IMPROVING GRAPH TREND FILTERING WITH NON-CONVEX PENALTIES

Rohan Varma[†], *Harlin Lee*[†], *Yuejie Chi*[†], *Jelena Kovačević*[*]

[†]Dept. of Electrical and Computer Engineering, Carnegie Mellon University
[*]Tandon School of Engineering, New York University

ABSTRACT

In this paper, we study the denoising of piecewise smooth graph signals that exhibit inhomogeneous levels of smoothness over a graph. We extend the graph trend filtering framework to a family of non-convex regularizers that exhibit superior recovery performance over existing convex ones. We present theoretical results in the form of asymptotic error rates for both generic and specialized graph models. We further present an ADMM-based algorithm to solve the proposed optimization problem and analyze its convergence. Numerical performance of the proposed framework with non-convex regularizers on both synthetic and real-world data are presented for denoising, support recovery, and semi-supervised classification.

***Index Terms*—** graph signal processing, graph trend filtering, piecewise smooth graph signals, semi-supervised classification, non-convex penalties

1. INTRODUCTION

Signal estimation from noisy observations is a well-studied problem in signal processing and has applications for signal inpainting, collaborative filtering, recommender systems and other large-scale data completion problems. Since noise can have deleterious, cascading effects in many downstream tasks, being able to efficiently and accurately reconstruct a signal is of significant importance.

With the explosive growth of information and communication, signals are generated at an unprecedented rate from various sources, including social networks, citation networks, biological networks, and physical infrastructure [1]. Unlike time-series signals or images, these signals lie on complex, irregular graph structures, and require novel processing techniques, leading to the emerging field of signal processing on graphs [2–4]. The associated graph-structured data are referred to as *graph signals*. In graph signal processing, a canonical assumption is that the graph signal is smooth with respect to the graph, that is, the signal coefficients do not vary much over local neighborhoods of the graph. However, this characterization is insufficient for many real-world signals. There are often localized discontinuities and patterns in the signal, and the signal is smooth in a *piecewise* manner over the graph. In community detection, for example, the label is constant within each group, but discontinuous over the edges that connect nodes in different groups. As a result, it is necessary to develop representations and algorithms to process and analyze such *piecewise smooth* graph signals.

In this work, we study the denoising of piecewise smooth graph signals that exhibit an inhomogeneous level of smoothness over the graph and have abrupt, localized discontinuities. The class of piecewise smooth signals, which includes piecewise constant graph signals, is complementary to the class of smooth graph signals that exhibit homogeneous levels of smoothness over the graph. The reconstruction of such smooth signals has been well studied in previous work both within the field of graph signal processing as well as in the context of Laplacian regularization.

The graph trend filtering (GTF) framework [5], which applies total variation denoising on graphs [6], is a particularly flexible and attractive approach that is based on minimizing the ℓ_1 norm of discrete graph differences. In this work, we present an extension to the GTF framework and apply a family of non-convex regularizers that exhibit superior recovery performance over ℓ_1 norm minimization. Although the ℓ_1 norm based regularization has many attractive properties [7], it is well-known that the estimates are biased toward zero for large coefficients. To reduce the bias, nonconvex penalties such as the smoothly clipped absolute deviation (SCAD) penalty [8] and the minimax concave penalty (MCP) [9] are proposed as alternatives with the attractive *oracle property*: in the asymptotic sense, they perform as well as the case where we know in advance the support of the sparse vectors [10–14]. These penalties behave similarly to the ℓ_1 norm when the signal values are small, but tend to a constant when the signal values are large. Through theoretical analyses and empirical performance, we demonstrate the improved performance of GTF using non-convex penalties such as SCAD and MCP in terms of both reduced reconstruction error as well as improved support recovery, i.e. how accurately we can localize the boundaries and discontinuities of the piecewise smooth signals.

