Active Semi-supervised Learning Using Sampling Theory for Graph Signals

Akshay Gadde, Aamir Anis and Antonio Ortega

University of Southern California

Nov 12, 2014
Motivation and Problem Definition

▶ Unlabeled data is abundant. Labeled data is expensive and scarce.
▶ Solution: **Active Semi-supervised Learning** (SSL).

▶ **Problem setting**: Offline, pool-based, batch-mode active SSL via graphs

1. *How to predict unknown labels from the known labels?*
2. *What is the optimal set of nodes to label given the learning algorithm?*
Graph Signal Processing

- **Graph** $G = (\mathcal{V}, \mathcal{E})$ with $N$ nodes
- nodes $\equiv$ data points; $w_{ij}$: similarity between $i$ and $j$.

- Adjacency matrix $\mathbf{W} = [w_{ij}]_{n \times n}$.
- Degree matrix $\mathbf{D} = \text{diag}\{\sum_j w_{ij}\}$.
- Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
- Normalized Laplacian $\mathcal{L} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$.
Graph Signal Processing

- **Graph** $G = (\mathcal{V}, \mathcal{E})$ with $N$ nodes
- nodes $\equiv$ data points; $w_{ij}$: similarity between $i$ and $j$.

- Adjacency matrix $\mathbf{W} = [w_{ij}]_{n \times n}$.
- Degree matrix $\mathbf{D} = \text{diag}\{\sum_j w_{ij}\}$.
- Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
- Normalized Laplacian $\mathcal{L} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$.

- **Graph signal** $f : \mathcal{V} \rightarrow \mathbb{R}$, denoted as $f \in \mathbb{R}^N$.
- Class membership functions are graph signals.

$$f^c(j) = \begin{cases} 1, & \text{if node } j \text{ is in class } c \\ 0, & \text{otherwise} \end{cases}$$
Notion of Frequency for Graph Signals

Spectrum of $\mathcal{L}$ provides frequency interpretation:

- $\lambda_k \in [0, 2]$: graph frequencies.
- $u_k$: graph Fourier basis.

- Fourier coefficients of $f$: $\tilde{f}(\lambda_i) = \langle f, u_i \rangle$.
- Graph Fourier Transform (GFT):
  $$\tilde{f} = U^T f.$$
Bandlimited Signals on Graphs

- **ω-bandlimited signal**: GFT has support [0, ω].
- **Paley-Wiener space** $PW_ω(G)$: Space of all $ω$-bandlimited signals.
  - $PW_ω(G)$ is a subspace of $\mathbb{R}^N$.
  - $ω_1 \leq ω_2 \Rightarrow PW_{ω_1}(G) \subseteq PW_{ω_2}(G)$.

- **Bandwidth of a signal**:
  \[
  ω(f) = \arg \max_\lambda \tilde{f}(λ) \text{ s.t. } |\tilde{f}(λ)| \geq 0
  \]
Bandlimited Signals on Graphs

- **ω-bandlimited signal:** GFT has support $[0, \omega]$.
- **Paley-Wiener space** $\text{PW}_\omega(G)$: Space of all $\omega$-bandlimited signals.
  - $\text{PW}_\omega(G)$ is a subspace of $\mathbb{R}^N$.
  - $\omega_1 \leq \omega_2 \Rightarrow \text{PW}_{\omega_1}(G) \subseteq \text{PW}_{\omega_2}(G)$.  

