Supplementary Material for "Trend Filtering on Graphs"

Yu-Xiang Wang[1] James Sharpnack[3] Alex Smola[1,4] Ryan J. Tibshirani[1,2]

yuxiangw@cs.cmu.edu jsharpna@gmail.com alex@smola.org ryantibs@stat.cmu.edu

[1] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213
[2] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
[3] Mathematics Department, University of California at San Diego, La Jolla, CA 10280
[4] Google, Strategic Technologies, Mountain View, CA 94303

We provide additional explanations and proofs for "Trend Filtering on Graphs".

A Graph Trend Filtering Interpretations

A.1 Piecewise Polynomials over Graphs

Here we give some insight for our definition of the family of graph difference operators (5) and (6), based on the idea of piecewise polynomials over graphs. In the univariate case, sparsity of β under the difference operator $D^{(k+1)}$ implies a specific kth order piecewise polynomial structure for the components of β [6, 8]. Since the components of β correspond to (real-valued) input locations $x = (x_1, \ldots x_n)$, the interpretation of a piecewise polynomial here is unambiguous. But for a graph, does sparsity of $\Delta^{(k+1)}\beta$ mean that the components of β are piecewise polynomial? And what does the latter even mean, as the components of β are defined over the nodes? To address these questions, we intuitively *define* a piecewise polynomial over a graph, and show that it implies sparsity under our constructed graph difference operators.

- **Piecewise constant ($k = 0$):** we say that a signal β is piecewise constant over a graph G if many of the differences $\beta_i - \beta_j$ are zero across edges $(i, j) \in E$ in G. Note that this is exactly the property associated with sparsity of $\Delta^{(1)}\beta$, since $\Delta^{(1)} = D$, the oriented incidence matrix of G.

- **Piecewise linear ($k = 1$):** we say that a signal β has a piecewise linear structure over G if β satisfies

$$\beta_i - \frac{1}{n_i} \sum_{(i,j)\in E} \beta_j = 0,$$

  for many nodes $i \in V$, where n_i is the number of nodes adjacent to i. In words, we are requiring that the signal components can be linearly interpolated from its neighboring values at many nodes in the graph. This is quite a natural notion of (piecewise) linearity: requiring that β_i be equal to the average of its neighboring values would enforce linearity at β_i under an appropriate embedding of the points in Euclidean space. Again, this is the same as requiring $\Delta^{(2)}\beta$ to be sparse, since $\Delta^{(2)} = L$, the graph Laplacian.

- **Piecewise polynomial ($k \geq 2$):** We say that β has a piecewise quadratic structure over G if the first differences $\alpha_i - \alpha_j$ of the second differences $\alpha = \Delta^{(2)}\beta$ are mostly zero, over edges $(i, j) \in E$. Likewise, β has a piecewise cubic structure over G if the second differences $\alpha_i - \frac{1}{n_i}\sum_{(i,j)\in E} \alpha_j$ of the second differences $\alpha = \Delta^{(2)}\beta$ are mostly zero, over nodes $i \in V$. This argument extends, alternating between leading first and second differences for even and odd k. Sparsity of $\Delta^{(k+1)}\beta$ in either case exactly corresponds to many of these differences being zero, by construction.

A.2 Electrical Network Interpretation of GTF Structure

Lemma 1 reveals a mathematical structure for GTF estimates $\hat{\beta}$, which satisfy $\hat{\beta} \in \Delta_{-A}^{(k+1)}$ for some set A. It is interesting to interpret the results using the electrical network perspective for graphs [7]. In this perspective, we think of replacing each edge in the graph with a resistor of value 1. If $c \in \mathbb{R}^n$ is a vector that describes how much current is going in at each node in the graph, then $v = Lc$ describes the induced voltage at each node. Provided that $\mathbb{1}^\top c = 0$, which means that the total accumulation of current in the network is 0, we can solve for the current values from the voltage values: $c = L^\dagger v$.

