SEMI-SUPERVISED LEARNING WITH SPECTRAL GRAPH WAVELETS

David I Shuman, Mohammad Javad Faraji, and Pierre Vandergheynst

Signal Processing Laboratory (LTS2)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Emails: {david.shuman, mohammadjavad.faraji, pierre.vandergheynst}@epfl.ch

ABSTRACT

We consider the transductive learning problem when the labels belong to a continuous space. Through the use of spectral graph wavelets, we explore the benefits of multiresolution analysis on a graph constructed from the labeled and unlabeled data. The spectral graph wavelets behave like discrete multiscale differential operators on graphs, and thus can sparsely approximate piecewise smooth signals. Therefore, rather than enforce a prior belief that the labels are globally smooth with respect to the intrinsic structure of the graph, we enforce sparse priors on the spectral graph wavelet coefficients. One issue that arises when the proportion of data with labels is low is that the fine scale wavelets that are useful in sparsely representing discontinuities are largely masked, making it difficult to recover the high frequency components of the label sequence. We discuss this challenge, and propose one method to use the structured sparsity of the wavelet coefficients to aid label reconstruction.

***Keywords*—** Sparse approximation, spectral graph theory, structured sparsity, transductive regression, wavelets

1. INTRODUCTION

1.1. The Transductive Learning Problem

The goal of *semi-supervised learning* is to learn a mapping from a set of data points $X = \{x_1, x_2, \ldots, x_N\}$ to the corresponding labels $Y = \{y_1, y_2, \ldots, y_N\}$. The pairs (x_i, y_i) are sampled in an independent and identically distributed (iid) fashion according to a joint distribution $p(x, y)$ over the sample space $\mathcal{X} \times \mathcal{Y}$. $\mathcal{Y}$ may be equal to $\{-1, 1\}$ in the case of binary classification, $\{1, 2, \ldots, c\}$ in the general classification problem, or a continuous space such as $\mathbb{R}$ in a regression problem. The data $X = \{x_1, x_2, \ldots, x_N\}$ is split into the labeled data $X_l = \{x_1, x_2, \ldots, x_l\}$ and the unlabeled data $X_u = \{x_{l+1}, x_{l+2}, \ldots, x_N\}$. In addition to X, the labels associated with the labeled data, $Y_l = \{y_1, y_2, \ldots, y_l\}$, are provided. Usually, $l << N$; i.e., a small portion of the data is labeled. The *transductive learning* problem is to predict the labels Y_u associated with the unlabeled data. The primary motivation for semi-supervised learning is that in many applications, unlabeled data is "cheap," but labeled data may be "expensive," either in monetary cost or time required to assemble the labels. References [1, 2] survey approaches to semi-supervised learning problems, as well as common applications.

1.2. Enforcing Global Smoothness

The main idea behind semi-supervised learning is that the unlabeled data provides information about $p(x)$, which may in turn provide information about $p(y \mid x)$. To make this latter inference (i.e., for the unlabeled data to be useful), the problem must satisfy some structural properties. Three such properties targeted by different methods are (i) the *smoothness assumption* that if two data points are connected by a path of high density (i.e., $p(x)$ is high along the path), then the labels for the two points are similar; (ii) the *cluster assumption* that the data are clustered, and points within the same cluster likely have the same label; and (iii) the *manifold assumption* that the high-dimensional data X lie on a low-dimensional manifold [2].

Accordingly, a number of semi-supervised learning methods (e.g., [3, 4, 5]) proceed by representing the data X by a weighted, undirected graph $G = \{V, \mathcal{E}, w\}$, which consists of a set of vertices V, a set of edges $\mathcal{E}$, and a weight matrix w whose entries $w_{u,v}$ represent a non-negative weight if there is an edge connecting vertices u and v, and are zero otherwise.[1] They then force the labels to be smooth with respect to the intrinsic structure of this graph by, for example, solving a regularization problem of the form:

$$\min_{f=[f_l;f_u]} S(f) \quad \text{s.t.} \quad f_l = Y_l,$$

where $S(f)$ penalizes local variation of the labels between connected points on the graph. For example, [3, 5] consider:

$$S(f) := \sum_{(u,v)\in\mathcal{E}} w_{u,v} \left[f(v) - f(u)\right]^2 = f^{\mathrm{T}} \mathcal{L} f. \tag{1}$$

