AN EFFICIENT ALTERNATING DIRECTION METHOD FOR GRAPH LEARNING FROM SMOOTH SIGNALS

Xiaolu Wang[*], *Chaorui Yao*[†], *Haoyu Lei*[†], *Anthony Man-Cho So*[*]

[*]Department of Sys. Eng. & Eng. Mgmt, The Chinese University of Hong Kong, HKSAR
[†]Department of Information Engineering, The Chinese University of Hong Kong, HKSAR

ABSTRACT

We consider the problem of identifying the graph topology from a set of smooth graph signals. A well-known approach to this problem is minimizing the Dirichlet energy accompanied with some Frobenius norm regularization. Recent works have incorporated the logarithmic barrier on the node degrees to improve the overall graph connectivity without compromising graph sparsity, which is shown to be quite effective in enhancing the quality of the learned graphs. Although a primal-dual algorithm has been proposed in the literature to solve this type of graph learning formulations, it lacks a rigorous convergence analysis and appears to have a slow empirical performance. In this paper, we cast the graph learning formulation as a nonsmooth, strictly convex optimization problem and develop an efficient alternating direction method of multipliers to solve it. We show that our algorithm converges to the global minimum with arbitrary initialization. We conduct extensive experiments on various synthetic and real-world graphs, the results of which show that our method exhibits sharp linear convergence and is substantially faster than the commonly adopted primal-dual method.

Index Terms— Graph Learning, Graph Signal Processing, ADMM, Optimization Algorithms

1. INTRODUCTION

Graphs play a central role in characterizing the structural information of data. Various types of data from real-world applications, including social networks, brain signal analysis, urban traffic flows, etc., can be regarded as signals that reside on graphs [1, 2]. The edge weights of the graph capture the inter-node relationships. Numerous algorithms in signal processing and machine learning have been developed to cope with graph-structured data [3]. Nevertheless, in many scenarios, the concrete graph connectivities and edge weights are not known a priori. This hinders further representation, processing, and analysis of the graph data. In some other applications, such as brain networks [4], the graph structure itself is the sought information. Therefore, it is crucial to infer the graph from a given set of observed signals.

A graph signal is usually represented by a common s-dimensional vector while its entries are closely related to the hidden graph. In order to discover the graph topology from the observed graph signals, it is often assumed that the signals vary across the graph smoothly [5]. This is a reasonable abstraction of many real-world graph-supported signals, which means that when two nodes are connected with a large edge weight, their corresponding node values tend to be close to each other. Various formulations of graph learning from smooth signals have been proposed [5–9]. Among them, the model proposed in [6], which combines the smoothness-promoting Dirichlet energy with a logarithmic barrier on the node degrees, exhibits superior performance in learning high-quality graphs. This convex formulation improves the overall graph connectivity without involving any spectral properties of the Laplacian matrix. Thus, it is possible to obtain the globally optimal solution by algorithms with low per-iteration computational cost.

A type of primal-dual method is generally adopted to solve the graph learning formulation in [7] and other related variants [10, 11]. However, the convergence of the adopted primal-dual algorithm has not been rigorously established. More critically, the iterative procedures do not guarantee the node degrees to lie in the domain of the logarithm barrier in the formulation. In practice, we observe that the primal-dual algorithm appears to converge rather slowly in many cases. Hence, there is a strong motivation for developing an efficient algorithm with provable convergence guarantee to solve this concise yet effective graph learning formulation.

In this paper, we cast the graph learning problem as a nonsmooth, strictly convex optimization problem with an equality constraint. We develop an efficient linearized alternating direction method of multipliers (ADMM) [12] to solve the problem, which can be much faster than the traditional primal-dual algorithms. Our theoretical analysis shows that the proposed ADMM converges to the global minimum from an arbitrary initial point. Sharp linear convergence of the ADMM is observed from our experiments. The convergence and runtime comparisons indicate that our method is significantly superior to the existing primal-dual method.

2. PROBLEM FORMULATION

Let $x \in \mathbb{R}^s$ be a graph signal, whose relations between different entries are characterized by an undirected graph $\mathcal{G} =$

$\langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ denotes the set of nodes with $|\mathcal{V}| = s$ and $\mathcal{E}$ denotes the set of edges. Let $W \in \mathbb{R}^{s\times s}$ denote the weight matrix of $\mathcal{G}$, where $W_{ij} \geq 0$ represents the weight of the edge $(i,j) \in \mathcal{E}$. Suppose that we have observed n (possibly noisy) graph signals $x_1, \ldots, x_n \in \mathbb{R}^s$ that reside on $\mathcal{G}$. We denote the data matrix by $X \coloneqq [x_1, \ldots, x_n] = [\tilde{x}_1^\top, \ldots, \tilde{x}_s^\top]^\top \in \mathbb{R}^{s\times n}$. We want to recover the underlying graph $\mathcal{G}$, or equivalently, the weight matrix W, from the given set of graph signals X. For this goal, it is generally assumed that the graph signal varies smoothly across the graph. More precisely, the following Dirichlet energy is used to measure the smoothness of the set of graph signals X:

$$\sum_{i=1}^{s}\sum_{j=1}^{s} W_{ij}\,\|\tilde{x}_i - \tilde{x}_j\|_2^2 = \|W \circ D\|_{1,1},$$

where $D_{ij} = \|\tilde{x}_i - \tilde{x}_j\|_2^2$ is the squared pairwise distance of node vectors $\tilde{x}_i$ and $\tilde{x}_j$, $\circ$ denotes the Hadamard product, and $\|\cdot\|_{1,1}$ is the element-wise ℓ_1 norm. Based on the smoothness assumption, the following formulation originally proposed in [7] has been widely used for learning the weight matrix of the graph:

$$\begin{aligned} \min_{W\in\mathbb{R}^{s\times s}} \;& \|W\circ D\|_{1,1} - \alpha \mathbf{1}^\top \log(W\mathbf{1}) + \frac{\beta}{2}\|W\|_F^2 \\ \text{s.t.} \;& \quad W \geq \mathbf{0},\; W = W^\top,\; \text{diag}(W) = \mathbf{0}. \end{aligned} \tag{1}$$