The odd case in Lemma 1 asserts that

$$\mathrm{null}(\Delta_{-A}^{(k+1)}) = \mathrm{span}\{\mathbb{1}\} + \{(L^\dagger)^{\frac{k+1}{2}} v : v_{-A} = 0\}.$$

For $k = 1$, this says that GTF estimates are formed by assigning a sparse number of nodes in the graph a nonzero voltage v, then solving for the induced current $L^\dagger v$ (and shifting this entire current vector by a constant amount). For $k = 3$, we assign a sparse number of nodes a nonzero voltage, solve for the induced current, and then *repeat this*: we relabel the induced current as input voltages to the nodes, and compute the new induced current. This process is again iterated for $k = 5, 7, \ldots$.

The even case in Lemma 1 asserts that

$$\mathrm{null}(\Delta_{-A}^{(k+1)}) = \mathrm{span}\{\mathbb{1}\} + (L^\dagger)^{\frac{k}{2}} \mathrm{span}\{\mathbb{1}_{C_1}, \ldots \mathbb{1}_{C_s}\}.$$

For $k = 2$, this result says that GTF estimates are given by choosing a partition $C_1, \ldots C_s$ of the nodes, and assigning a constant input voltage to each element of the partition. We then solve for the induced current (and potentially shift this by an overall constant amount). The process is iterated for $k = 4, 6, \ldots$ by relabeling the induced current as input voltage.

The comparison between the structure of estimates for $k = 2$ and $k = 3$ is informative: in a sense, the 2nd order GTF estimates will be *smoother* than the 3rd order estimates, because a sparse input voltage vector need not induce a current that is piecewise constant over nodes in the graph. E.g., an input voltage vector with only a few nodes having very large nonzero values will induce a current that is peaked around these nodes, but not piecewise constant.

B Proofs of Theoretical Results

B.1 Proof of Theorem 3

By assumption we can write

$$y = \beta_0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Let $\Delta \in \mathbb{R}^{r \times n}$, and denote $R = \mathrm{row}(\Delta)$, the row space of Δ, and $R^\perp = \mathrm{null}(\Delta)$, the null space of Δ. Also let P_R be the projection onto R, and $P_{R^\perp}$ the projection onto $R^\perp$. Consider

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|\Delta\beta\|_1,$$
$$\tilde{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2}\|P_R y - \beta\|_2^2 + \lambda \|\Delta\beta\|_1.$$

The first quantity $\hat{\beta} \in \mathbb{R}^n$ is the estimate of interest, the second one $\tilde{\beta} \in R$ is easier to analyze. Note that

$$\hat{\beta} = P_{R^\perp} y + \tilde{\beta},$$

and write $\|x\|_R = \|P_R x\|_2$, $\|x\|_{R^\perp} = \|P_{R^\perp} x\|_2$. Then

$$\|\hat{\beta} - \beta_0\|_2^2 = \|\epsilon\|_{R^\perp}^2 + \|\tilde{\beta} - \beta_0\|_R^2,$$

so assuming that $R^\perp$ is of constant dimension, it suffices to bound the first term. Now we establish a basic inequality for $\tilde\beta$. By optimality, we have

$$\frac{1}{2}\|y-\tilde\beta\|_R^2 + \lambda\|\Delta\tilde\beta\|_1 \le \frac{1}{2}\|y-\beta_0\|_R^2 + \lambda\|\Delta\beta_0\|_1,$$

and after rearranging terms,

$$\|\tilde\beta-\beta_0\|_R^2 \le 2\epsilon^\top P_R(\tilde\beta-\beta_0) + 2\lambda\|\Delta\beta_0\|_1 - 2\lambda\|\Delta\tilde\beta\|_1. \tag{B.1}$$

This is our basic inequality. In the first term above, we use $P_R = \Delta^\dagger\Delta$, and apply Holder's inequality:

$$\epsilon^\top\Delta^\dagger\Delta(\tilde\beta-\beta_0) \le \|(\Delta^\dagger)^\top\epsilon\|_\infty \|\Delta(\tilde\beta-\beta_0)\|_1. \tag{B.2}$$

