Active Semi-supervised Learning Using Sampling Theory for Graph Signals

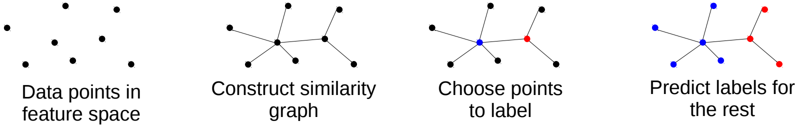
Akshay Gadde, Aamir Anis and Antonio Ortega

University of Southern California

August 26, 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

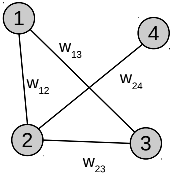


1. *How to predict unknown labels from the known labels?*
2. *What is the optimal set of nodes to label given the learning algorithm?*

Graph Signal Processing

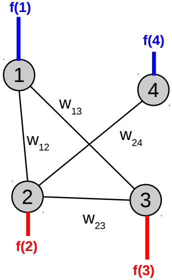
- **Graph** $G = (\mathcal{V}, \mathcal{E})$ with N nodes
- nodes $\equiv$ data points; w_{ij}: similarity between i and j.

- 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 $\boldsymbol{\mathcal{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.



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 $\boldsymbol{\mathcal{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.

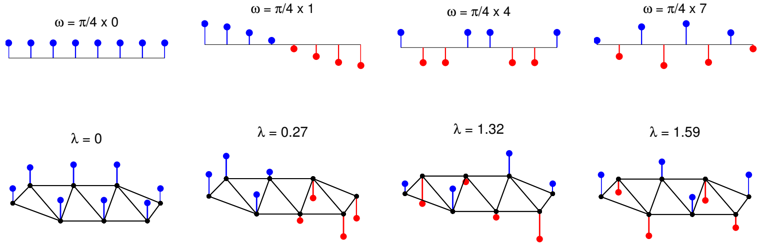
- **Graph signal** $f : \mathcal{V} \to \mathbb{R}$, denoted as $\mathbf{f} \in \mathbb{R}^N$.
- Class membership functions are graph signals.

$$\mathbf{f}^c(j) = \begin{cases} 1, & \text{if node } j \text{ is in class } c \\ 0, & \text{otherwise} \end{cases}$$

Notion of Frequency for Graph Signals

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

- $\lambda_k \in [0, 2]$: *graph frequencies*.
- $\mathbf{u}_k$: *graph Fourier basis*.



- *Fourier coefficients* of $\mathbf{f}$: $\tilde{\mathbf{f}}(\lambda_i) = \langle \mathbf{f}, \mathbf{u}_i \rangle$.
- *Graph Fourier Transform* (GFT):

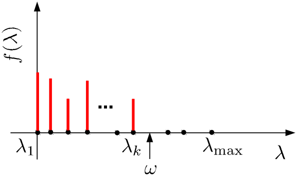
$$\tilde{\mathbf{f}} = \mathbf{U}^T \mathbf{f}.$$

Bandlimited Signals on Graphs

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

- **Bandwidth of a signal:**

$$\omega(\mathbf{f}) = \arg\max_\lambda \tilde{\mathbf{f}}(\lambda) \text{ s.t. } |\tilde{\mathbf{f}}(\lambda)| \geq 0$$

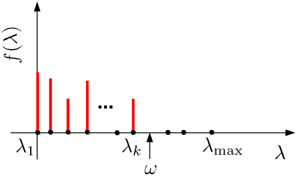


Bandlimited Signals on Graphs

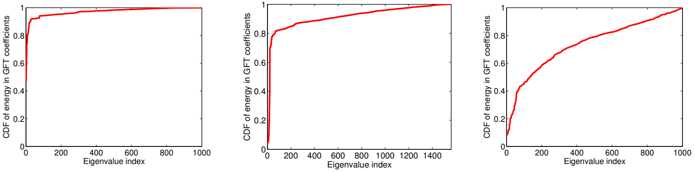
- ω-**bandlimited signal**: GFT has support $[0, \omega]$.
- **Paley-Wiener space** $PW_\omega(G)$: Space of all ω-bandlimited signals.
  - $PW_\omega(G)$ is a subspace of $\mathbb{R}^N$.
  - $\omega_1 \leq \omega_2 \Rightarrow PW_{\omega_1}(G) \subseteq PW_{\omega_2}(G)$.

