e., $\bigcup_{k=1}^{\kappa(\mathcal{G})} c_k(\mathcal{G}) = \mathcal{V}_0 = \{1, \dots, N\}$ and $c_k(\mathcal{G}) \cap c_{k'}(\mathcal{G}) = \varnothing$ $(k \ne k')$, we have

$$\sum_{k=1}^{\kappa(\mathcal{G})} \sum_{n \in c_k(\mathcal{G})} \phi(x_n, \tau_k) = \sum_{n=1}^{N} \phi(x_n, \sigma_n).$$

Thus, we can further rewrite (4) to

$$\psi_d^{\mathcal{G}_0}(\boldsymbol{x}) = \min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N \\ \|\boldsymbol{D}\boldsymbol{\sigma}\|_{0,\boldsymbol{w}}\le d}} \sum_{n=1}^{N} \phi(x_n, \sigma_n).$$

By replacing the weighted ℓ_0 pseudo-norm in the constraint with its convex envelope, i.e., the weighted ℓ_1 norm, the proposed convex penalty is designed as

$$\Psi_\alpha^{\mathcal{G}_0}(\boldsymbol{x}) := \min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N \\ \|\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}\|_1\le\alpha}} \sum_{n=1}^{N} \phi(x_n, \sigma_n), \tag{5}$$

where $\alpha \in \mathbb{R}_+$ is a tuning parameter related to the sum of the weights of edges eliminated from $\mathcal{G}_0$, and $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w}) \in \mathbb{R}^{|\mathcal{E}_0|\times|\mathcal{E}_0|}$ is the diagonal matrix containing $\boldsymbol{w}$ on the main diagonal.

3.2. Optimization Algorithm

We present a graph-structured sparse recovery model using the proposed penalty (5), and develop its first-order solver. Specifically, we consider the following regularization model:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{L}\boldsymbol{x}) + \lambda\Psi_\alpha^{\mathcal{G}_0}(\boldsymbol{x}), \tag{6}$$

where $f(\boldsymbol{L}\boldsymbol{x})$ is some convex data-fidelity function with $f\colon \mathbb{R}^J \to \mathbb{R}_+$ and $\boldsymbol{L} \in \mathbb{R}^{J\times N}$, and $\lambda \in \mathbb{R}_{++}$ is the regularization parameter. We suppose that $f \in \Gamma_0(\mathbb{R}^J)$, and its proximity operator can be computed efficiently. Such examples include the square error $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2$ and the absolute error $f(\boldsymbol{u}) = \|\boldsymbol{y}-\boldsymbol{u}\|_1$, where $\boldsymbol{y}$ is the known observation vector and $\boldsymbol{L}$ is set to the known measurement matrix. We can also apply (6) to machine learning problems by using, e.g., the logistic loss $f(\boldsymbol{u}) = \sum_{j=1}^J \log(1+\exp(-y_j u_j))$ and the hinge loss $f(\boldsymbol{u}) = \sum_{j=1}^J \max\{1-y_j u_j, 0\}$, where $y_j \in \{-1,+1\}$ is a label for the j-th training data $\boldsymbol{a}_j \in \mathbb{R}^N$, and $\boldsymbol{L} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_J)^\top$.

Plugging the definition of $\Psi_\alpha^{\mathcal{G}_0}(\boldsymbol{x})$ in (5) into (6), we can solve the optimization problem (6) as

$$\underset{(\boldsymbol{x},\boldsymbol{\sigma})\in\mathbb{R}^N\times\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{L}\boldsymbol{x}) + \lambda\sum_{n=1}^{N}\phi(x_n,\sigma_n) \text{ s.t. } \|\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}\|_1 \le \alpha.$$