Therefore if we choose $\lambda \ge \|(\Delta^\dagger)^\top\epsilon\|_\infty$, then we see from (B.1) that

$$\|\tilde\beta-\beta_0\|_R^2 \le 2\lambda\|\Delta(\tilde\beta-\beta_0)\|_1 + 2\lambda\|\Delta\beta_0\|_1 - 2\lambda\|\Delta\tilde\beta\|_1,$$

i.e.,

$$\|\tilde\beta-\beta_0\|_R^2 \le 4\lambda\|\Delta\beta_0\|_1. \tag{B.3}$$

Finally, $\|(\Delta^\dagger)^\top\epsilon\|_\infty \le O_{\mathbb{P}}(B\sqrt{\log r})$ by a standard result on maxima of Gaussians, where recall B is the maximum ℓ_2 norm of the columns of $\Delta^\dagger$. Thus with $\lambda = \Theta(B\sqrt{\log r})$, we have from (B.3),

$$\|\tilde\beta-\beta_0\|_R^2 = O_{\mathbb{P}}\big(B\sqrt{\log r}\|\Delta\beta_0\|_1\big),$$

or

$$\frac{\|\tilde\beta-\beta_0\|_R^2}{n} = O_{\mathbb{P}}\left(\frac{B\sqrt{\log r}}{n}\|\Delta\beta_0\|_1\right),$$

as desired.

B.2 Proof of Corollary 4

Case 1. We consider $\Delta = D^{(k+1)}$, the univariate trend filtering operator of order $k+1$. Here the number of rows of the Δ is $r = n-k-1$, and the dimension of its null space is $k+1$. Further, it is not hard to verify that $\Delta^\dagger = P_R H/k!$, where recall $R = \mathrm{row}(\Delta)$, and $H \in \mathbb{R}^{n\times(n-k-1)}$ contains the last $n-k-1$ columns of the falling factorial basis matrix, evaluated over inputs $x_1 = 1, \dots x_n = n$ [8]. The largest column norm of $P_R H/k!$ is on the order of $n^{k+1/2}$, which proves the result.

Cases 2 and 3. When G is the Ramanujan d-regular graph, the number of edges in the graph is $O(nd)$. The operator $\Delta = \Delta^{(k+1)}$ has number of rows $r = n$ when k is odd and $r = O(nd)$ when k is even; overall this is $O(nd)$. The dimension of the null space of Δ is constant (it is in fact 1, since the graph is connected). When G is the Erdos-Renyi random graph, the same bounds apply to the number of rows and the dimension of the null space, except that the bounds become probabilistic ones.

Now we apply the crude inequality

$$B = \max_{i=1,\dots r} \Delta^\dagger e_i \le \max_{\|x\|_2\le 1} \Delta^\dagger x = \|\Delta^\dagger\|_2,$$

the right-hand side being the maximum singular value of $\Delta^\dagger$. As $\Delta = \Delta^{(k+1)}$, the graph difference operator of order $k+1$, we claim that

$$\|\Delta^\dagger\|_2 \le 1/\lambda_{\min}(L)^{\frac{k+1}{2}}, \tag{B.4}$$

where $\lambda_{\min}(L)$ denotes the smallest nonzero eigenvalue of the graph Laplacian L. To see this, note first that $\|\Delta^\dagger\|_2 = 1/\sigma_{\min}(\Delta)$, where the denominator is the smallest nonzero singular value of Δ. Now for odd k, we have $\Delta^{(k+1)} = L^{\frac{k+1}{2}}$, and the claim follows as

$$\sigma_{\min}(L^{\frac{k+1}{2}}) = \min_{x\in R:\|x\|_2\le 1} L^{\frac{k+1}{2}} \ge \big(\sigma_{\min}(L)\big)^{\frac{k+1}{2}},$$