- **Bandwidth of a signal:**

$$\omega(\mathbf{f}) = \arg\max_\lambda \tilde{\mathbf{f}}(\lambda) \text{ s.t. } |\tilde{\mathbf{f}}(\lambda)| \geq 0$$



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

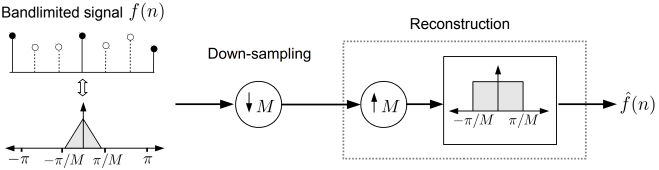


(a) USPS (b) Isolet (c) 20 newsgroups

USC

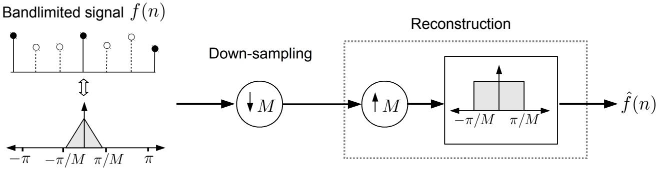
Sampling Theory for Graph Signals

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

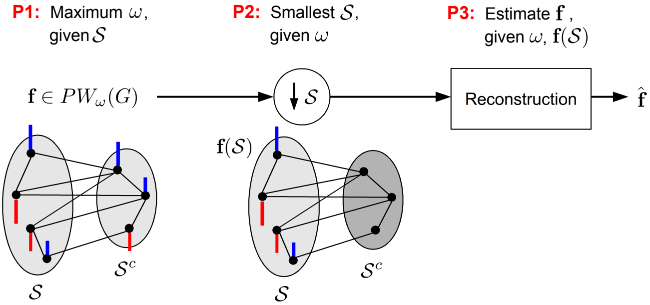


Sampling Theory for Graph Signals

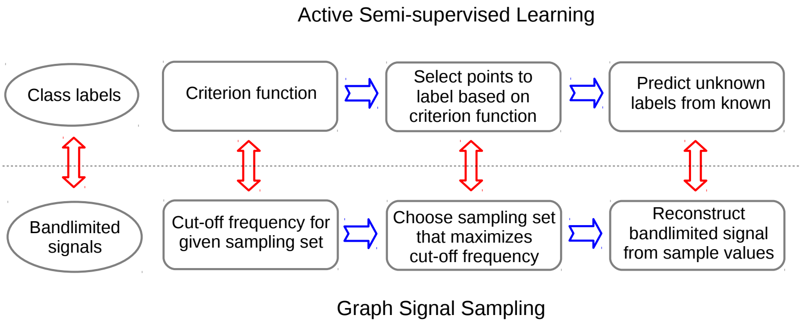
Sampling theorem: bandwidth $\omega \Leftrightarrow$ sampling rate for unique representation



Sampling theory for graph signals:

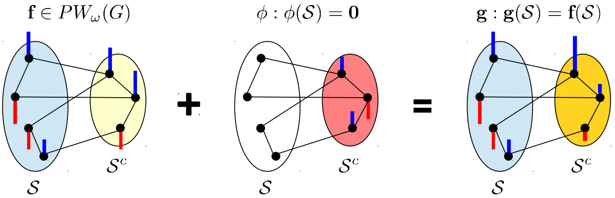


Relevance of Sampling Theory to Active SSL



P1: Cut-off Frequency

How "smooth" the label set information have to be to reconstruct from $\mathcal{S}$?

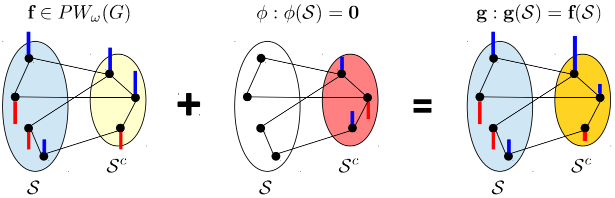


Condition for unique sampling of $PW_\omega(G)$ on $\mathcal{S}$

Let $L_2(\mathcal{S}^c) = \{\phi : \phi(\mathcal{S}) = \mathbf{0}\}$. Then, we need $PW_\omega(G) \cap L_2(\mathcal{S}^c) = \{\mathbf{0}\}$.