Since problem (1) does not involve any spectral structure of the matrix variables, it is convenient to convert it into the vector form:

$$\begin{aligned} \min_{w\in\mathbb{R}^m} \;& 2b^\top w - \alpha \mathbf{1}^\top \log(Qw) + \beta\|w\|_2^2 \\ \text{s.t.} \;& \quad w \geq 0, \end{aligned} \tag{2}$$

where $m = s(s-1)/2$, b (resp. w) is the vector that stacks all entries above the main diagonal of D (resp. W), and $Q \in \{0,1\}^{s\times m}$ is a sparse binary matrix that satisfies $Qw = W\mathbf{1}$.

Instead of using the primal-dual algorithm proposed in [7] to solve (2), we resort to the more efficient ADMM. Let $Qw = v$. By penalizing the non-negative constraint in (2), we obtain the following reformulation, which falls into a standard form that ADMM can deal with:

$$\begin{aligned} \min_{w\in\mathbb{R}^m, v\in\mathbb{R}^s} \;& f(w) + g(v) \\ \text{s.t.} \;& \quad Qw - v = \mathbf{0}, \end{aligned} \tag{3}$$

where $f(w) = 2b^\top w + \beta\|w\|_2^2 + \mathbb{I}_{\{w\geq 0\}}$ with $\mathbb{I}_{\{w\geq 0\}} = \begin{cases} 0, & w \geq \mathbf{0} \\ +\infty, & \text{otherwise} \end{cases}$, and $g(v) = -\alpha\mathbf{1}^\top \log(v)$. Note that $w \mapsto f(w)$ is strongly convex in w but nonsmooth due to the indicator function. Besides, $v \mapsto g(v)$ is strictly convex in v, since $\nabla^2 g(v) = \text{Diag}\left(\frac{1}{v_1^2}, \ldots, \frac{1}{v_m^2}\right) \succ \mathbf{0}$ for all $v > \mathbf{0}$. In summary, problem (3) is a nonsmooth, strictly convex optimization problem with a linear equality constraint. It can be shown that the problem has a unique global minimizer.

3. OPTIMIZATION ALGORITHM

In this section, we develop the detailed procedures of the linearized ADMM for solving problem (3). We first provide the following two propositions as a preparation.

Proposition 1. *If $f(w) = 2b^\top w + \beta\|w\|_2^2 + \mathbb{I}_{\{w\geq 0\}}$, then for $\tau_1 > 0$, the closed-form proximal mapping of f is given by*

$$\text{prox}_{\tau_1 f}(w) = \max\left\{\frac{1}{2\tau_1\beta + 1}w - \frac{2\tau_1}{2\tau_1\beta+1}b, \mathbf{0}\right\}.$$

Proposition 2. *If $g(v) = -\alpha\mathbf{1}^\top \log(v)$, then for $\tau_2 > 0$, the closed-form proximal mapping of g is given by*

$$\text{prox}_{\tau_2 g}(v) = \frac{v + \sqrt{v^2 + 4\alpha\tau_2\mathbf{1}}}{2},$$

where the square and the square root are both taken element-wise.

Introducing the dual variable $\lambda \in \mathbb{R}^s$ for the equality constraint in (3), we have the augmented Lagrangian function with penalty parameter $t > 0$ as follows:

$$\begin{aligned} \mathcal{L}_t(w, v; \lambda) =& f(w) + g(v) - \langle \lambda, Qw - v\rangle + \frac{t}{2}\|Qw - v\|_2^2 \\ =& 2b^\top w + \beta\|w\|_2^2 + \mathbb{I}_{\{w\geq 0\}} - \alpha\mathbf{1}^\top \log(v) \\ & - \langle \lambda, Qw - v\rangle + \frac{t}{2}\|Qw - v\|_2^2. \end{aligned}$$

Fixing v^k and λ^k in the k-th iteration, the subproblem for w is

$$\begin{aligned} & \min_w \; f(w) - \left\langle Q^\top\lambda^k, w\right\rangle + \frac{t}{2}\left\|Qw - v^k\right\|_2^2 \\ \Leftrightarrow & \min_w \; f(w) + \frac{t}{2}\left\|Qw - v^k - \frac{\lambda^k}{t}\right\|_2^2. \end{aligned} \tag{4}$$

In view of Proposition 1, we linearize the quadratic term in (4) to perform one step of the proximal gradient iteration to update w:

$$\begin{aligned} w^{k+1} =& \text{prox}_{\tau_1 f}\left(w^k - \tau_1 Q^\top\left(Qw^k - v^k - \frac{\lambda^k}{t}\right)\right) \\ =& \max\left\{\tilde{w}^{k+1}, \mathbf{0}\right\}, \end{aligned} \tag{5}$$

where

$$\tilde{w}^{k+1} = \frac{w^k - \tau_1 Q^\top\left(Qw^k - v^k - \frac{\lambda^k}{t}\right) - 2\tau_1 b}{2\tau_1\beta + 1}.$$