Applying the primal-dual method [25–30] to the reformulated problem with further slight reformulations, we obtain Algorithm 1 where $(\boldsymbol{x}^{(i)})_{i=0}^\infty$ converges to the solution of (6). Representing $\phi(x,\sigma)$ as the sum of the perspective function of $|x|^2/2$ and $\sigma/2$, from [34, Example 2.4], we can compute the proximity operator of $\gamma\lambda\phi$ in Algorithm 1 by

$$\mathrm{prox}_{\gamma\lambda\phi}(x,\sigma) = \begin{cases} (0,0), & \text{if } 2\gamma\lambda\sigma + |x|^2 \le \gamma^2\lambda^2; \\ (0, \sigma - \frac{\gamma\lambda}{2}), & \text{if } x = 0 \text{ and } 2\sigma > \gamma\lambda; \\ \left(x - \gamma\lambda s\frac{x}{|x|}, \sigma + \gamma\lambda\frac{s^2-1}{2}\right), & \text{otherwise,} \end{cases}$$

Algorithm 1: Solver for the proposed regularization model (6)

Input: $\gamma > 0$, $\mu_1 \in \left(0, \frac{1}{\sqrt{\|\boldsymbol{L}\|_{\mathrm{op}}^2+1}}\right]$, $\mu_2 \in \left(0, \frac{1}{\sqrt{\|\boldsymbol{W}\boldsymbol{D}\|_{\mathrm{op}}^2+1}}\right]$.

while *a stopping criterion is not satisfied* **do**

$\tilde{\boldsymbol{x}}^{(i+1)} = \boldsymbol{x}^{(i)} + \mu_1\boldsymbol{L}^\top(\boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i)} - \boldsymbol{u}^{(i)}))$

$\tilde{\boldsymbol{\sigma}}^{(i+1)} = \boldsymbol{\sigma}^{(i)} + \mu_2\boldsymbol{D}^\top\boldsymbol{W}(\boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)}))$

$\tilde{\boldsymbol{u}}^{(i+1)} = \boldsymbol{u}^{(i)} - \mu_1(\boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i)} - \boldsymbol{u}^{(i)}))$

$\tilde{\boldsymbol{\eta}}^{(i+1)} = \boldsymbol{\eta}^{(i)} - \mu_2(\boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)}))$

for $n = 1, \dots, N$ **do**

$(x_n^{(j+1)}, \sigma_n^{(j+1)}) = \mathrm{prox}_{\gamma\lambda\phi}(\tilde{x}_n^{(j+1)}, \tilde{\sigma}_n^{(j+1)})$

$\boldsymbol{u}^{(i+1)} = \mathrm{prox}_{\gamma f}(\tilde{\boldsymbol{u}}^{(i+1)})$

$\boldsymbol{\eta}^{(i+1)} = P_{B_1^\alpha}(\tilde{\boldsymbol{\eta}}^{(i+1)})$

$\boldsymbol{r}_1^{(i+1)} = \boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i+1)} - \boldsymbol{u}^{(i+1)})$

$\boldsymbol{r}_2^{(i+1)} = \boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}^{(i+1)} - \boldsymbol{\eta}^{(i+1)})$

where $s \in \mathbb{R}_{++}$ is the unique positive root of

$$s^3 + \left(\tfrac{2}{\gamma\lambda}\sigma + 1\right)s - \tfrac{2}{\gamma\lambda}|x| = 0,$$

and can be explicitly given via Cardano's formula as follows. Let $p = \frac{2}{\gamma\lambda}\sigma + 1$, $q = -\frac{2}{\gamma\lambda}|x|$, and $D = -\frac{q^2}{4} - \frac{p^3}{27}$. Then,

$$s = \begin{cases} \sqrt[3]{-\frac{q}{2} + \sqrt{-D}} + \sqrt[3]{-\frac{q}{2} - \sqrt{-D}}, & \text{if } D < 0; \\ 2\sqrt[3]{-\frac{q}{2}}, & \text{if } D = 0; \\ 2\sqrt[3]{\sqrt{\frac{q^2}{4} + D}}\cos\left(\frac{\arctan(-2\sqrt{D}/q)}{3}\right), & \text{if } D > 0, \end{cases}$$

where $\sqrt[3]{\cdot}$ denotes the real cubic root. The proximity operator of γf depends on the employed data-fidelity function (see, e.g., [35], for examples). The ℓ_1 ball projection $P_{B_1^\alpha}$ can be computed as