The rest of this paper is organized as follows. In Section 2, we provide some background and definitions on graph signal processing and GTF. Section 3 presents the proposed GTF framework with non-convex penalties, its performance guarantee, and an efficient algorithm based on ADMM. Numerical performances of the proposed approach are examined on both synthetic and real-world data for denoising and semi-supervised classification in Section 4. Finally, we conclude in Section 5.

2. GRAPH SIGNAL PROCESSING, PIECEWISE SMOOTH SIGNALS, AND GRAPH TREND FILTERING

We consider an undirected graph $G = (\mathcal{V}, \mathcal{E}, \boldsymbol{A})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of nodes, $\mathcal{E} = \{e_1, \ldots, e_m\}$ is the set of edges, and $\boldsymbol{A} = [A_{j,k}] \in \mathbb{R}^{n \times n}$ is the *graph shift operator* [2], or the weighted adjacency matrix. The edge set $\mathcal{E}$ represents the connections of the undirected graph G, and the positive edge weight $A_{j,k}$ between nodes v_j and v_k measures the underlying relation between the jth and the kth node, such as a similarity, a dependency, or a communication pattern. Let a *graph signal* be defined as

$$\boldsymbol{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_n\right]^T \in \mathbb{R}^n,$$

The first two authors contributed equally.

Emails:{rohanv, harlinl, yuejiec}@andrew.cmu.edu, jelenak@nyu.edu.

This work is supported in part by NSF under grants CCF-1563918, CCF-1826519 and ECCS-1818571, by ONR under grant N00014-18-1-2142, and by ARO under grant W911NF-18-1-0303.

where β_i denotes the signal coefficient at the ith node.

Let $\boldsymbol{\Delta} \in \mathbb{R}^{m \times n}$ be the oriented incidence matrix of G, where each row corresponds to an edge. That is, if the edge $e_i = (j,k) \in \mathcal{E}$ connects the jth node to the kth node ($j < k$), the entries in the ith row of $\boldsymbol{\Delta}$ is then given as

$$\Delta_{i,\ell} = \begin{cases} -\sqrt{A_{j,k}}, & \ell = j; \\ \sqrt{A_{j,k}}, & \ell = k; \\ 0, & \text{otherwise} \end{cases}.$$

The entries of the signal $\boldsymbol{\Delta\beta} = [\sqrt{A_{j,k}}(\beta_k - \beta_j)]_{(j,k)\in\mathcal{E}}$ specifies the weighted pairwise differences of the graph signal over each edge. As a result, $\boldsymbol{\Delta}$ can be interpreted as a *graph difference operator*. In graph signal processing, a signal is called smooth over a graph G if $\|\boldsymbol{\Delta\beta}\|_2^2 = \sum_{(j,k)\in\mathcal{E}} A_{j,k}(\beta_k - \beta_j)^2$ is small.

2.1. Piecewise Smooth Graph Signals

In practice, the graph signal may not be necessarily smooth over the entire graph, but only locally within different pieces of the graph. To model inhomogeneous levels of smoothness over a graph, we say that a graph signal $\boldsymbol{\beta}$ is piecewise constant over a graph G if many of the differences $\beta_k - \beta_j$ are zero for $(j,k) \in \mathcal{E}$. Consequently, the difference signal $\boldsymbol{\Delta\beta}$ is sparse and $\|\boldsymbol{\Delta\beta}\|_0$ is small.

We can generalize this notion to characterize *piecewise kth order polynomial* signals on a graph, where the piecewise constant case corresponds to $k = 0$, by generalizing the notion of graph difference operators. Specifically, we use the following recursive definition of the kth order graph difference operator $\boldsymbol{\Delta}^{(k+1)}$ [5]. Let $\boldsymbol{\Delta}^{(1)} = \boldsymbol{\Delta}$ for $k = 0$. For $k \geq 1$, let

$$\boldsymbol{\Delta}^{(k+1)} = \begin{cases} \boldsymbol{\Delta}^{(1)T}\boldsymbol{\Delta}^{(k)} \in \mathbb{R}^{n\times n}, & \text{odd } k \\ \boldsymbol{\Delta}^{(1)}\boldsymbol{\Delta}^{(k)} \in \mathbb{R}^{m\times n}, & \text{even } k \end{cases}.$$