Fixing λ^k and the newly updated w^{k+1}, the subproblem for v is

$$\begin{aligned} & \min_v \; g(v) + \left\langle \lambda^k, v\right\rangle + \frac{t}{2}\left\|Qw^{k+1} - v\right\|_2^2 \\ \Leftrightarrow & \min_v \; g(v) + \frac{t}{2}\left\|Qw^{k+1} - v - \frac{\lambda^k}{t}\right\|_2^2. \end{aligned} \tag{6}$$

Algorithm 1 ADMM for Graph Learning

Input: penalty parameter t, step sizes τ_1 and τ_2, primal residual tolerance ε_p, dual residual tolerance ε_d
Initialize: $k = 0$, randomly pick w^0, v^0, and λ^0, and pick sufficiently large r_p, r_d
while $r_p > \varepsilon_p$ or $r_d > \varepsilon_d$ **do**
  update w according to (5)
  update v according to (7)
  update λ according to (8)
  set primal residual $r_p = \|tQ^\top \left(v^{k+1} - v^k\right)\|_2$
  set dual residual $r_d = \|Qw^k - v^k\|_2$
  $k \leftarrow k + 1$
end while

Now, in view of Proposition 2, we perform a similar proximal gradient iteration to update v in (6):

$$\begin{aligned} v^{k+1} =& \operatorname{prox}_{\tau_2 g}\left(v^k + \tau_2\left(Qw^{k+1} - v^k - \frac{\lambda^k}{t}\right)\right) \\ =& \frac{\tilde{v}^{k+1} + \sqrt{\left(\tilde{v}^{k+1}\right)^2 + 4\alpha\tau_2 \mathbf{1}}}{2}, \end{aligned} \tag{7}$$

where $\tilde{v}^{k+1} = (1 - \tau_2 t)v^k + \tau_2 t Q w^{k+1} - \tau_2 \lambda^k / t$. Subsequently, the dual variable λ is updated as

$$\lambda^{k+1} = \lambda^k - t\left(Qw^{k+1} - v^{k+1}\right). \tag{8}$$

The overall description of our linearized ADMM is presented in Algorithm 1. The stopping criterion is that the primal and dual residuals attain certain pre-specified tolerances. The per-iteration computational cost of our linearized ADMM is $O(m)$, which is comparable to that of the primal-dual method in [7].

4. CONVERGENCE ANALYSIS

Suppose that $\alpha, \beta > 0$. As discussed in Section 2, problem (3) has a unique globally optimal solution (w^*, v^*). In this section, we establish the global iterate convergence of Algorithm 1 following the techniques in [13]. To proceed, we first state the following lemma.

Lemma 1. *Suppose that (w^*, v^*) is the optimal solution of problem (3) and λ^* is the corresponding optimal dual variable. If the step sizes satisfy $\tau_1 < \frac{1}{\sigma_{\max}(Q)}$ and $\tau_2 < 1$, then there exists*

$$c = \min\left\{\frac{t}{\tau_1} - t\sigma_{\max}(Q), \frac{t}{\tau_2} - \gamma, \frac{1}{t} - \frac{1}{\gamma}\right\} > 0$$

such that the sequence $\left\{\left(w^k, v^k, \lambda^k\right)\right\}_{k=0}^{\infty}$ generated by Algorithm 1 satisfies

$$\|z^k - z^*\|_M^2 - \|z^{k+1} - z^*\|_M^2 \geq c\|z^k - z^{k+1}\|_2^2,$$

where $\sigma_{\max}(\cdot)$ denotes the largest singular value, $z^k = \begin{pmatrix} w^k \\ v^k \\ \lambda^k \end{pmatrix}$, $z^* = \begin{pmatrix} w^* \\ v^* \\ \lambda^* \end{pmatrix}$, $M = \begin{pmatrix} \frac{1}{\tau_1}\mathbf{I} - tQ^\top Q & & \\ & \frac{1}{\tau_2}\mathbf{I} & \\ & & \frac{1}{t}\mathbf{I} \end{pmatrix}$, and $\|u\|_M = \sqrt{u^\top M u}$ for $u \in \mathbb{R}^{m+2s}$.

Lemma 1 indicates that the distances of the iterates to the optimal solution decrease monotonically, and it further implies that the iterates lie in a bounded set. Equipped with Lemma 1, we can obtain the following theorem, which reveals the global convergence of Algorithm 1.

Theorem 1. *Suppose that the step sizes satisfy $\tau_1 < \frac{1}{\sigma_{\max}(Q)}$ and $\tau_2 < 1$. Then, the sequence $\left\{z^k = \left(w^k, v^k, \lambda^k\right)\right\}_{k=0}^{\infty}$ generated by Algorithm 1 converges for all $t > 0$, and the limit point is optimal for problem (3).*

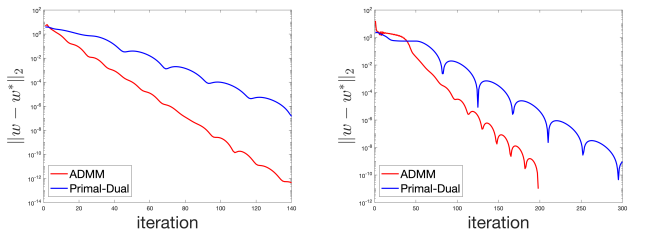
Due to space limitation, the proofs of all the aforementioned results are deferred to the full version of this paper.