$$P_{B_1^\alpha}(\boldsymbol{\eta}) = \begin{cases} \boldsymbol{\eta}, & \text{if } \|\boldsymbol{\eta}\|_1 \le \alpha; \\ (a_n\mathrm{sign}(\eta_n))_{n=1}^{|\mathcal{E}_0|}, & \text{otherwise,} \end{cases}$$

with $a_n := \max\left\{|\eta_n| - \sum_{t=1}^T \frac{\rho_t - \alpha}{T}, 0\right\}$, where $\rho_1, \dots, \rho_{|\mathcal{E}_0|}$ are obtained by sorting $|\eta_1|, \dots, |\eta_{|\mathcal{E}_0|}|$ in descending order, and $T := \max\left\{t \in \{1, \dots, |\mathcal{E}_0|\} \;\middle|\; \sum_{n=1}^t \frac{\rho_n - \alpha}{t} < \rho_t\right\}$. Per an iteration, the proposed algorithm has linear computational complexity with respect to the signal length N and the number of edges $|\mathcal{E}_0|$ because matrix multiplications and the operators in Algorithm 1 can be computed in linear complexity of N and $|\mathcal{E}_0|$.[2]

Remark 1 (Comparison with nonconvex graph-structured sparse model)**.** Graph-structured sparse model is originally presented in [1] as a nonconvex set which has three parameters: the maximum number of nonzero entries, the maximum

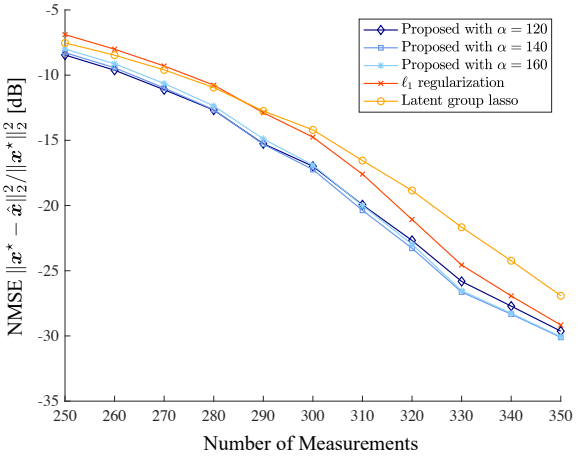
[2]The ℓ_1 ball projection can be computed in $\mathcal{O}(|\mathcal{E}_0|)$ expected complexity (see, e.g., [36]).

number of nonzero connected components, and the total cost of edges in nonzero connected components. Due to the nonconvexity, the existing methods, e.g., [1–3], rely on approximate solutions. On the other hand, we newly develop a convex regularization model (6) where the way of controlling the graph-structured sparsity is modified. In the proposed model, the total cost of eliminated edges from the underlying graph $\mathcal{G}_0$ is controlled by α. This part roughly corresponds to the total cost of edges in the existing framework [1], if we set the weights of edges of $\mathcal{G}_0$ inversely proportional to those in the existing framework. The maximum number of nonzero entries and the maximum number of nonzero connected components are jointly controlled by the regularization parameter λ because $\Psi_{\alpha}^{\mathcal{G}_0}(\boldsymbol{x})$ penalizes larger connected components more heavily (see (1)).