- **Bandwidth of a signal:**
  
  $\omega(f) = \arg \max \tilde{f}(\lambda) \text{ s.t. } |\tilde{f}(\lambda)| \geq 0$

- **Class membership functions can be approximated by bandlimited graph signals.**

![CDF of energy in GFT coefficients](image1)

![CDF of energy in GFT coefficients](image2)

(a) USPS  
(b) Isolet  
(c) 20 newsgroups
Sampling Theory for Graph Signals

Sampling theorem: bandwidth $\omega \Leftrightarrow$ sampling rate for unique representation
Sampling Theory for Graph Signals

Sampling theorem: bandwidth $\omega \Leftrightarrow$ sampling rate for unique representation

Sampling theory for graph signals:

**P1:** Maximum $\omega$, given $S$

**P2:** Smallest $S$, given $\omega$

**P3:** Estimate $\hat{f}$, given $\omega$, $f(S)$

Bandlimited signal $f(n)$

Down-sampling

Reconstruction

$\frac{-\pi}{M} \rightarrow \frac{\pi}{M}$

Reconstruction

$\hat{f}(n)$

$\textbf{USC}$

PCSJ/IMPS 2014

Active SSL using sampling theory for graph signals

6 / 21
Relevance of Sampling Theory to Active SSL

Active Semi-supervised Learning

Class labels

Criterion function

Select points to label based on criterion function

Predict unknown labels from known

Bandlimited signals

Cut-off frequency for given sampling set

Choose sampling set that maximizes cut-off frequency

Reconstruct bandlimited signal from sample values

Graph Signal Sampling
P1: Cut-off Frequency

How “smooth” the label set information have to be to reconstruct from $S$?

Condition for unique sampling of $PW_\omega(G)$ on $S$

Let $L_2(S^c) = \{\phi : \phi(S) = 0\}$. Then, we need $PW_\omega(G) \cap L_2(S^c) = \{0\}$. 
P1: Cut-off Frequency

How “smooth” the label set information have to be to reconstruct from $S$?

Let $L^2(S^c) = \{\phi : \phi(S) = 0\}$. Then, we need $PW_\omega(G) \cap L^2(S^c) = \{0\}$.

**Condition for unique sampling of $PW_\omega(G)$ on $S$**

**Sampling Theorem**

$f$ can be perfectly recovered from $f(S)$ iff

$$\omega(f) \leq \omega_c(S) \triangleq \inf_{\phi \in L^2(S^c)} \omega(\phi)$$

- Cut-off frequency = smallest bandwidth that a $\phi \in L^2(S^c)$ can have.
P1: Computing the Cut-off Frequency for Given $S$

Approximate bandwidth of a signal $\omega_k(f) = (f^T L_k f f^T f)^{1/k}$, where $k \in \mathbb{Z}^+$. 

Monotonicity: $\forall f, k_1 < k_2 \Rightarrow \omega_{k_1}(f) \leq \omega_{k_2}(f)$.

Convergence: $\lim_{k \to \infty} \omega_k(f) = \omega(f)$.

Minimize approximate bandwidth over $L_2(S_c)$ to estimate cut-off frequency $\Omega_k(S) = \min_{\phi \in L_2(S_c)} \omega_k(\phi) = \min_{\phi}: \phi(S) = 0$. 

Let $\{\sigma_1, k, \psi_1, k\} \to$ smallest eigen-pair of $(L_k S_c)$. 

Estimated cutoff frequency $\Omega_k(S) = (\sigma_1, k)$, $\phi_{opt_k}(S) = 0$. 

Corresponding smoothest signal $\phi_{opt_k}(S_c) = \psi_1, k$, $\phi_{opt_k}(S) = 0$. 

Rayleigh quotient.
Approximate bandwidth of a signal

\[ \omega_k(f) \triangleq \left( \frac{f^\top \mathcal{L}^k f}{f^\top f} \right)^{1/k} \], \text{ where } k \in \mathbb{Z}^+ 

- **Monotonicity:** \( \forall f, k_1 < k_2 \Rightarrow \omega_{k_1}(f) \leq \omega_{k_2}(f) \).
- **Convergence:** \( \lim_{k \to \infty} \omega_k(f) = \omega(f) \).
P1: Computing the Cut-off Frequency for Given $S$

Approximate bandwidth of a signal

$$
\omega_k(f) \triangleq \left( \frac{f^T \mathbf{L}^k f}{f^T f} \right)^{1/k}, \text{ where } k \in \mathbb{Z}^+
$$

- **Monotonicity:** $\forall f, k_1 < k_2 \Rightarrow \omega_{k_1}(f) \leq \omega_{k_2}(f)$.
- **Convergence:** $\lim_{k \to \infty} \omega_k(f) = \omega(f)$.