The signal $\boldsymbol{\beta}$ is said to be a piecewise kth order polynomial graph signal if $\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}\|_0$ is small. To further illustrate, let us consider the piecewise linear graph signal, corresponding to $k = 1$, as a signal whose value at a node can be linearly interpolated from the weighted average of the values at neighboring nodes. It is easy to see that this is the same as requiring the second-order differences $\boldsymbol{\Delta}^T\boldsymbol{\Delta\beta}$ to be sparse. Similarly, we say that a signal has a piecewise quadratic structure if the differences between the second-order differences defined for piecewise linear signals are mostly zero, that is, if $\boldsymbol{\Delta\Delta}^T\boldsymbol{\Delta\beta}$ is sparse.

2.2. Denoising Piecewise Smooth Graph Signals via GTF

Assume we observe a noisy signal $\boldsymbol{y}$ over the graph under i.i.d Gaussian noise:

$$\boldsymbol{y} = \boldsymbol{\beta}^\star + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}), \tag{1}$$

and seek to reconstruct $\boldsymbol{\beta}^\star$ from $\boldsymbol{y}$ by leveraging the graph structure. When $\boldsymbol{\beta}$ is a smooth graph signal, Laplacian smoothing [15–19] can be used, which solves the following problem:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\Delta\beta}\|_2^2, \tag{2}$$

where $\lambda > 0$. However, it cannot localize abrupt changes in the graph signal when the signal is piecewise smooth.

Graph trend filtering (GTF) [5] is a flexible framework for estimation on graphs that is adaptive to inhomogeneity in the level of smoothness of an observed signal across nodes. The kth order GTF estimate is defined as:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}\|_1, \tag{3}$$

which can be regarded as applying total variation or fused lasso with the graph difference operator $\boldsymbol{\Delta}^{(k+1)}$ [6, 20]. The sparsity-promoting properties of the ℓ_1 norm have been well-studied [21]. Consequently, applying the ℓ_1 penalty in GTF sets many of the (higher-order) graph differences to zero while keeping a small fraction of nonzero values such that the GTF estimate is *locally adaptive* over the graph.

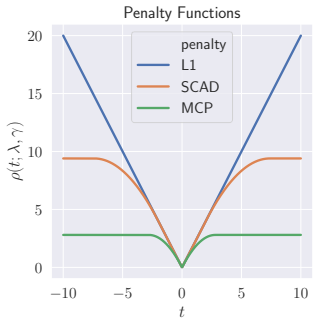

Fig. 1: *Illustration of $\rho(\cdot;\lambda,\gamma)$ for ℓ_1, SCAD ($\gamma = 3.7$), and MCP ($\gamma = 1.4$), where $\lambda = 2$.*

3. GTF WITH NON-CONVEX PENALTIES

The ℓ_1 norm penalty considered in (3) is well-known to produce biased estimates [22], which motivates us to extend the GTF framework to a broader class of sparsity-promoting regularizers that are not necessarily convex. We wish to solve the following generalized kth order GTF problem:

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}\,;\lambda,\gamma), \tag{4}$$

where $g(\boldsymbol{v};\lambda,\gamma) = \sum_{t=1}^T \rho(v_t;\lambda,\gamma)$ for $\boldsymbol{v} \in \mathbb{R}^T$ is a regularizer defined as the sum of the penalty function $\rho(\cdot;\lambda,\gamma) : \mathbb{R} \to \mathbb{R}$ applied on each coordinate v_t of $\boldsymbol{v}$. Similar to [10, 12, 23], we consider a family of penalty functions $\rho(\cdot;\lambda,\gamma)$ that satisfies the following assumptions.