1.3. Beyond Global Smoothness

In this paper, we are specifically interested in regression problems ($\mathcal{Y}$ is continuous) where the labels may not be globally smooth with respect to the underlying graph structure, but rather

[1] The wide range of methods to construct graph weights includes those based on the Euclidean distances between data points and those based on the k-nearest neighbor graph.

piecewise smooth signals with discontinuities or large local variations. While simple methods based on global smoothness, such as the interpolated regularization algorithm of [5], do surprisingly well empirically on label functions with large local variations, our goal is to explore the benefits of multiresolution analysis on the graph. Reference [6] takes a similar approach by defining multiscale wavelets on trees. We extend the multiresolution approach to arbitrary graphs without restrictions on the underlying structure by leveraging spectral graph wavelets.

2. SPECTRAL GRAPH WAVELETS

In this section, we review some basic definitions from spectral graph theory [7] and the construction of the spectral graph wavelets introduced in [8]. We again start with a model of the data X as an undirected, weighted graph $G = \{\mathcal{E}, V, w\}$ with $|V| = N$, and assume that the graph is connected. The non-normalized Laplacian is defined as $\mathcal{L} := D - w$, where the off-diagonal elements of the degree matrix D are zeros, and the diagonal element of D corresponding to the degree of each vertex is the sum of the weights of all the edges incident to it.

We denote the complete set of orthonormal eigenvectors of $\mathcal{L}$ and their associated real eigenvalues by χ_ℓ and λ_ℓ for $\ell = 0, \ldots, N-1$. Without loss of generality, we assume the eigenvalues of the Laplacian of the connected graph to be ordered as $0 = \lambda_0 < \lambda_1 \leq \lambda_2 ... \leq \lambda_{N-1} := \lambda_{\max}$. The graph Fourier transform $\hat{f}$ of a function $f \in \mathbb{R}^N$ on the vertices of G is defined by $\hat{f}(\ell) := \langle \chi_\ell, f \rangle = \sum_{n=1}^{N} \chi_\ell^*(n) f(n)$, and the inverse transform is given by $f(n) = \sum_{\ell=0}^{N-1} \hat{f}(\ell)\chi_\ell(n)$.

The spectral graph wavelet transform [8] is generated by wavelet operators that are operator-valued functions of the Laplacian. The transform is determined by the choice of a kernel function $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, which is analogous to the Fourier transform of a wavelet in the classical setting. This kernel g behaves as a band-pass filter, satisfying $g(0) = 0$ and $\lim_{x\rightarrow\infty} g(x) = 0$. The wavelet operator $T_g = g(\mathcal{L})$ is defined through its action on a given function f as $\widehat{T_g f}(\ell) := g(\lambda_\ell)\hat{f}(\ell)$. The wavelet operator at scale t is then defined by $T_g^t = g(t\mathcal{L})$.

To form the spectral graph wavelets, we localize the wavelet operators at different scales by applying them to the impulse on a single vertex. That is, $\psi_{t,n} := T_g^t \delta_n$, or, equivalently, $\psi_{t,n}(m) = \sum_{\ell=0}^{N-1} g(t\lambda_\ell)\chi_\ell^*(n)\chi_\ell(m)$. The wavelet coefficients of a function f are computed by taking the inner products with the wavelets:

$$W_f(t,n) = \langle \psi_{t,n}, f \rangle = \left(T_g^t f\right)(n) = \sum_{\ell=0}^{N-1} g(t\lambda_\ell)\hat{f}(\ell)\chi_\ell(n).$$

The spectral graph wavelet transform also includes a second class of waveforms called scaling functions, which are analogous to the low-pass residual scaling functions from classical wavelet analysis. Introduced to stably represent the low frequency content of signals defined on the vertices, they are constructed in a manner analogous to the wavelets, with the scaling function at vertex n defined by $\phi_n := T_h\delta_n = h(\mathcal{L})\delta_n$. The scaling function generator $h : \mathbb{R}^+ \rightarrow \mathbb{R}$ acts as a low-pass filter, satisfying $h(0) > 0$ and $\lim_{x\rightarrow\infty} h(x) = 0$.