and $\sigma_{\min}(L) = \lambda_{\min}(L)$, since L is symmetric. Above, R denotes the row space of L (the space orthogonal to the vector $\mathbb{1}$ of all 1s). For even k, we have $\Delta^{(k+1)} = DL^{\frac{k}{2}}$, and again

$$\sigma_{\min}(DL^{\frac{k}{2}}) = \min_{x \in R : \|x\|_2 \le 1} DL^{\frac{k+1}{2}} \ge \sigma_{\min}(D)\big(\sigma_{\min}(L)\big)^{\frac{k}{2}},$$

where $\sigma_{\min}(D) = \sqrt{\lambda_{\min}(L)}$, since $D^\top D = L$. This verifies the claim.

Hence having established (B.4), it suffices to lower bound $\lambda_{\min}(L)$ for the two graphs in question. Indeed, for both graphs, we have the lower bound

$$\lambda_{\min}(L) = \Omega(d - \sqrt{d}).$$

e.g., see Lubotzky et al. [3], Marcus et al. [4] for the Ramanujan graph and Feige and Ofek [2], Chung and Radcliffe [1] for the Erdos-Renyi graph. This completes the proof.

B.3 Proof of Theorem 5

A modification of the Holder bound (B.2) in the proof of Theorem 3 leads to potentially a sharper bound. Suppose that we were able to argue that

$$\epsilon^\top P_R(\tilde{\beta} - \beta_0) \le C_1 \|\tilde{\beta} - \beta_0\|_R + C_2 \|\Delta(\tilde{\beta} - \beta_0)\|_1, \tag{B.5}$$

with probability tending to 1, for some C_1, C_2. Following the proof strategy of Theorem 3, then we would take $\lambda \ge C_2/2$, and arrive at

$$\|\tilde{\beta} - \beta_0\|_R^2 \le C_1 \|\tilde{\beta} - \beta_0\|_R + 4\lambda \|\Delta\beta_0\|_1.$$

This is a quadratic of the form $ax^2 - bx - c \le 0$ in $x = \|\tilde{\beta} - \beta_0\|_R$. As $a > 0$, the larger of its two roots serves as a bound for x. That is, $x \le (b + \sqrt{b^2 + 4ac})/(2a) \le b/a + \sqrt{c/a}$, or

$$\|\tilde{\beta} - \beta_0\|_R \le C_1 + \sqrt{4\lambda \|\Delta\beta_0\|_1}. \tag{B.6}$$

Depending on C_1, C_2, the above bound can be significantly stronger than the previous one in (B.3); if $C_1 = 0$, then (B.6) simply reduces to (B.3); if C_1 is too large, then (B.6) could be actually weaker than (B.3); but for C_1 somewhere in the middle, (B.6) can substantially improve on (B.3), if C_2 is much smaller than $B\sqrt{\log r}$. Our next lemma shows that a bound of the form (B.5) is possible under the incoherence assumption Δ. Plugging in the appropriate quantities C_1, C_2 into (B.6) (with $\lambda = C_2/2$) then gives the final result.

Lemma B.1. *Let $\xi_1 \le \ldots \le \xi_n$ be the singular values of Δ, and let $\psi_1, \ldots \psi_r$ be the left singular vectors, satisfying the incoherence condition:*

$$\|\psi_i\|_\infty \le \mu/\sqrt{n}, \quad i = 1, \ldots r,$$

for some $\mu > 0$. For an index $i_0 \in \{1, \ldots n\}$, let

$$C = \mu \sqrt{\frac{2 \log 2r}{n} \sum_{i=i_0+1}^{n} \frac{1}{\xi_i^2}}.$$

Then, assuming that $i_0 \to \infty$, we have

$$\epsilon^\top P_R(\tilde{\beta} - \beta_0) \le 1.001\sigma\big(\sqrt{i_0}\|\tilde{\beta} - \beta_0\|_R + C\|\Delta(\tilde{\beta} - \beta_0)\|_1\big),$$

with probability tending to 1.