Assumptions 1. (a) $\rho(t;\lambda,\gamma)$ satisfies $\rho(0;\lambda,\gamma) = 0$, is symmetric around 0, and is nondecreasing on the real line.

(b) For $t \geq 0$, the function $t \to \frac{\rho(t;\lambda,\gamma)}{t}$ is non-increasing in t. Also, $\rho(t;\lambda,\gamma)$ is differentiable for all $t \neq 0$ and subdifferentiable at $t = 0$, with $\lim_{t\to 0^+}\rho'(t;\lambda,\gamma) = \lambda$. This upper bounds $\rho(t;\lambda,\gamma) \leq \lambda|t|$.

(c) There exists $\mu > 0$ such that $\rho(t;\lambda,\gamma) + \frac{\mu}{2}t^2$ is convex.

Besides the ℓ_1 penalty, the non-convex SCAD [8] penalty

$$\rho_{\text{SCAD}}(t;\lambda,\gamma) = \lambda\int_0^{|t|} \min(1, \frac{(\gamma - u/\lambda)_+}{\gamma - 1})du, \quad \gamma \geq 2,$$

and MCP [9]

$$\rho_{\text{MCP}}(t;\lambda,\gamma) = \lambda\int_0^{|t|} (1 - \frac{u}{\lambda\gamma})_+ du, \quad \gamma \geq 1$$

also satisfy these assumptions, among others. We note that for SCAD, $\mu \geq \frac{1}{\gamma-1}$ and for MCP, $\mu \geq \frac{1}{\gamma}$. Fig. 1 illustrates the ℓ_1, SCAD and MCP penalties for comparisons. SCAD and MCP primarily differ from ℓ_1 penalty in that they apply less penalty over large signal values, and as a result mitigate the bias effect.

3.1. Error Bounds

We present asymptotic error rates on the generalized GTF problem (4) under the noise model in (1). Furthermore, we specialize the error rates to a few common graph models.

Theorem 1. Let C be the number of connected components in the graph G, or equivalently, the dimension of the null space of $\boldsymbol{\Delta}^{(k+1)}$. Further, let r be the number of rows of $\boldsymbol{\Delta}^{(k+1)}$, and ζ the maximum ℓ_2 norm of the columns of $\boldsymbol{\Delta}^{(k+1)\dagger}$. Setting $\lambda = \Theta(\zeta\sqrt{\log r})$, and for a penalty function $\rho(\cdot;\lambda,\gamma)$ such that $\mu < \frac{1}{\|\boldsymbol{\Delta}^{(k+1)}\|_2^2}$,

$$\frac{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2}{n} \leq O\left(\frac{C}{n}\right) + \frac{4g(\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}^\star;\lambda,\gamma)}{n(1-\mu\|\boldsymbol{\Delta}^{(k+1)}\|_2^2)} \tag{5}$$

$$\leq O\left(\frac{C}{n}\right) + \frac{4\Theta(\zeta\sqrt{\log r})\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}^\star\|_1}{n(1-\mu\|\boldsymbol{\Delta}^{(k+1)}\|_2^2)}. \tag{6}$$

We note that error rates for GTF with a non-convex regularizer $g(\cdot;\lambda,\gamma)$ are at least as fast as those using the ℓ_1 regularizer. Particularly, the rates with non-convex regularizers are faster when there are large coefficients on which they apply less shrinkage such that $g(\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}^\star;\lambda,\gamma) \ll \lambda\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}^\star\|_1$. It is shown in [5] that $\zeta \leq \frac{1}{\lambda_2(\boldsymbol{L})^{(k+1)/2}}$, where $\lambda_2(\boldsymbol{L})$ is the smallest non-zero eigenvalue of the graph Laplacian matrix $\boldsymbol{L} = \boldsymbol{\Delta}^{(1)T}\boldsymbol{\Delta}^{(1)}$ and quantifies the algebraic connectivity of the graph [24]. Moreover, one can bound $\lambda_2(\boldsymbol{L}) \geq \frac{4}{nD}$, where D is the diameter of the graph. Consequently, we get faster rates when the graph is well-connected and has a small diameter. We can further specialize the rates in Theorem 1 for some representative graphs to gain further insights.