To summarize, given a fixed set of wavelet scales $\{t_j\}_{j=1}^J$ and the wavelet and scaling generators g and h, the spectral graph wavelet transform is a linear map $W : \mathbb{R}^N \rightarrow \mathbb{R}^{N(J+1)}$ defined by $Wf = \left((T_h f)^T, (T_g^{t_1} f)^T, \cdots, (T_g^{t_J} f)^T\right)^T$.

3. REGULARIZATION

3.1. Promoting Sparsity

An important property of the spectral graph wavelets is that their localization at small scales is guaranteed by simple constraints on the kernel g. This property ensures that the graph wavelets behave like discrete multiscale differential operators on graphs, and thus can sparsely approximate piecewise smooth signals. The scaling coefficients, on the other hand, are not expected to be sparse, as they represent the smoothed signal. Therefore, one method to determine the labels for the unlabeled data is to incorporate the sparse prior on the wavelet coefficients into the following regularization problem, which is a weighted version of lasso or basis pursuit denoising:

$$\min_{\alpha=[\alpha^S;\, \alpha^D]} \frac{1}{2}\|f - MW^*\alpha\|_2^2 + \lambda\|\alpha^D\|_1. \qquad (2)$$

In (2), $f \in \mathbb{R}^N$ is a column vector with the labels y_i at all locations where these labels are available, and zeros elsewhere; $M \in \mathbb{R}^{N\times N}$ is a matrix that has 1s on all diagonal elements corresponding to the locations of the labels, and zeros elsewhere; $W^* \in \mathbb{R}^{N\times N(J+1)}$ is the adjoint of the wavelet transform; $\alpha^S \in \mathbb{R}^N$ represents the scaling coefficients; and $\alpha^D \in \mathbb{R}^{NJ}$ represents the wavelet coefficients. The reconstructed labels are given by $W^*\alpha_*$, where α_* minimizes (2).

Intuitively, from a sparse approximation theory point of view, we can view problem (2) as trying to represent the signal f as a sparse linear superposition of atoms from the "masked dictionary" comprising the columns of MW^*. Unfortunately, the support of many of the high frequency wavelets (that is, those wavelets associated with small t_j's) is contained within vertices of the graph associated with unlabeled data. Therefore, the associated columns of MW^* are vectors of zeros, and these wavelets are not useful in synthesizing the signal. Empirical results confirm this intuition, and the wavelet coefficients associated with the high frequency wavelets are usually set to zero in (2). As a result, the above method tends to work best on problem instances that are also well-suited to existing regularization methods (i.e., where the label functions are globally smooth).

3.2. Using Structured Sparsity to Recover High Frequency Components

By design, wavelets whose support overlaps a discontinuity will have high coefficients. Therefore, a discontinuity in the label values induces a block of high wavelet coefficients *at all scales* at vertices close to the discontinuity. Based on this *structured* *sparsity*, one way to more accurately recover the high frequency components of the signal is to enforce sparsity across spatial locations, but persistence across scales at the same location. By persistence, we mean that if a wavelet coefficient at one scale at a fixed location is non-zero (active), then the wavelet coefficients at the other scales at that same location are also likely to be non-zero. The weighted mixed norm $\|\cdot\|_{\tau;p,q}$ discussed in [9, 10] provides the mathematical tool required to do this. Let $x = \{x_{k,l}\}_{k\in 1,2,\ldots,K;\; l\in 1,2,\ldots,L}$ be a doubly-indexed sequence of coefficients comprising K groups of L coefficients per group. Then the weighted mixed norm of x with weights τ is given by

$$\|x\|_{\tau;p,q} := \left(\sum_{k=1}^{K}\left(\sum_{l=1}^{L}\tau_{k,l}\,|x_{k,l}|^{p}\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}. \qquad (3)$$