5. NUMERICAL RESULTS

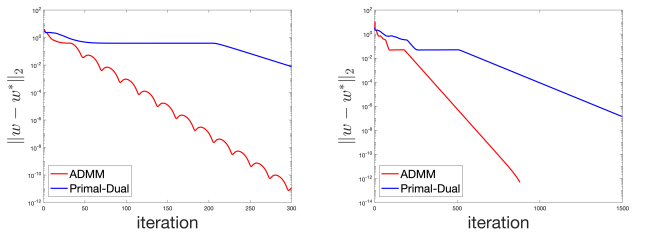
In this section, we present the numerical performance of our linearized ADMM and compare it with that of the primal-dual method [7]. All algorithms are implemented in MATLAB. In particular, we test the primal-dual algorithm based on the Graph Signal Processing toolbox [14] and use the scaling trick given in [9, Proposition 1] to accelerate the convergence. All reported results are based on the best-tuned α and β so that the learned graphs have the highest quality in terms of the F-measure [5, 15]. Moreover, the parameters t, τ_1, τ_2 in the ADMM, and the step sizes in the primal-dual algorithm, are also best-tuned to achieve the best possible convergence results. We follow [5] to generate the graph signals. Suppose that the Laplacian matrix of the ground-truth graph is $L = \operatorname{Diag}(W\mathbf{1}) - W$ and admits the eigen-decomposition $L = \chi\Lambda\chi^\top$. Then, the graph signal is generated as $x = \chi h + \delta$, where $\delta \sim \mathcal{N}(\mathbf{0}, \epsilon\mathbf{I})$ is the Gaussian noise with noise level ϵ and $h \sim \mathcal{N}\left(\mathbf{0}, \Lambda^\dagger\right)$, where $\Lambda^\dagger$ is the pseudo-inverse of Λ.

5.1. Experiments on Synthetic Graphs

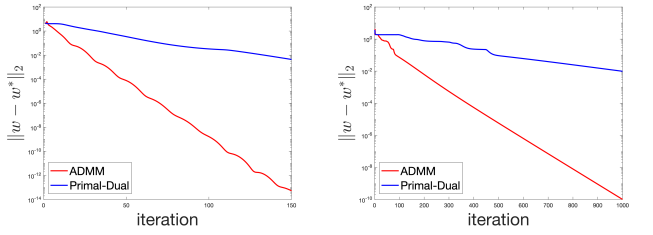
We first carry out experiments on three types of synthetic graphs, namely, Gaussian graph, ER graph, and PA graph. For the Gaussian graph, the coordinates of the nodes are sampled uniformly from the unit square, and the edge weights are determined by the radial basis function $\exp\left(-d(i,j)^2/2\rho^2\right)$, where $d(i,j)$ is the Euclidean distance between node i and node j and $\rho = 0.5$ is the kernel width parameter. All edges whose weights are smaller than 0.75 are removed. The ER graph is generated according to the Erdős-Rényi (ER) model [16], where each possible edge is independently added to the graph with probability 0.2. The PA graph is generated according to the preferential attachment (PA) model [17], where



(a) Gaussian ($s = 20$, $n = 100$) (b) Gaussian ($s = 50$, $n = 400$)



(c) ER ($s = 20$, $n = 100$) (d) ER ($s = 50$, $n = 400$)



(e) PA ($s = 20$, $n = 100$) (f) PA ($s = 50$, $n = 400$)

Fig. 1: Convergence performance on synthetic graphs

one new node is added to the graph at a time and connected to an existing node. We generate different sets of graph signals with the same noise level $\epsilon = 0.5$.

Figure 1 illustrates the suboptimality gap $\|w^k - w^*\|_2$ in logarithmic scale against the number of iterations k of the two algorithms with different s and n. It is observed that the ADMM always exhibits notably sharper linear convergence rates than the primal-dual algorithm. For some cases, e.g., Figure 1c, the primal-dual algorithm converges rather slowly, while the ADMM still performs quite well.

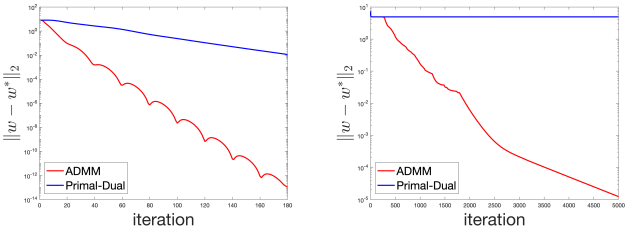
We also compare the runtime of the ADMM and the primal-dual algorithm by stopping them when the residuals are less than 10^{-10}. Since problem (1) is convex, we input it into the convex optimization package CVX [18] by using the default SDPT3 solver with the "highest" precision (provided by CVX). The runtime of CVX is provided as a baseline. The runtime comparison is reported in Table 1. In all cases, the ADMM consumes considerably less time than the primal-dual algorithm to achieve the common precision.

5.2. Experiments on Real-World Graphs

We also test the numerical performance on several real-world graphs from the *SuiteSparse Matrix Collection* [19]. In particular, we select the bcspwr graph with $s = 118$ from power network problems and the lshp graph with $s = 1561$ from thermal problems. We generate 100 and 2000 graph signals for them, respectively.

Graph	s	CVX	Primal-Dual	ADMM
Gaussian	20	1.94	0.04	0.01
	50	13.00	0.06	0.03
ER	20	3.57	0.23	0.02
	50	12.42	0.44	0.04
PA	20	2.03	0.07	0.004
	50	11.98	0.63	0.06

Table 1: Comparison of runtime (in seconds)



(a) bcspwr ($s = 118$, $n = 100$) (b) lshp ($s = 1561$, $n = 2000$)

Fig. 2: Convergence performance on real-world graphs

Figure 2 shows the numerical results. For these sparse real-world graphs, the ADMM still converges much faster than the primal-dual algorithm. Especially for the relatively large lshp graph, the primal-dual algorithm can hardly obtain a desirable suboptimal solution, while ADMM can still achieve 10^{-5} precision within thousands of iterations.