Proof. We will abuse notation and define for a scalar a the pseudoinverse to be $a^\dagger = 1/a$ for $a \neq 0$ and 0 otherwise. Throughout this proof let $[n] = \{1, \ldots, n\}$. Let the singular value decomposition of Δ be

$$\Delta = \Psi \Xi \Phi^\top.$$

where $\Psi \in \mathbb{R}^{r \times r}$, $\Phi \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Xi \in \mathbb{R}^{r \times n}$ has diagonal elements $(\Xi)_{ii} = \xi_i$, $i \in [n]$. First, let us establish that

$$\Delta^\dagger = \Phi \Xi^\dagger \Psi^\top,$$

where $\Xi^\dagger \in \mathbb{R}^{n \times r}$ and $(\Xi^\dagger)_{ii} = \xi_i^\dagger$ for $i \in [n]$. Consider a vector $\delta \in \mathbb{R}^n$ such that

$$\sqrt{i_0} \|\delta\|_2 + C \|\Delta \delta\|_1 \leq 1.$$

Denote the projection $P_{i_0} = \Phi_{[i_0]} \Phi_{[i_0]}^\top$ where $\Phi_{[i_0]}$ contains the first i_0 right singular vectors. We can decompose

$$\epsilon^\top P_R \delta = \epsilon^\top P_{i_0} P_R \delta + \epsilon^\top (I - P_{i_0}) P_R \delta.$$

The first term can be bounded by

$$\epsilon^\top P_{i_0} P_R \delta \leq \|P_{i_0} \epsilon\|_2 \|P_R \delta\|_2 \leq 1.001 \sigma \sqrt{i_0} \|\delta\|_2,$$

via the fact that $\|P_{i_0}\epsilon\|_2^2 \overset{d}{=} \sum_{i=1}^{i_0} \epsilon_i^2$ and the law of large numbers. We can bound the second term by

$$\epsilon^\top (I - P_{i_0}) P_R \delta = \epsilon^\top (I - P_{i_0}) \Delta^\dagger \Delta \delta \leq \|(\Delta^\dagger)^\top (I - P_{i_0}) \epsilon\|_\infty \|\Delta \delta\|_1,$$

using $P_R = \Delta^\dagger \Delta$ and Holder's inequality. Define $g_j = (I - P_{i_0}) \Delta^\dagger e_j$ for $j \in [r]$ with e_j the jth canonical basis vector. So,

$$\|g_j\|_2^2 = \|\Phi_{[n] \backslash [i_0]} \Xi^\dagger \Psi^\top e_j\|_2^2 \leq \frac{\mu^2}{n} \sum_{k=i_0+1}^{n} (\xi_k^\dagger)^2,$$

by rotational invariance of $\|\cdot\|_2$ and the incoherence assumption. Then, by a standard result of the maxima of Gaussians,

$$\|(\Delta^\dagger)^\top (I - P_{i_0}) \epsilon\|_\infty = \max_{j \in [r]} |g_j^\top \epsilon| \leq 1.001 \sigma \sqrt{2 \log(2r) \frac{\mu^2}{n} \sum_{k=i_0+1}^{n} (\xi_k^\dagger)^2} = 1.001 \sigma C,$$

with probability approaching 1. Hence with probability tending to 1,

$$\epsilon^\top P_R \delta \leq 1.001 \sigma \big(\sqrt{i_0} \|\delta\|_2 + C \|\Delta \delta\|_1 \big) \leq 1.001 \sigma,$$

for all δ such that $\sqrt{i_0} \|\delta\|_2 + C \|\Delta \delta\|_1 \leq 1$. Applying this to the particular choice

$$\delta = (\tilde{\beta} - \beta_0) / (\sqrt{i_0} \|(\tilde{\beta} - \beta_0)\|_2 + C \|\Delta(\tilde{\beta} - \beta_0)\|_1),$$