To promote sparsity across the K groups and persistence within each group of L coefficients, we can take $q < 2$ and $p \geq 2$ in (3). Thus, to better recover the higher frequency components of the masked signal, we propose to solve the regularization problem

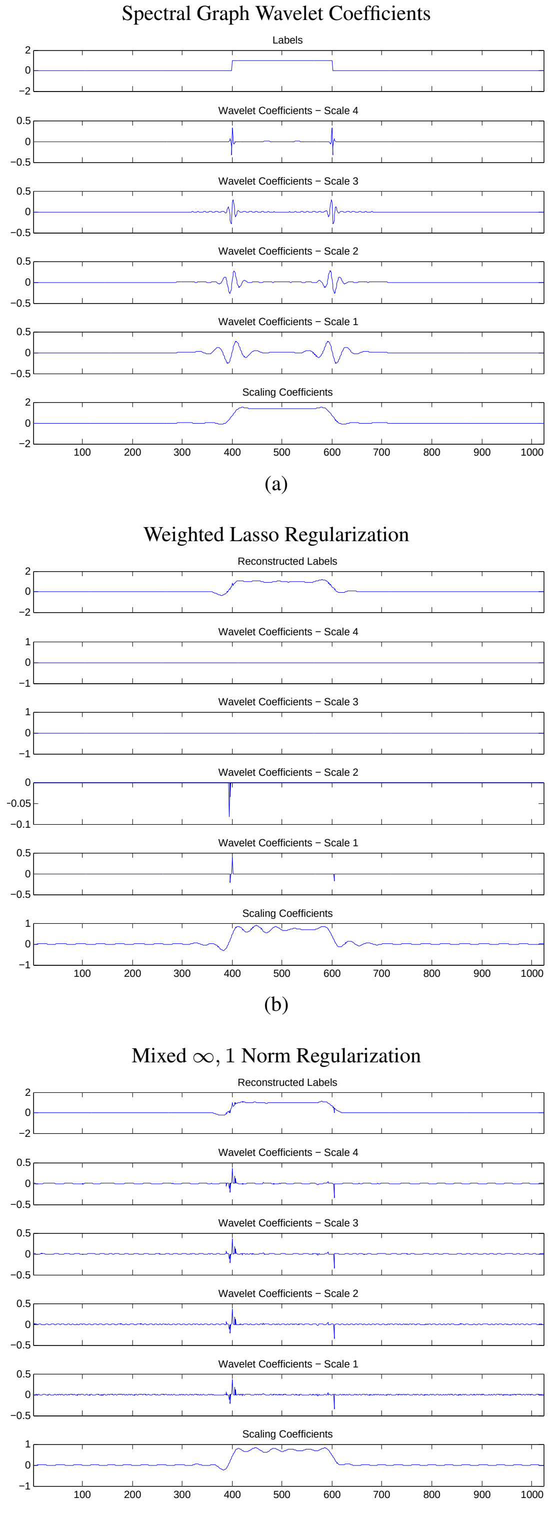
$$\min_{\alpha=[\alpha^S;\,\alpha^D]} \quad \|\alpha^D\|_{\tau;p^D,1} \quad \text{s.t.} \quad \|f - MW^*\alpha\|_2^2 \leq \epsilon. \qquad (4)$$

If τ is a vector of ones, then $\|\alpha^D\|_{\tau;p^D,1} = \sum_{n=1}^{N}\|\alpha^{D,n}\|_{p^D}$, where $\alpha^{D,n} \in \mathbb{R}^J$ represents the wavelet coefficients at location n and all scales $\{t_j\}_{j=1}^{J}$. The parameter $p^D \geq 2$ corresponds to different priors on the distributions of the wavelet coefficients across all scales at the same vertex. Problem (4) can be solved by proximal splitting methods (see, e.g., [11]).