We do not provide any runtime comparison in this subsection, since learning graphs beyond medium scale by CVX or the primal-dual algorithm takes too long.

6. CONCLUSION AND FUTURE WORK

In this paper, we developed an efficient linearized ADMM algorithm to solve a popular graph learning formulation. The global convergence is guaranteed in theory, and superb performance is verified by numerical experiments. We note that the objective function in (3) has no Lipschitz gradient due to the logarithmic term, which brings difficulty in analyzing the convergence rate of Algorithm 1. Nevertheless, our numerical experiments suggest that Algorithm 1 may enjoy linear convergence. Besides, it is interesting to extend the algorithmic framework in this paper to a wider range of graph learning scenarios, such as learning time-varying graphs [10, 11]. We leave these theoretical and algorithmic issues as future work.

7. REFERENCES

[1] Xiaowen Dong, Dorina Thanou, Michael Rabbat, and Pascal Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 44–63, 2019.

[2] Gonzalo Mateos, Santiago Segarra, Antonio G Marques, and Alejandro Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 16–43, 2019.

[3] David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[4] Chenhui Hu, Lin Cheng, Jorge Sepulcre, Georges El Fakhri, Yue M Lu, and Quanzheng Li, "A graph theoretical regression model for brain connectivity learning of Alzheimer's disease," in *IEEE International Symposium on Biomedical Imaging*, 2013, pp. 616–619.

[5] Xiaowen Dong, Dorina Thanou, Pascal Frossard, and Pierre Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Transactions on Signal Processing*, vol. 64, no. 23, pp. 6160–6173, 2016.

[6] Chenhui Hu, Lin Cheng, Jorge Sepulcre, Keith A Johnson, Georges E Fakhri, Yue M Lu, and Quanzheng Li, "A spectral graph regression model for learning brain connectivity of Alzheimer's disease," *PLoS One*, vol. 10, no. 5, pp. e0128136, 2015.

[7] Vassilis Kalofolias, "How to learn a graph from smooth signals," in *Artificial Intelligence and Statistics (AISTATS)*, 2016, pp. 920–929.

[8] Hilmi E. Egilmez, Eduardo Pavez, and Antonio Ortega, "Graph learning from data under Laplacian and structural constraints," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 6, pp. 825–841, 2017.

[9] Vassilis Kalofolias and Nathanaël Perraudin, "Large scale graph learning from smooth signals," in *International Conference on Learning Representations (ICLR)*, 2019.

[10] Vassilis Kalofolias, Andreas Loukas, Dorina Thanou, and Pascal Frossard, "Learning time varying graphs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2826–2830.

[11] Koki Yamada, Yuichi Tanaka, and Antonio Ortega, "Time-varying graph learning with constraints on graph temporal variation," *arXiv preprint arXiv:2001.03346*, 2020.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] Shiqian Ma, "Alternating proximal gradient method for convex minimization," *Journal of Scientific Computing*, vol. 68, no. 2, pp. 546–572, 2016.

[14] Nathanaël Perraudin, Johan Paratte, David Shuman, Lionel Martin, Vassilis Kalofolias, Pierre Vandergheynst, and David K Hammond, "GSPbox: A toolbox for signal processing on graphs," *arXiv preprint arXiv:1408.5781*, 2014.

[15] Christopher D. Manning, Hinrich Schütze, and Prabhakar Raghavan, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[16] Paul Erdős and Alfréd Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[17] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[18] Michael Grant and Stephen Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014.

[19] Timothy A. Davis and Yifan Hu, "The University of Florida sparse matrix collection," *ACM Transactions on Mathematical Software*, vol. 38, no. 1, pp. 1–25, 2011.

A. SUPPLEMENTARY PROOFS

A.1. Proof of Proposition 1

Proof. Note that $f(w)$ is separable in each dimension of w, i.e., $f(w) = \sum_{i=1}^{m} f_i(w_i)$ where $f_i(w_i) = 2b_i w_i + \beta w_i^2 + \mathbb{I}_{\{w_i \geq 0\}}$. Then, for $i = 1, \ldots, m$,

$$\begin{aligned}
&\text{prox}_{\tau_1 f_i}(w_i) \\
=&\,\text{argmin}_{v \in \mathbb{R}} \left\{ f_i(v) + \frac{1}{2\tau_1}(v - w_i)^2 \right\} \\
=&\,\text{argmin}_{v \in \mathbb{R}} \left\{ 2b_i v + \beta v^2 + \mathbb{I}_{\{v \geq 0\}} + \frac{1}{2\tau_1}(v - w_i)^2 \right\} \\
=&\,\max\left\{ \frac{w_i - 2\tau_1 b_i}{2\tau_1 \beta + 1}, \mathbf{0} \right\},
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\text{prox}_{\tau_1 f}(w) =&\,\text{argmin}_v \left\{ \tau_1 f(v) + \frac{1}{2}\|v - w\|_2^2 \right\} \\
=&\,\max\left\{ \frac{1}{2\tau_1 \beta + 1} w - \frac{2\tau_1}{2\tau_1 \beta + 1} b, \mathbf{0} \right\}.
\end{aligned}$$