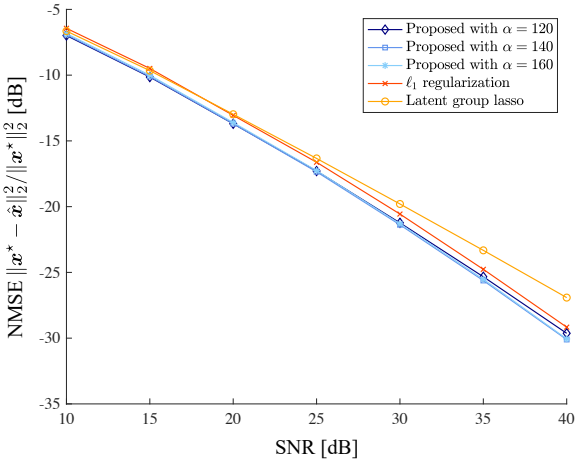
4. NUMERICAL EXPERIMENTS

We consider the estimation of a graph-structured sparse signal $\boldsymbol{x}^\star \in \mathbb{R}^N$ from noisy compressive measurements. The underlying graph $\mathcal{G}_0$ is randomly generated as a *connected 3-regular graph* [24]. Then, 2 subgraphs of $\mathcal{G}_0$ are randomly chosen for 150 nonzero entries of $\boldsymbol{x}^\star$, where N is set to 500. Amplitudes of the nonzero entries are drawn from i.i.d. Gaussian distribution $\mathcal{N}(0,1)$. The selection scheme of subgraphs are as follows. First, the numbers of nodes of subgraphs are randomly chosen from $\{1,2,\ldots,149\}$ under the constraint that their sum equals to 150. Each subgraph is initialized with a node randomly chosen from $\mathcal{G}_0$. Then, adjacent nodes of a randomly chosen node are added into the subgraph until the number of nodes becomes the prescribed number. If some nodes are shared by the subgraphs chosen by this process, we reject these subgraphs, and repeat the process until the prescribed condition $\mathrm{supp}(\boldsymbol{x}^\star) = 150$ is met. For simplicity, we set the edge weights as $w_0(e) = 1$ for every $e \in \mathcal{E}_0$. The measurements are generated by $\boldsymbol{y} := \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{\varepsilon}$, where the entries of $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ $(d < N)$ are drawn from i.i.d. $\mathcal{N}(0,1)$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is the white Gaussian noise.

We compare the proposed regularization model (6) with the existing convex regularizations: ℓ_1 regularization and latent group lasso (LGL) [21] which is an advanced convex group-sparse regularization model. Since the LGL needs predefined groups, we set the groups to adjacent nodes of $\mathcal{G}_0$. In the proposed model (6), we use the square error by setting $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2$ and $\boldsymbol{L} = \boldsymbol{A}$. The existing penalties are also combined with the square error in similar ways. Note that the regularization parameter is tuned independently for each model to obtain the best result. Figs. 1(a) and (b) show the normalized mean square error (NMSE) $\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}\|_2^2/\|\boldsymbol{x}^\star\|_2^2$ respectively against the number d of measurements and the SNR $E[\|\boldsymbol{A}\boldsymbol{x}^\star\|_2^2]/E[\|\boldsymbol{\varepsilon}\|_2^2]$, where $\hat{\boldsymbol{x}}$ is the solution of each model and the results are averaged over 100 independent trials. LGL shows better performances than the ℓ_1 regularization for small number of measurements and low SNR, but not for large number of measurements and high SNR. This is hypothesized because the estimation bias is caused by the mismatch between the ideal groups and the predefined groups consisting of adjacent nodes. On the other hand, the proposed model for several choices of α achieves better estimation accuracy than the existing models for all conditions in Fig. 1.



(a) NMSE against the number of measurements where SNR is fixed to 40dB.



(b) NMSE against SNR where the number of measurements is fixed to 350.

Fig. 1: Comparison of the regularization models where the results are averaged over 100 independent trials.

5. CONCLUSION

We presented a novel convex framework for the graph-structured sparse recovery. More precisely, we proposed a convex penalty function (5) which can identify the relevant subgraph of an underlying graph. For the proposed graph-structured recovery model (6), we developed Algorithm 1 which is a first-order iterative algorithm guaranteed to converge to an optimal solution. Numerical experiments illustrate the efficacy of the proposed method compared with the existing convex regularizations based on fixed structures.

Acknowledgment: This work was supported in part by JSPS Grants-in-Aid 21K17827 and 19K20361.

6. REFERENCES

[1] C. Hegde, P. Indyk, and L. Schmidt, "A nearly-linear time framework for graph-structured sparsity," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2015, pp. 928–937.

[2] F. Chen and B. Zhou, "A generalized matching pursuit approach for graph-structured sparsity," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 1389–1395.

[3] B. Zhou, F. Chen, and Y. Ying, "Stochastic iterative hard thresholding for graph-structured sparsity optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, vol. 97, pp. 7563–7573.

[4] J. Peng, L. Zhu, Y. Wang, and J. Chen, "Mining relationships among multiple entities in biological networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 3, pp. 769–776, 2020.