4. NUMERICAL ILLUSTRATIONS

In this section, we illustrate the reconstruction issues discussed in the previous section with two simple toy examples. In Example 1, we consider one dimensional data: $X = \{1, 2, \ldots, 1024\}$, y_i equals 1 if $400 \leq x_i \leq 600$ and 0 otherwise, and 154 (15%) of the data points are selected at random to be the labeled set. We construct the weighted graph based on the thresholded Gaussian kernel weighting function: $w_{u,v} = e^{-\frac{[x_u - x_v]^2}{2\sigma^2}}$ if $\|x_u - x_v\| \leq \kappa$ and $w_{u,v} = 0$ otherwise. We let $\sigma = 2.0$ and $\kappa = 3.1$. We consider a spectral graph wavelet transform W with $J = 4$ wavelet scales in addition to the scaling functions. All wavelet design parameters are set to the defaults of the spectral graph wavelet toolbox.[2]

Figure 1(a) shows the spectral graph wavelet coefficients, WY, of the full set of target labels Y (assuming they were all known). The key takeaway is that the wavelet coefficients are sparse, with the active coefficients located around the discontinuities in the label sequence. Figure 1(b) shows the wavelet and scaling coefficients recovered by the weighted lasso regularization (2) with $\lambda = 0.1$. This method tends to only recover large scale wavelet coefficients, as the finer scale wavelets are largely masked by the matrix M.[3] Figure 1(c) shows the wavelet and



(a)

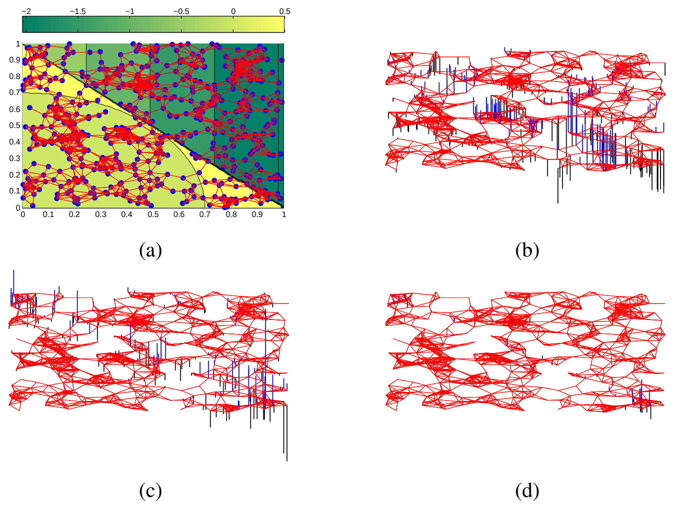
(b)

(c)

Fig. 1. Example 1. (a) The spectral graph wavelet coefficients of the unmasked labels Y. (b) The optimal coefficients and reconstructed labels from the weighted lasso (2). (c) The optimal coefficients and reconstructed labels from the mixed norm regularization problem (4) with $p^D = \infty$.

[2] Available at http://wiki.epfl.ch/sgwt

[3] Note that the more localized, finer scale wavelets are actually indexed by higher scale numbers. In this example, the finest scale wavelets are at scale 4.



(a)

(b)

(c)

(d)

Fig. 2. Example 2. (a) The label values, which are piecewise smooth with a discontinuity along $x_2 = 1 - x_1$. (b)-(d) The wavelet coefficients of the full label signal on the graph at scales 1, 2, and 3, respectively, are clustered around the discontinuity.

scaling coefficients recovered by the mixed norm regularization (4) with $p^D = \infty$, τ a vector of ones, and $\epsilon = 10^{-4}$. Note that the $\infty, 1$ mixed norm objective promotes a uniform distribution of wavelet coefficients across all scales at the same location. We repeated the experiment 50 times with different random label patterns each time. The average mean-square errors from the weighted lasso, the mixed norm regularization problem, and the global smoothness-promoting interpolated regularization algorithm of [5] were 0.0079, 0.0062, and 0.0048, respectively.