□

A.2. Proof of Proposition 2

Proof. Note that $g(v)$ is separable in each dimension of v, i.e., $g(v) = \sum_{j=1}^{s} g_j(v_j)$ where $g_j(v_j) = -\alpha \log(v_j)$. Then, for $j = 1, \ldots, s$,

$$\begin{aligned}
&\text{prox}_{\tau_2 g_j}(v_j) \\
=&\,\text{argmin}_{\nu \in \mathbb{R}} \left\{ g_j(\nu) + \frac{1}{2\tau_2}(\nu - v_j)^2 \right\} \\
=&\,\text{argmin}_{\nu \in \mathbb{R}} \left\{ -\alpha \log(v_j) + \frac{1}{2\tau_2}(\nu - v_j)^2 \right\} \\
=&\,\frac{v_j + \sqrt{v_j^2 + 4\alpha\tau_2}}{2}.
\end{aligned}$$

Combining all dimensions gives

$$\text{prox}_{\tau_2 g}(v) = \frac{v + \sqrt{v^2 + 4\alpha\tau_2 \mathbf{1}}}{2}.$$

□

A.3. Proof of Lemma 1

Proof. The optimal solution (w^*, v^*, λ^*) satisfies the following KKT conditions

$$\mathbf{0} \in \partial f(w^*) - Q^\top \lambda^*, \tag{9}$$

$$\mathbf{0} = \nabla g(v^*) + \lambda^*, \tag{10}$$

$$\mathbf{0} = Qw^* - v^*. \tag{11}$$

The optimality condition for the subproblem (5) is

$$\mathbf{0} \in \tau_1 \partial f(w^{k+1}) + w^{k+1} - w^k + \tau_1 Q^\top \left(Qw^k - v^k - \frac{\lambda^k}{t} \right). \tag{12}$$

Plugging (8) into (12) gives

$$\begin{aligned}
&\mathbf{0} \in \tau_1 \partial f(w^{k+1}) + w^{k+1} - w^k + \tau_1 Q^\top \left(Qw^k - v^k - \frac{\lambda^{k+1}}{t} - Qw^{k+1} + v^{k+1} \right) \\
&= \tau_1 \partial f(w^{k+1}) + w^{k+1} - w^k + \tau_1 Q^\top Q(w^k - w^{k+1}) - \tau_1 Q^\top (v^k - v^{k+1}) - \frac{\tau_1}{t} Q^\top \lambda^{k+1} \\
\Leftrightarrow\ & \frac{1}{\tau_1} \left(w^k - w^{k+1} \right) - Q^\top Q(w^k - w^{k+1}) + Q^\top (v^k - v^{k+1}) + \frac{1}{t} Q^\top \lambda^{k+1} \in \partial f(w^{k+1}).
\end{aligned} \tag{13}$$

Since ∂f is a monotone operator, combining (13) with $Q^\top \lambda^* \in \partial f(w^*)$ by (9) yields

$$\left\langle \frac{1}{\tau_1} \left(w^k - w^{k+1} \right) - Q^\top Q \left(w^k - w^{k+1} \right) + Q^\top \left(v^k - v^{k+1} \right) + \frac{1}{t} Q^\top \left(\lambda^{k+1} - \lambda^* \right), w^{k+1} - w^* \right\rangle \geq 0. \tag{14}$$

The optimality condition for the subproblem (7) is

$$\mathbf{0} = \tau_2 \nabla g(v^{k+1}) + v^{k+1} - v^k - \tau_2 \left(Qw^{k+1} - v^k - \frac{\lambda^k}{t} \right). \tag{15}$$

Plugging (8) into (15) gives

$$\begin{aligned}
&\mathbf{0} = \tau_2 \nabla g(v^{k+1}) + v^{k+1} - v^k - \tau_2 \left(v^{k+1} - v^k - \frac{\lambda^{k+1}}{t} \right) \\
\Leftrightarrow\ & \frac{1}{\tau_2} \left(v^k - v^{k+1} \right) - \left(v^k - v^{k+1} \right) - \frac{\lambda^{k+1}}{t} = \nabla g(v^{k+1}).
\end{aligned} \tag{16}$$

∇g is a monotone operator due to the convexity of g, thus combining (16) with $-\lambda^* = \nabla g(v^*)$ by (10), we have

$$\left\langle \frac{1}{\tau_2} \left(v^k - v^{k+1} \right) - \left(v^k - v^{k+1} \right) - \frac{1}{t} \left(\lambda^{k+1} - \lambda^* \right), v^{k+1} - v^* \right\rangle \geq 0. \tag{17}$$

Summing (14) and (17) gives

$$\begin{aligned}
&\frac{1}{\tau_1} \left(w^k - w^{k+1} \right)^\top \left(w^{k+1} - w^* \right) - \left(w^k - w^{k+1} \right)^\top Q^\top Q \left(w^{k+1} - w^* \right) + \left(v^k - v^{k+1} \right)^\top \left(Qw^{k+1} - Qw^* \right) \\
&+ \frac{1}{\tau_2} \left(v^k - v^{k+1} \right)^\top \left(v^{k+1} - v^* \right) - \left(v^k - v^{k+1} \right)^\top \left(v^{k+1} - v^* \right) + \frac{1}{t} \left(\lambda^{k+1} - \lambda^* \right)^\top \left(Qw^{k+1} - v^{k+1} - Qw^* + v^* \right) \\
&\geq 0.
\end{aligned} \tag{18}$$