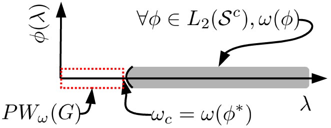
P1: Cut-off Frequency

How "smooth" the label set information have to be to reconstruct from $\mathcal{S}$?



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

Let $L_2(\mathcal{S}^c) = \{\phi : \phi(\mathcal{S}) = \mathbf{0}\}$. Then, we need $PW_\omega(G) \cap L_2(\mathcal{S}^c) = \{\mathbf{0}\}$.



Sampling Theorem

$\mathbf{f}$ can be perfectly recovered from $\mathbf{f}(\mathcal{S})$ iff

$$\omega(\mathbf{f}) \leq \omega_c(\mathcal{S}) \triangleq \inf_{\phi_{L_2(\mathcal{S}^c)}} \omega(\phi)$$

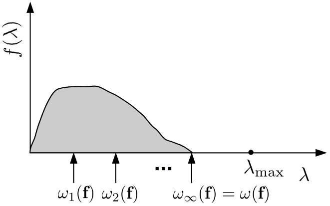
- Cut-off frequency = smallest bandwidth that a $\phi \in L_2(\mathcal{S}^c)$ can have.

P1: Computing the Cut-off Frequency for Given $\mathcal{S}$

P1: Computing the Cut-off Frequency for Given $\mathcal{S}$

Approximate bandwidth of a signal

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

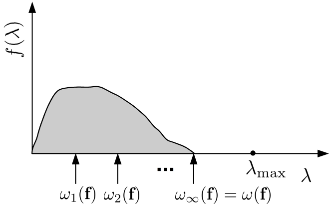


- *Monotonicity:* $\forall \mathbf{f}, k_1 < k_2 \Rightarrow \omega_{k_1}(\mathbf{f}) \leq \omega_{k_2}(\mathbf{f})$.
- *Convergence:* $\lim_{k\to\infty} \omega_k(\mathbf{f}) = \omega(\mathbf{f})$.

P1: Computing the Cut-off Frequency for Given $\mathcal{S}$

Approximate bandwidth of a signal

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



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

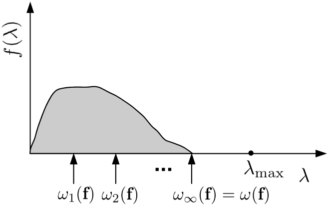
Minimize approximate bandwidth over $L_2(\mathcal{S}^c)$ to estimate cut-off frequency

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

P1: Computing the Cut-off Frequency for Given $\mathcal{S}$

Approximate bandwidth of a signal

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



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

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

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

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

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

Corresponding smoothest signal $\phi_k^{\text{opt}}(\mathcal{S}^c) = \psi_{1,k},\ \phi_k^{\text{opt}}(\mathcal{S}) = \mathbf{0}$.

P2: Sampling Set Selection

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

P2: Sampling Set Selection

- Optimal sampling set should maximally capture signal information.
- $\mathcal{S}_{\text{opt}} = \arg\max_{|\mathcal{S}|=m} \Omega_k(\mathcal{S}) \rightarrow$ combinatorial!
- Greedy gradient-based approach.
  - Start with $\mathcal{S} = \{\emptyset\}$.
  - Add nodes one by one while ensuring maximum increase in $\Omega_k(\mathcal{S})$.

binary relaxation

$$(\Omega_k(\mathcal{S}))^k = \min_{\phi(\mathcal{S})=\mathbf{0}} \frac{\phi^\top \mathcal{L}^k \phi}{\phi^\top \phi} \approx \min_{\mathbf{x}} \left(\frac{\mathbf{x}^\top \mathcal{L}^k \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} + \alpha \frac{\mathbf{x}^\top \text{diag}(\mathbf{t})\mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \right) \Bigg|_{\mathbf{t}=\mathbf{1}_{\mathcal{S}}} = \lambda_k^\alpha(\mathbf{t})|_{\mathbf{t}=\mathbf{1}_{\mathcal{S}}}$$

relax the constraint

- $\left.\frac{d\lambda_\alpha^k(\mathbf{t})}{d\mathbf{t}(i)}\right|_{\mathbf{t}=\mathbf{1}_{\mathcal{S}}} \approx \alpha(\phi_k^{\text{opt}}(i))^2.$

P2: Sampling Set Selection

- Optimal sampling set should maximally capture signal information.
- $\mathcal{S}_{\text{opt}} = \arg\max_{|\mathcal{S}|=m} \Omega_k(\mathcal{S}) \rightarrow$ combinatorial!
- Greedy gradient-based approach.
  - Start with $\mathcal{S} = \{\emptyset\}$.
  - Add nodes one by one while ensuring maximum increase in $\Omega_k(\mathcal{S})$.