- **Chain graph:** For univariate trend filtering,

$$\frac{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2}{n} = O\left(\sqrt{\frac{\log n}{n}} n^k \left\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}^\star\right\|_1\right).$$

- **d-regular graphs and Erdős-Rényi random graphs:** For d-regular graphs as well as Erdős-Rényi random graphs with edge probability $p \in (0,1)$ such that $d = np$,

$$\frac{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2}{n} = O\left(\frac{\sqrt{\log(nd)}}{nd^{\frac{k+1}{2}}} \left\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}^\star\right\|_1\right).$$

3.2. ADMM Algorithm

We optimize the generalized GTF formulation in (4) via the alternating direction method of multipliers (ADMM) framework for solving separable optimization problems [25]. Via a change of variable defining $\boldsymbol{\eta} = \boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}$, we can write the transformed problem

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\eta};\lambda,\gamma) \quad \text{s.t. } \boldsymbol{\eta} = \boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta}$$

and its corresponding Lagrangian as:

$$L(\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2 + g(\boldsymbol{\eta};\lambda,\gamma) + \frac{\tau}{2}\|\boldsymbol{\Delta}^{(k+1)}\boldsymbol{\beta} - \boldsymbol{\eta} + \boldsymbol{u}\|_2^2 - \frac{\tau}{2}\|\boldsymbol{u}\|_2^2 \tag{7}$$

where $\boldsymbol{u}$ is the Lagrangian multiplier, and τ the parameter. Algorithm 1 shows the ADMM updates based on the Lagrangian in (7). Note that both SCAD and MCP admit closed-form proximal operators. We have the following convergence guarantee for Alg. 1.

Theorem 2. Alg. 1 converges to a local minimum if $\tau \geq -\rho(\cdot;\lambda,\gamma)''$.

Algorithm 1 ADMM Optimization for Non-Convex GTF

Inputs: $\boldsymbol{y}$, $\boldsymbol{\Delta}^{(k+1)}$, and parameters λ, γ, τ
Initialize:
  $\boldsymbol{D} \leftarrow \boldsymbol{\Delta}^{(k+1)}$, $\boldsymbol{\eta} \leftarrow \boldsymbol{D\beta}$, $\boldsymbol{u} \leftarrow \boldsymbol{D\beta} - \boldsymbol{\eta}$,
  $\boldsymbol{\beta} \leftarrow \boldsymbol{y}$ or $\boldsymbol{\beta}_{init}$ if given.
repeat
  $\boldsymbol{\beta} \leftarrow (\boldsymbol{I} + \tau\boldsymbol{D}^T\boldsymbol{D})^{-1}(\tau\boldsymbol{D}^T(\boldsymbol{\eta} - \boldsymbol{u}) + \boldsymbol{y})$
  for `i ← 1 to length(`$\boldsymbol{D\beta}$`)` **do**
    $\eta_i \leftarrow \text{prox}_\rho([\boldsymbol{D\beta}]_i + u_i; \lambda/\tau)$
    ▷ $\text{prox}_\rho(t;\alpha)$ = proximal operator on t with $\alpha\rho$
  end for
  $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{D\beta} - \boldsymbol{\eta}$
until termination

4. NUMERICAL EXPERIMENTS

For the following experiments, we fixed $\gamma = 3.7$ for SCAD, $\gamma = 1.4$ for MCP, and tuned λ and $\frac{\tau}{\lambda}$ for each experiment. To meet the convergence criteria in Theorem 2, we enforce $\tau \geq \frac{1}{\gamma}$. SCAD/MCP were warm-started with the GTF estimate with ℓ_1 penalty.