Minimize approximate bandwidth over $L_2(S^c)$ to estimate cut-off frequency

$$
\Omega_k(S) \triangleq \min_{\phi \in L_2(S^c)} \omega_k(\phi) = \min_{\phi: \phi(S) = 0} \left( \frac{\phi^T \mathbf{L}^k \phi}{\phi^T \phi} \right)^{1/k}
$$
P1: Computing the Cut-off Frequency for Given $S$

Approximate bandwidth of a signal

$$\omega_k(f) \triangleq \left( \frac{f^T \mathcal{L}^k f}{f^T f} \right)^{1/k}, \text{ where } k \in \mathbb{Z}^+$$

- **Monotonicity:** $\forall f, k_1 < k_2 \Rightarrow \omega_{k_1}(f) \leq \omega_{k_2}(f)$.
- **Convergence:** $\lim_{k \to \infty} \omega_k(f) = \omega(f)$.

Minimize approximate bandwidth over $L_2(S^c)$ to estimate cut-off frequency

$$\Omega_k(S) \triangleq \min_{\phi \in L_2(S^c)} \omega_k(\phi) = \min_{\phi: \phi(S) = 0} \left( \frac{\phi^T \mathcal{L}^k \phi}{\phi^T \phi} \right)^{1/k} = \left( \min_{\psi} \frac{\psi^T (\mathcal{L}^k)_{S^c} \psi}{\psi^T \psi} \right)^{1/k} \text{ (Rayleigh quotient)}$$

Let \( \{\sigma_{1,k}, \psi_{1,k}\} \rightarrow \) smallest eigen-pair of \((\mathcal{L}^k)_{S^c}\).

Estimated cutoff frequency \( \Omega_k(S) = (\sigma_{1,k})^{1/k} \),

Corresponding smoothest signal \( \phi^{\text{opt}}_k(S^c) = \psi_{1,k}, \phi^{\text{opt}}_k(S) = 0 \).
P2: Sampling Set Selection

- Optimal sampling set should maximally capture signal information.
- $S_{\text{opt}} = \arg\max_{|S|=m} \Omega_k(S) \rightarrow \text{combinatorial!}$
P2: Sampling Set Selection

- Optimal sampling set should maximally capture signal information.
- $S_{opt} = \arg \max_{|S|=m} \Omega_k(S) \rightarrow \text{combinatorial!}$

- Greedy gradient-based approach.
  - Start with $S = \{\emptyset\}$.
  - Add nodes one by one while ensuring maximum increase in $\Omega_k(S)$.

\[
(\Omega_k(S))^k = \min_{\phi(S)=0} \frac{\phi^T \mathcal{L}^k \phi}{\phi^T \phi} \approx \min_{x} \left( \frac{x^T \mathcal{L}^k x}{x^T x} + \alpha \frac{x^T \text{diag}(t)x}{x^T x} \right) \bigg|_{t=1_S} = \lambda_k^\alpha(t)|_{t=1_S}
\]

- relax the constraint

\[
\frac{d\lambda_k^\alpha(t)}{dt(i)} \bigg|_{t=1_S} \approx \alpha(\phi_k^{opt}(i))^2.
\]
P2: Sampling Set Selection

- Optimal sampling set should maximally capture signal information.
- $S_{\text{opt}} = \arg \max_{|S|=m} \Omega_k(S) \rightarrow \text{combinatorial!}$

- Greedy gradient-based approach.
  - Start with $S = \{\emptyset\}$.
  - Add nodes one by one while ensuring maximum increase in $\Omega_k(S)$.