proves the lemma. □

B.4 Proof of Corollary 6

We can associate to every vertex in the torus a pair $i_1, i_2 \in [\ell] \times [\ell]$. Recall that Δ in this context is the combinatorial Laplacian L. It can be shown that the eigenvalues and eigenvectors of Δ associated with the pair are

$$2(2 - \cos(2\pi i_1) - \cos(2\pi i_2), N_\ell^{-2} (\sin(2\pi k_1 i_1 / \ell))_{k_1 \in [\ell]} \otimes (\sin(2\pi k_2 i_2 / \ell))_{k_2 \in [\ell]} = \psi_{(i_1, i_2)},$$

where N_ℓ is a normalizing constant that forces the eigenvectors to be of unit norm. Due to this constraint, $N_\ell \sim \sqrt{\ell}$, where $a \sim b$ indicates that $a = b(1 + o(1))$. We have that

$$\|\psi_{(i_1,i_2)}\|_\infty \sim n^{-1/2}$$

uniformly, and so it obeys the coherence condition with μ arbitrarily close to 1 for n large enough. The remainder of this proof comes from Sharpnack et al. [5], but we include it here for completeness. Now, we turn to calculating the functional $\sum_{i=i_0+1}^n \xi_i^{-2}$. For $i \in [\ell]$, we have $|\{(i_1, i_2) : i_1 \vee i_2 = i\}| \leq 2i$ and we know that $\xi_{i_1,i_2} \geq 2(1 - \cos(2\pi i_1 \vee i_2/\ell))$. Letting $j_0 \in [\ell]$ such that $j_0 = o(\ell)$, then

$$\begin{aligned}\frac{1}{n} \sum_{i_1,i_2:i_1\vee i_2>j_0} \frac{1}{\xi^2_{i_1,i_2}} &\leq \frac{1}{n} \sum_{j=j_0}^{\ell} \frac{2j}{(2(1-\cos(2\pi j/\ell)))^2} \\ &\leq \frac{1}{2\ell} \sum_{j=j_0}^{\ell} \frac{j/\ell}{(1-\cos(2\pi j/\ell))^2} \\ &\sim \frac{1}{2} \int_{j_0/\ell}^{1} \frac{x}{(1-\cos(2\pi x))^2} dx \sim \frac{1}{2}\left(\frac{\ell}{j_0}\right)^3,\end{aligned}$$

by a Taylor expansion about $x = 0$. Moreover, $i_0 = |\{i_1, i_2 \in [\ell] : i_1 \vee i_2 \leq j_0\}| = j_0^2$ and so we will seek to balance $i_0 = j_0^2$ with $\sqrt{(\log n)\ell^3/j_0^3}$. This is accomplished by

$$j_0 \sim (\log n)^{1/7} n^{3/14},$$

which is the order of C. Applying Theorem 5 with $i_0 = j_0^2$ and C as above gives us our result.

References

[1] F. Chung and M. Radcliffe. On the spectra of general random graphs. *The Electronic Journal of Combinatorics*, 18 (1), 2011. #P215.

[2] U. Feige and E. Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27 (2):251–275, 2005.

[3] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[4] A. W. Marcus, D. A. Spielman, and N. Srivastava. Ramanujan graphs and the solution of the Kadison-Singer problem. arXiv: 1408.4421, 2014.

[5] J. Sharpnack, A. Rinaldo, and A. Singh. Detecting anomalous activity on networks with the graph Fourier scan statistic. arXiv: 1311.7217, 2014.

[6] R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

[7] N. Vishnoi. $Lx = b$: Laplacian solvers and their algorithmic applications. *Foundations and Trends in Theoretical Computer Science*, 8(1–2):1–141, 2012.

[8] Y.-X. Wang, A. Smola, and R. J. Tibshirani. The falling factorial basis and its statistical properties. *International Conference on Machine Learning*, 31, 2014.