Then by using (8) $\lambda^{k+1} = \lambda^k - t \left(Qw^{k+1} - v^{k+1} \right)$ and (11) $Qw^* - v^* = \mathbf{0}$, we have

$$\begin{aligned}
&\frac{1}{\tau_1} \left(w^k - w^{k+1} \right)^\top \left(w^{k+1} - w^* \right) - \left(w^k - w^{k+1} \right)^\top Q^\top Q \left(w^{k+1} - w^* \right) + \frac{1}{t} \left(v^k - v^{k+1} \right)^\top \left(\lambda^k - \lambda^{k+1} \right) \\
&+ \frac{1}{\tau_2} \left(v^k - v^{k+1} \right)^\top \left(v^{k+1} - v^* \right) + \frac{1}{t^2} \left(\lambda^{k+1} - \lambda^* \right)^\top \left(\lambda^k - \lambda^{k+1} \right) \geq 0 \\
\Leftrightarrow\ & \left(w^k - w^{k+1} \right)^\top \left(\frac{t}{\tau_1} \mathbf{I} - tQ^\top Q \right) \left(w^{k+1} - w^* \right) + \frac{t}{\tau_2} \left(v^k - v^{k+1} \right)^\top \left(v^{k+1} - v^* \right) \\
&+ \frac{1}{t} \left(\lambda^{k+1} - \lambda^* \right)^\top \left(\lambda^k - \lambda^{k+1} \right) \geq - \left(v^k - v^{k+1} \right)^\top \left(\lambda^k - \lambda^{k+1} \right),
\end{aligned} \tag{19}$$

which can be written in the matrix form as

$$\begin{aligned}
&\left(\left(w^k - w^{k+1} \right)^\top, \left(v^k - v^{k+1} \right)^\top, \left(\lambda^k - \lambda^{k+1} \right)^\top \right) \begin{pmatrix} \frac{1}{\tau_1} \mathbf{I} - tQ^\top Q & & \\ & \frac{1}{\tau_2} \mathbf{I} & \\ & & \frac{1}{t} \mathbf{I} \end{pmatrix} \begin{pmatrix} w^{k+1} - w^* \\ v^{k+1} - v^* \\ \lambda^{k+1} - \lambda^* \end{pmatrix} \\
&\geq - \left(v^k - v^{k+1} \right)^\top \left(\lambda^k - \lambda^{k+1} \right).
\end{aligned} \tag{20}$$

Using notations z^k, z^* and M, we have

$$\begin{aligned}&\left\langle z^k - z^{k+1}, z^{k+1} - z^* \right\rangle_M \geq -\left\langle v^k - v^{k+1}, \lambda^k - \lambda^{k+1} \right\rangle \\ \Leftrightarrow &\left\langle z^k - z^{k+1}, z^k - z^* \right\rangle_M \geq \|z^k - z^{k+1}\|_M^2 - \left\langle v^k - v^{k+1}, \lambda^k - \lambda^{k+1} \right\rangle.\end{aligned}$$

Therefore,

$$\begin{aligned}&\|z^k - z^*\|_M^2 - \|z^{k+1} - z^*\|_M^2 \\ =&2\left\langle z^k - z^{k+1}, z^k - z^* \right\rangle_M - \|z^{k+1} - z^k\|_M^2 \\ \geq&2\left(\|z^k - z^{k+1}\|_M^2 - \left\langle v^k - v^{k+1}, \lambda^k - \lambda^{k+1} \right\rangle\right) - \|z^{k+1} - z^k\|_M^2 \\ =&\|z^k - z^{k+1}\|_M^2 - 2\left\langle v^k - v^{k+1}, \lambda^k - \lambda^{k+1} \right\rangle \\ \geq&\|z^k - z^{k+1}\|_M^2 - 2\|v^k - v^{k+1}\|_2 \|\lambda^k - \lambda^{k+1}\|_2 \\ \geq&\|z^k - z^{k+1}\|_M^2 - \gamma\|v^k - v^{k+1}\|_2^2 - \frac{1}{\gamma}\|\lambda^k - \lambda^{k+1}\|_2^2 \\ =&\left(w^k - w^{k+1}\right)^\top \left(\frac{t}{\tau_1}\mathbf{I} - tQ^\top Q\right)\left(w^k - w^{k+1}\right) + \left(v^k - v^{k+1}\right)^\top \left(\frac{t}{\tau_2} - \gamma\right)\mathbf{I}\left(v^k - v^{k+1}\right) \\ &+ \left(\lambda^k - \lambda^{k+1}\right)^\top \left(\frac{1}{t} - \frac{1}{\gamma}\right)\mathbf{I}\left(\lambda^k - \lambda^{k+1}\right) \\ \geq&\left(\frac{t}{\tau_1} - t\sigma_{\max}(Q)\right)\|w^k - w^{k+1}\|_2^2 + \left(\frac{t}{\tau_2} - \gamma\right)\|v^k - v^{k+1}\|_2^2 + \left(\frac{1}{t} - \frac{1}{\gamma}\right)\|\lambda^k - \lambda^{k+1}\|_2^2 \\ \geq&\min\left\{\frac{t}{\tau_1} - t\sigma_{\max}(Q), \frac{t}{\tau_2} - \gamma, \frac{1}{t} - \frac{1}{\gamma}\right\}\left(\|w^k - w^{k+1}\|_2^2 + \|v^k - v^{k+1}\|_2^2 + \|\lambda^k - \lambda^{k+1}\|_2^2\right) \\ =&c\|z^k - z^{k+1}\|_2^2 \\ \geq&0,\end{aligned}$$

where the second last inequality is because $\tau_1 < \frac{1}{\sigma_{\max}(Q)}$, and by letting letting $\gamma = \frac{1+\tau_2}{2\tau_2}t$ and $\tau_2 < 1$ we have $\frac{t}{\tau_2} - \gamma > 0$ and $\frac{1}{t} - \frac{1}{\gamma} > 0$. □