\[
(\Omega_k(S))^k = \min_{\phi(S)=0} \frac{\phi^\top L^k \phi}{\phi^\top \phi} \approx \min_x \left( \frac{x^\top L^k x}{x^\top x} + \alpha \frac{x^\top \text{diag}(t)x}{x^\top x} \right) \bigg|_{t=1_S} = \lambda_k^\alpha(t)|_{t=1_S}
\]

\[
\frac{d\lambda_k^\alpha(t)}{dt(i)} \bigg|_{t=1_S} \approx \alpha(|\phi_k^{\text{opt}}(i)|^2).
\]

Greedy algorithm

$S \leftarrow S \cup \nu$, where $\nu = \arg \max_j (\phi^{\text{opt}}_k(j))^2$
Connection with Active Learning

- Cut-off function $\Omega_k(S) \equiv$ variation of smoothest signal in $L_2(S^c)$.
- Larger cut-off function $\Rightarrow$ more variation in $\phi_{opt} \Rightarrow$ more cross-links.

![Diagram showing cut-off functions and cross-links]

**Intuition**

Unlabeled nodes are strongly connected to labeled nodes!
\( C_1 = \{ x : x(S) = f(S) \} \) and \( C_2 = PW_\omega(G) \).

We need to find a unique \( f \in C_1 \cap C_2 \Rightarrow \) sampling theorem guarantees uniqueness.

Projection onto convex sets

\[
f_{i+1} = P_{C_2} P_{C_1} f_i, \text{ where } f_0 = [f(S)^T, 0]^T.
\]
P3: Label Prediction as Signal Reconstruction

- $C_1 = \{ x : x(S) = f(S) \}$ and $C_2 = PW_\omega(G)$.
- We need to find a unique $f \in C_1 \cap C_2 \Rightarrow$ sampling theorem guarantees uniqueness.

Projection onto convex sets

$f_{i+1} = P_{C_2} P_{C_1} f_i$, where $f_0 = [f(S)^T, 0]^T$.

- $P_{C_1}$ resets the samples on $S$ to $f(S)$.

- $P_{C_2} = U h(\Lambda) U^T$ sets $\tilde{f}(\lambda) = 0$ if $\lambda > \omega$.

$$h(\lambda) = \begin{cases} 1, & \text{if } \lambda \leq \omega \\ 0, & \text{if } \lambda \geq \omega \end{cases}$$
\[ C_1 = \{ x : x(S) = f(S) \} \text{ and } C_2 = PW_\omega(G). \]

- We need to find a unique \( f \in C_1 \cap C_2 \Rightarrow \) sampling theorem guarantees uniqueness.

**Projection onto convex sets**

\[ f_{i+1} = P_{C_2} P_{C_1} f_i, \text{ where } f_0 = [f(S)^T, 0]^T. \]

- \( P_{C_1} \) resets the samples on \( S \) to \( f(S) \).

- \( P_{C_2} = U h(\Lambda) U^T \) sets \( \tilde{f}(\lambda) = 0 \) if \( \lambda > \omega \).

\[
h(\lambda) = \begin{cases} 
1, & \text{if } \lambda < \omega \\
0, & \text{if } \lambda \geq \omega 
\end{cases}
\]

- \( P_{C_2} \approx \sum_{i=1}^n \left( \sum_{j=0}^p a_j \lambda_i^j \right) u_i u_i^T = \sum_{j=0}^p a_j \mathcal{L}^j \rightarrow p\text{-hop localized} \)

Predicted class of node \( n = \arg\max_c f^c(n) \).
Summary of the Algorithm

Input data

Construct graph

Choose nodes to label by maximizing cut-off frequency

Query labels of chosen nodes

Predict labels by signal reconstruction

Input: $G = \{V, E\}$, $L$, target size $m$, parameter $k \in \mathbb{Z}^+$
Initialize: $S = \{\emptyset\}$
while $|S| \leq m$
    For $S$, compute the smoothest signal $\phi_k^{\text{opt}} \in L_2(S^c)$
    $v \leftarrow \arg \max_i [(\phi_k^{\text{opt}}(i))^2]$
    $S \leftarrow S \cup v$
end while