In Example 2, we consider 500 vertices placed randomly in the $[0, 1] \times [0, 1]$ square. The graph and wavelet constructions are the same as Example 1, with $\sigma = .074$, $\kappa = .075$, and $J = 3$. 25 (5%) of the data points are selected randomly to be the labeled set. The labels, shown in Figure 2(a), are given by

$$y_i = \begin{cases} x_{i,1}^2 + x_{i,2}^2 & \text{if} \quad x_{i,2} < 1 - x_{i,1} \\ -2x_{i,1} & \text{otherwise} \end{cases} .$$

Here, $x_{i,1}$ and $x_{i,2}$ are the coordinates of x_i in the square. Figures 2(b)-2(d) show the wavelet coefficients of Y at different scales, and we see that they are once again clustered around the discontinuities in the labels. We consider the same three reconstruction methods as Example 1, except that we use $\lambda = 0.3$ for the weighted lasso. We repeated the experiment 20 times with different random graphs and label patterns each time. The average mean-square errors were 0.325 for the weighted lasso, 0.317 for the mixed norm regularization problem, and 0.283 for the interpolated regularization algorithm of [5].

5. DISCUSSION

In test problems on larger examples and standard databases, the label prediction performance of the proposed method (4) is competitive with methods based on global smoothness priors (sometimes slightly better, sometimes slightly worse, depending on the data set, method of graph construction, parameter selection, etc.). However, this is somewhat disappointing due to the significant additional complexity of the proposed spectral graph wavelet method. The core issue is that we do not yet fully understand the best way to leverage the structured sparsity of the spectral graph wavelet transform to fill in the high frequency information that is masked out by the matrix M. Other reconstruction options we continue to investigate include: i) incorporating persistence within groups of coefficients across neighboring locations at the same scale (i.e., if a wavelet coefficient at a given scale and location is active, the coefficients at the same scale at neighboring locations in the underlying graph should also be active); ii) making the group definitions depend on the locations of the labeled data, so that the mixed norm penalty specifically promotes persistence in the neighborhoods most affected by the mask M; and iii) incorporating different penalization weights at different scales by adjusting τ in (4).

6. REFERENCES

[1] X. Zhu, “Semi-supervised learning literature survey,” *Technical Report TR-1530, University of Wisconsin-Madison Department of Computer Sciences*, 2005.

[2] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, 2006.

[3] X. Zhu and Z. Ghahramani, “Semi-supervised learning using Gaussian fields and harmonic functions,” in *Proc. International Conference on Machine Learning*, Washington, D.C., 2003, pp. 912–919.

[4] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, “Learning with local and global consistency,” in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds. 2004, pp. 321–328, MIT Press.

[5] M. Belkin, I. Matveeva, and P. Niyogi, “Regularization and semi-supervised learning on large graphs,” *Learning Theory, Lecture Notes in Comptuer Science*, pp. 624–638, 2004.

[6] M. Gavish, B. Nadler, and R. R. Coifman, “Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning,” in *Proc. International Conference on Machine Learning*, Haifa, Israel, 2010.

[7] F. K. Chung, *Spectral Graph Theory*, Vol. 92 of the CBMS Regional Conference Series in Mathematics, AMS Bokstore, 1997.

[8] D. K. Hammond, P. Vandergheynst, and R. Gribonval, “Wavelets on graphs via spectral graph theory,” *Applied and Computational Harmonic Analysis*, vol. 30, pp. 129–150, March 2011.

[9] M. Kowalski, “Sparse regression using mixed norms,” *Applied and Computational Harmonic Analysis*, vol. 27, pp. 303–324, November 2009.

[10] M. Kowalski and B. Torrésani, “Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients,” *Signal, Image and Video Processing*, vol. 3, pp. 251–264, 2009.

[11] 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. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. 2011, Springer-Verlag.