4.1. Denoising via GTF with Non-Convex Regularizers

In this experiment, we compare the performance of GTF using non-convex regularizers such as SCAD and MCP with that using the ℓ_1 norm. For the ground truth, we construct a piecewise constant signal on a 20×20 2d-grid graph and the Minnesota road graph, and add different levels of noise as (1). We recover the signal with Alg. 1, and plot the SNR of the reconstructed signal versus the SNR of the input signal in Fig. 2. SCAD/MCP consistently outperforms ℓ_1 in denoising both regular and irregular graph signals. Below we further highlight two important advantages of non-convex regularizers.

Bias Reduction: We demonstrate the reduction in signal bias in Fig. 3 for the graph signal defined over a 12×12 2d-grid graph, using both the ℓ_1 penalty and the MCP penalty. Clearly, the MCP estimate (orange) has less bias than the ℓ_1 estimate (blue), and can recover the ground truth surface (purple) more closely.

Support Recovery: We illustrate the improved support recovery performance of non-convex regularizers [26] on localizing the boundaries for a piecewise constant signal on the Minnesota road graph as in Fig. 2. We aim to classify an edge as being 1) between two nodes in the same piece or 2) a cut edge across two pieces. By sweeping the regularization parameter λ, we obtain the ROC curve, the true positive rate versus the false positive rate of classifying an edge correctly, and see that MCP and SCAD consistently outperform the ℓ_1 penalty.

4.2. Semi-Supervised Classification

Graph-based learning provides a flexible and attractive way to model data in semi-supervised classification problems when labels are expensive to acquire [15, 16, 19], where a nearest-neighbors graph can be constructed based on the similarity between each pair of samples.

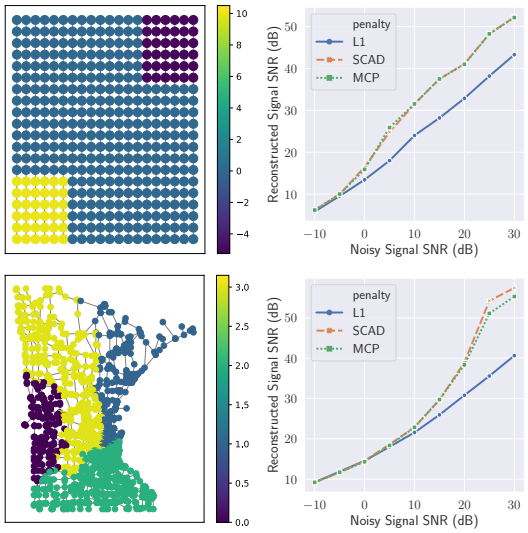

Fig. 2: *The ground truth piecewise constant signal on* 20×20 *2d-grid graph (top), and Minnesota road graph (bottom), and their corresponding plots of input signal SNR versus reconstructed signal SNR, averaged over 10 and 20 repetitions, respectively.*

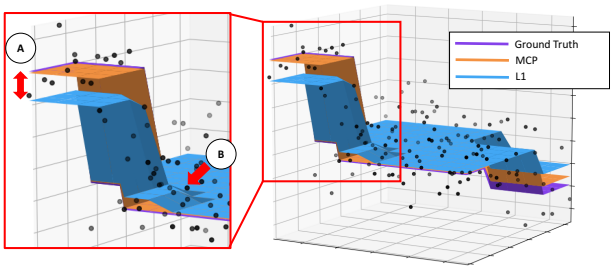

Fig. 3: *GTF using MCP (orange) has much lower bias than GTF using* ℓ_1 *(blue) when estimating a piecewise constant signal over a* 12×12 *grid graph. See highlighted regions pointed by red arrows in A and B. The scatter points are the noisy signal with 5dB SNR.*