POCS iteration: $f_{i+1} = P_{c_1} P_{c_2} f_i$
Label of node $n = \arg \max_c f^c(n)$
Related Work

Submodular optimization:
- Optimizing “strength” of a network ($\Psi$-max) [Guillory and Bilmes, 2011]
  - computationally complex
- Graph partitioning based heuristic (METIS) [Guillory and Bilmes, 2009]

Generalization error bound minimization:
- Minimizing generalization error bound for LLGC [Gu and Han, 2012]
  - contains a regularization parameter that needs to be tuned.

Optimal experiment design:
- Local linear reconstruction (LLR) [Zhang et al., 2011]
  - does not consider the learning algorithm
Results: Toy Example

Task

Pick 8 data points for labeling.
Results: Toy Example

Task
Pick 8 data points for labeling.

- 4 data points picked from each circle.
- Maximally separated points within one circle.
- Maximal spacing between selected data points in different circles.
Results: Real Datasets

- **USPS**: handwritten digits
  - $x_i = 16 \times 16$ image
  - number of classes = 10
  - $K$-NN graph with $K = 10$
  - $w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

- **ISOLET**: spoken letters
  - $x_i \in \mathbb{R}^{617}$ speech features
  - number of classes = 26
  - $K$-NN graph with $K = 10$
  - $w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

- **Newsgroups**: documents
  - $x_i \in \mathbb{R}^{3000}$ tf-idf of words
  - number of classes = 10
  - $K$-NN graph with $K = 10$
  - $w_{ij} = \frac{x_i^\top x_j}{\|x_i\| \|x_j\|}$
Results: Effect of $k$

Larger $k \Rightarrow$ better estimate of cut-off frequency is optimized.

![Graphs showing effect of k on accuracy and CDF of energy in GFT coefficients for USPS, Isolet, and 20 newsgroups datasets.](image-url)
Conclusion and Future Work

- Application of graph signal sampling theory to active SSL
  - Class labels $\Rightarrow$ bandlimited graph signals
  - Choosing nodes $\Rightarrow$ Best sampling set selection
  - Predicting unknown labels $\Rightarrow$ Signal reconstruction from samples

- Proposed approach gives significantly better results.

- Future work:
  - Approximate optimality of proposed sampling set selection.
  - Robustness against noise
A. Anis, A. Gadde, and A. Ortega.  
Towards a sampling theorem for signals on arbitrary graphs.  
In *ICASSP*, 2014.

S.K. Narang, A. Gadde, and A. Ortega.  
Localized iterative methods for interpolation in graph structured data.  

A. Guillory and J. Bilmes.  
Active semi-supervised learning using submodular functions.  
In *UAI*, 2011.

A. Guillory and J. Bilmes.  
Label selection on graphs.  
In *NIPS*, 2009.

Active learning based on locally linear reconstruction.  
*TPAMI*, 2011.

Q. Gu and J. Han.  
Towards active learning on graphs, an error bound minimization approach.  
In *ICDM*, 2012.
Thank you!
Label Complexity

- Let $\hat{f}$ be the reconstruction of $f$ obtained from its samples on $S$.
- What is the minimum number of labels required so that $\|f - \hat{f}\| \leq \delta$?

Smoothness of a signal

Let $P_\theta$ be the projector for $PW_\theta(G)$. Then $\gamma(f) = \min \theta \text{ s.t. } \|f - P_\theta f\| \leq \delta$.

Theorem

The minimum number of labels $|S|$ required to satisfy $\|f - \hat{f}\| \leq \delta$ is greater than $p$, where $p$ is the number of eigenvalues of $L$ less than $\gamma(f)$. 