A.4. Proof of Theorem 1

Proof. 1° Firstly, we show that any limit point of $\left\{z^k\right\}_{k=0}^\infty$ is an optimal solution to (3). By Lemma 1, we know that

$$\|z^k - z^*\|_M^2 \geq \|z^{k+1} - z^*\|_M^2 \geq 0, \tag{21}$$

which indicates that the sequence $\left\{\|z^k - z^*\|_M^2\right\}_{k=0}^\infty$ is monotonically non-increasing and lower bounded, and thus converges. Hence by

$$0 \leq \|z^k - z^{k+1}\|_M^2 \leq \frac{1}{c}\left(\|z^k - z^*\|_M^2 - \|z^{k+1} - z^*\|_M^2\right) \to 0,$$

we have $w^k - w^{k+1} \to \mathbf{0}$, $v^k - v^{k+1} \to \mathbf{0}$, and $\lambda^k - \lambda^{k+1} \to \mathbf{0}$. Then by (8) we have the feasibility $Qw^k - v^k \to \mathbf{0}$. Besides,

$$\|z^k\|_M \leq \|z^k - z^*\|_M + \|z^*\|_M \leq \|z^0 - z^*\|_M + \|z^*\|_M,$$

which means that the sequence $\left\{z^k\right\}_{k=0}^\infty$ is bounded. Thus, $\left\{z^k\right\}_{k=0}^\infty$ contains a subsequence $\left\{u^{k_\ell}\right\}_{l=0}^\infty$ that converges to the limit point $\hat{z} = \left(\hat{w}, \hat{v}, \hat{\lambda}\right)$. Taking limit for both sides of (13) and (16), we have

$$\begin{aligned}\frac{1}{t}Q^\top\hat{\lambda} \in&\partial f(\hat{w}), \\ -\frac{1}{t}\hat{\lambda} =&\nabla g(\hat{v}).\end{aligned}$$

Combined with the feasibility $Q\hat{w} - \hat{v} = 0$, we know that $\hat{z} = \left(\hat{w}, \hat{v}, \hat{\lambda}\right)$ is a KKT point of problem (3) and thus it is optimal.

2° It suffices to further prove that the limit point of $\left\{z^k\right\}_{k=0}^{\infty}$ is unique. Suppose there are two subsequences $\{z^{p_\ell}\}_{l=0}^{\infty}$ and $\{z^{q_\ell}\}_{l=0}^{\infty}$ that converge to limit points $\hat{z}_1$ and $\hat{z}_2$, respectively. Similarly, we can show that $\hat{z}_1$ and $\hat{z}_2$ are both KKT points of (3). Analogous to (21), we have

$$\begin{aligned}\|z^k-\hat{z}_1\|_M^2 \geq \|z^{k+1}-\hat{z}_1\|_M^2 \geq 0,\\ \|z^k-\hat{z}_2\|_M^2 \geq \|z^{k+1}-\hat{z}_2\|_M^2 \geq 0,\end{aligned}$$

which implies that there exist ξ_1 and ξ_2 such that

$$\lim_{k\to\infty}\|z^k-\hat{z}_1\|_M^2=\xi_1, \tag{22}$$

$$\lim_{k\to\infty}\|z^k-\hat{z}_2\|_M^2=\xi_2. \tag{23}$$

Besides, we have the identity

$$\|z^k-\hat{z}_1\|_M^2-\|z^k-\hat{z}_2\|_M^2=-2\left\langle z^k,\hat{z}_1-\hat{z}_2\right\rangle_M+\|\hat{z}_1\|_M^2-\|\hat{z}_2\|_M^2. \tag{24}$$

Taking limits for both sides of (24) with regard to subsequences $\{z^{p_\ell}\}_{\ell=0}^{\infty}$ and $\{z^{q_\ell}\}_{\ell=0}^{\infty}$ respectively, we have

$$\begin{aligned}\lim_{\ell\to\infty}\left(\|z^{p_\ell}-\hat{z}_1\|_M^2-\|z^{p_\ell}-\hat{z}_2\|_M^2\right)=\lim_{\ell\to\infty}-2\left\langle z^{p_\ell},\hat{z}_1-\hat{z}_2\right\rangle_M+\|\hat{z}_1\|_M^2-\|\hat{z}_2\|_M^2,\\ \lim_{\ell\to\infty}\left(\|z^{q_\ell}-\hat{z}_1\|_M^2-\|z^{q_\ell}-\hat{z}_2\|_M^2\right)=\lim_{\ell\to\infty}-2\left\langle z^{q_\ell},\hat{z}_1-\hat{z}_2\right\rangle_M+\|\hat{z}_1\|_M^2-\|\hat{z}_2\|_M^2.\end{aligned}$$

Together with (22) and (23), we have

$$\begin{aligned}\xi_1-\xi_2=-2\left\langle\hat{z}_1,\hat{z}_1-\hat{z}_2\right\rangle_M+\|\hat{z}_1\|_M^2-\|\hat{z}_2\|_M^2=-\|\hat{z}_1-\hat{z}_2\|_M^2,\\ \xi_1-\xi_2=-2\left\langle\hat{z}_2,\hat{z}_1-\hat{z}_2\right\rangle_M+\|\hat{z}_1\|_M^2-\|\hat{z}_2\|_M^2=\|\hat{z}_1-\hat{z}_2\|_M^2,\end{aligned}$$

which implies that $\|\hat{z}_1-\hat{z}_2\|_M^2=0$, i.e., $\hat{z}_1=\hat{z}_2$. Thus, the bounded sequence $\left\{z^k\right\}_{k=0}^{\infty}$ has a unique limit point. Consequently, $\left\{z^k\right\}_{k=0}^{\infty}$ converges to that unique limit point.

Combining 1° and 2°, we conclude that $\left\{z^k\right\}_{k=0}^{\infty}$ converges, and the limit point is optimal to (3). $\square$