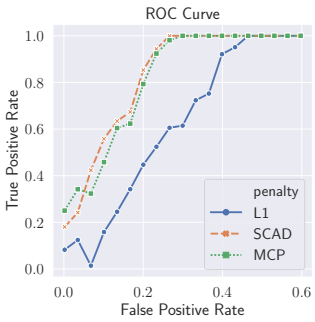

Fig. 4: *The ROC curve for classifying an edge as on the boundaries of pieces for the Minnesota road graph signal shown in Fig. 2. The noisy piecewise constant signal had input SNR = 7.8dB.*

We move beyond our original problem in (4) to a K-class classification problem in a semi-supervised learning setting, where for a given dataset with n samples, we observe a subset of the one-hot encoded class labels, $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K] \in \mathbb{R}^{n \times K}$, such that $Y_{i,j} = 1$ if ith sample has been observed to be in jth class, and $Y_{i,j} = 0$ otherwise. A diagonal, indicator matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ denotes samples whose class labels have been observed. We also let $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K] \in \mathbb{R}^{n \times K}$ (which set to be uniform in the experiment) be a fixed prior belief, and $\epsilon > 0$ determine how much emphasis to be given to the prior belief. Then, we can define the modified absorption problem [5, 16, 19] using the generalized GTF framework to estimate the unknown class probabilities $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K] \in \mathbb{R}^{n \times K}$:

$$\hat{\boldsymbol{B}} = \operatorname{argmin}_{\boldsymbol{B} \in \mathbb{R}^{n \times K}} \frac{1}{2} \sum_{j=1}^{K} \|\boldsymbol{M}(\boldsymbol{y}_j - \boldsymbol{\beta}_j)\|_2^2 + \sum_{j=1}^{K} g(\boldsymbol{\Delta}^{(k+1)} \boldsymbol{\beta}_j; \lambda, \gamma) + \epsilon \sum_{j=1}^{K} \|\boldsymbol{r}_j - \boldsymbol{\beta}_j\|_2^2 \quad (8)$$

The labels $\hat{\boldsymbol{Y}}$ can be estimated using $\hat{\boldsymbol{B}}$ such that $\hat{Y}_{i,j} = 1$ if and only if $j = \arg\max_{1 \le l \le K} \hat{B}_{i,l}$, and otherwise $\hat{Y}_{i,j} = 0$. We applied the algorithm in (8) to the 7 most popular UCI classification datasets [27] with $\epsilon = 0.01$, excluding *adult* which had more than 30,000 samples. For each dataset, we normalized each feature to have zero mean and unit variance, and constructed a 5NN graph of the samples based on the Euclidean distance between their features, with edge weights from Gaussian radial basis kernel. We randomly assigned 20% of samples in each class to be observed initially, and performed 10 repetitions. Table 1 shows that the misclassification rates from using non-convex penalties such as SCAD/MCP are competitive with those from ℓ_1.

		heart	wine-q.	wine	iris	breast	car
# of samples (n)		303	1599	178	150	569	1728
# of classes (K)		2	6	3	3	2	4
$k = 0$	L1	0.148	0.346	0.038	0.036	0.042	0.172
	SCAD	0.148	0.353	0.038	0.033	0.042	0.149
	p-value	1.	0.06	1.	0.27	1.	0.06
	MCP	0.144	0.351	0.037	0.035	0.040	0.148
	p-value	0.23	0.18	0.34	0.34	0.35	0.05
$k = 1$	L1	0.143	0.351	0.034	0.039	0.035	0.104
	SCAD	0.144	0.350	0.034	0.039	0.035	0.104
	p-value	0.30	0.43	0.34	1.	0.71	0.66
	MCP	0.146	0.350	0.034	0.039	0.034	0.103
	p-value	0.05	0.44	0.34	1.	0.02	0.23

Table 1: *Misclassification rates averaged over 10 trials, with p-values from running sampled t-tests between SCAD/MCP misclassification rates and the corresponding rates using* ℓ_1. *Cases where non-convex penalties perform better than* ℓ_1 *with p-value below* 0.1 *are highlighted in green, and where they perform worse are in red.*