[5] J. Cadena, F. Chen, and A. Vullikanti, "Graph anomaly detection based on Steiner connectivity and density," *Proc. IEEE*, vol. 106, no. 5, pp. 829–845, 2018.

[6] B. Zhou and F. Chen, "Graph-structured sparse optimization for connected subgraph detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2016, pp. 709–718.

[7] F. Chen, B. Zhou, A. Alim, and L. Zhao, "A generic framework for interesting subspace cluster detection in multi-attributed networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2017, pp. 41–50.

[8] F. Jie, C. Wang, F. Chen, L. Li, and X. Wu, "Block-structured optimization for anomalous pattern detection in interdependent networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2019, pp. 1138–1143.

[9] D. De Oliveira, D. Neill, J. Garrett, and L. Soibelman, "Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network," *J. Comput. Civ. Eng.*, vol. 25, no. 1, pp. 21–30, Jan. 2011.

[10] B. Anbaroğlu, T. Cheng, and B. Heydecker, "Non-recurrent traffic congestion detection on heterogeneous urban road networks," *Transportmetrica A: Transport Science*, vol. 11, no. 9, pp. 754–771, 2015.

[11] S. Speakman, E. McFowland, and D. B. Neill, "Scalable detection of anomalous patterns with connectivity constraints," *Journal of Computational and Graphical Statistics*, vol. 24, no. 4, pp. 1014–1033, 2015.

[12] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[13] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.

[14] X. Lv, G. Bi, and C. Wan, "The group lasso for stable recovery of block-sparse signal representations," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, Apr. 2011.

[15] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.

[16] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[17] Y. Nardi and A. Rinaldo, "On the asymptotic properties of the group lasso estimator for linear models," *Electron. J. Statist.*, vol. 2, pp. 605–633, 2008.

[18] F. Wei and J. Huang, "Consistent group selection in high-dimensional linear regression," *Bernoulli*, vol. 16, no. 4, pp. 1369–1384, 2010.

[19] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statist. Sci.*, vol. 27, no. 4, pp. 481–499, Nov. 2012.

[20] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.

[21] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2009, pp. 433–440.

[22] H. Kuroda and D. Kitahara, "Block-sparse recovery with optimal block partition," preprint TechRxiv:14538552, Nov. 2021.

[23] H. Kuroda, D. Kitahara, and A. Hirabayashi, "A convex penalty for block-sparse signals with unknown structures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2021, pp. 5430–5434.

[24] R. Diestel, *Graph Theory*, Springer, 5th edition, 2017.

[25] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[26] E. Esser, X. Zhang, and T. F. Chan, "A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science," *SIAM J. Imaging Sci.*, vol. 3, no. 4, pp. 1015–1046, 2010.

[27] I. Loris and C. Verhoeven, "On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty," *Inverse Problems*, vol. 27, no. 12, Nov. 2011, 15 pages.

[28] P. Chen, J. Huang, and X. Zhang, "A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration," *Inverse Problems*, vol. 29, no. 2, Jan. 2013, 33 pages.

[29] M. Yamagishi and I. Yamada, "Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization," *Inverse Problems*, vol. 33, no. 4, Mar. 2017, 35 pages.

[30] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Math. Comput.*, vol. 82, no. 281, pp. 301–329, 2013.

[31] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, "Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists," preprint arXiv:1912.00137, 2021.

[32] P. L. Combettes, "Perspective functions: Properties, constructions, and examples," *Set-Valued and Variational Analysis*, vol. 26, no. 2, pp. 247–264, 2018.

[33] P. L. Combettes and C. L. Müller, "Perspective functions: Proximal calculus and applications in high-dimensional statistics," *J. Math. Anal. Appl.*, vol. 457, no. 2, pp. 1283–1306, 2018.

[34] P. L. Combettes and C. L. Müller, "Perspective maximum likelihood-type estimation via proximal decomposition," *Electron. J. Statist.*, vol. 14, no. 1, pp. 207–238, 2020.

[35] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds., pp. 185–212. Springer, 2011.

[36] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Math. Programm.*, vol. 158, no. 1, pp. 575–585, 2016.