5. CONCLUSIONS

We presented a framework for denoising piecewise smooth signals on graphs that generalizes the graph trend filtering framework to a family of non-convex regularizers. We presented theoretical guarantees on the asymptotic error rates of our framework, and presented a general algorithm to solve this generalized graph trend filtering problem. Furthermore, we demonstrated the superior performance of these non-convex regularizers in terms of reconstruction error, bias reduction, and support recovery on both synthetic and real-world data. In the future, we plan to present further theoretical guarantees on support recovery. Due to space constraints, the detailed proofs of the presented theorems are deferred to the full version [28].

6. REFERENCES

[1] M. Newman, *Networks*. Oxford University Press, 2018.

[2] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1644–1656, 2013.

[3] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. 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] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, 2018.

[5] Y.-X. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani, "Trend filtering on graphs," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3651–3691, 2016.

[6] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "\ell_1 Trend Filtering," *SIAM Review*, vol. 51, no. 2, pp. 339–360, 2009.

[7] P. Buhlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[8] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[9] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[10] P.-L. Loh and M. J. Wainwright, "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," in *Advances in Neural Information Processing Systems*, 2013, pp. 476–484.

[11] P.-L. Loh, "Statistical consistency and asymptotic normality for high-dimensional robust $ M $-estimators," *The Annals of Statistics*, vol. 45, no. 2, pp. 866–896, 2017.

[12] C.-H. Zhang and T. Zhang, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576–593, 2012.

[13] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The Annals of Applied Statistics*, vol. 5, no. 1, p. 232, 2011.

[14] S. Ma and J. Huang, "A concave pairwise fusion approach to subgroup analysis," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 410–423, 2017.

[15] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *International Conference on Computational Learning Theory*. Springer, 2004, pp. 624–638.

[16] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 2003, pp. 912–919.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*, 2002, pp. 585–591.

[18] ——, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[19] P. P. Talukdar and K. Crammer, "New regularized algorithms for transductive learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 442–457.

[20] R. J. Tibshirani, *The solution path of the generalized lasso*. Stanford University, 2011.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[22] C.-H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[23] L. Chen and Y. Gu, "The convergence guarantees of a non-convex approach for sparse recovery," *IEEE Transactions on Signal Processing*, vol. 62, no. 15, pp. 3754–3767, 2014.

[24] F. R. Chung and F. C. Graham, *Spectral Graph Theory*. American Mathematical Soc., 1997.

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. 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.

[26] P.-L. Loh and M. J. Wainwright, "Support recovery without incoherence: A case for nonconvex regularization," *The Annals of Statistics*, vol. 45, no. 6, pp. 2455–2482, 2017.

[27] A. Asuncion and D. Newman, *UCI Machine Learning Repository*, 2007.

[28] R. Varma, H. Lee, Y. Chi, and J. Kovačević, "Vector-valued graph trend filtering with nonconvex penalties," 2019. [Online]. Available: https://users.ece.cmu.edu/~yuejiec/papers/noncvxGTF.pdf

[29] F. Chung and M. Radcliffe, "On the spectra of general random graphs," *the electronic journal of combinatorics*, vol. 18, no. 1, p. 215, 2011.

[30] A. Lubotzky, R. Phillips, and P. Sarnak, "Ramanujan graphs," *Combinatorica*, vol. 8, no. 3, pp. 261–277, 1988.

[31] R. Merris, "Laplacian matrices of graphs: a survey," *Linear algebra and its applications*, vol. 197, pp. 143–176, 1994.

[32] B. Mohar, Y. Alavi, G. Chartrand, and O. R. Oellermann, "The Laplacian spectrum of graphs," *Graph Theory, Combinatorics, and Applications*, vol. 2, no. 871-898, p. 12, 1991.

[33] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.