binary relaxation

$$(\Omega_k(\mathcal{S}))^k = \min_{\phi(\mathcal{S})=\mathbf{0}} \frac{\phi^\top \mathcal{L}^k \phi}{\phi^\top \phi} \approx \min_{\mathbf{x}} \left(\frac{\mathbf{x}^\top \mathcal{L}^k \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} + \alpha \frac{\mathbf{x}^\top \text{diag}(\mathbf{t}) \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \right) \Bigg|_{\mathbf{t}=\mathbf{1}_\mathcal{S}} = \lambda_k^\alpha(\mathbf{t})|_{\mathbf{t}=\mathbf{1}_\mathcal{S}}$$

relax the constraint

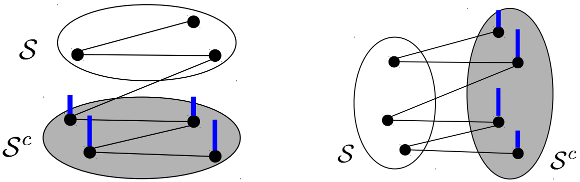
- $\left.\frac{d\lambda_\alpha^k(\mathbf{t})}{d\mathbf{t}(i)}\right|_{\mathbf{t}=\mathbf{1}_\mathcal{S}} \approx \alpha(\phi_k^{\text{opt}}(i))^2.$

Greedy algorithm

$\mathcal{S} \leftarrow \mathcal{S} \cup v$, where $v = \arg\max_j (\phi^{\text{opt}}(j))^2$

Connection with Active Learning

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

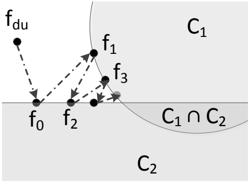


Intuition

Unlabeled nodes are strongly connected to labeled nodes!

P3: Label Prediction as Signal Reconstruction

- $\mathcal{C}_1 = \{\mathbf{x} : \mathbf{x}(\mathcal{S}) = \mathbf{f}(\mathcal{S})\}$ and $\mathcal{C}_2 = PW_\omega(G)$.
- We need to find a unique $\mathbf{f} \in \mathcal{C}_1 \cap \mathcal{C}_2 \Rightarrow$ sampling theorem guarantees uniqueness.

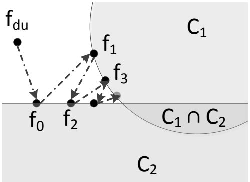


Projection onto convex sets

$\mathbf{f}_{i+1} = \mathbf{P}_{\mathcal{C}_2}\mathbf{P}_{\mathcal{C}_1}\mathbf{f}_i$, where $\mathbf{f}_0 = [\mathbf{f}(\mathcal{S})^\top, \mathbf{0}]^\top$.

P3: Label Prediction as Signal Reconstruction

- $\mathcal{C}_1 = \{\mathbf{x} : \mathbf{x}(\mathcal{S}) = \mathbf{f}(\mathcal{S})\}$ and $\mathcal{C}_2 = PW_\omega(G)$.
- We need to find a unique $\mathbf{f} \in \mathcal{C}_1 \cap \mathcal{C}_2 \Rightarrow$ sampling theorem guarantees uniqueness.



Projection onto convex sets

$\mathbf{f}_{i+1} = \mathbf{P}_{\mathcal{C}_2}\mathbf{P}_{\mathcal{C}_1}\mathbf{f}_i$, where $\mathbf{f}_0 = [\mathbf{f}(\mathcal{S})^\top, \mathbf{0}]^\top$.

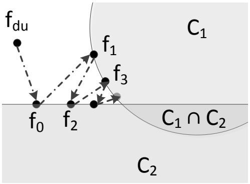
- $\mathbf{P}_{\mathcal{C}_1}$ resets the samples on $\mathcal{S}$ to $\mathbf{f}(\mathcal{S})$.
- $\mathbf{P}_{\mathcal{C}_2} = \mathbf{U}h(\mathbf{\Lambda})\mathbf{U}^\top$ sets $\tilde{\mathbf{f}}(\lambda) = 0$ if $\lambda > \omega$.

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

USC

P3: Label Prediction as Signal Reconstruction

- $\mathcal{C}_1 = \{\mathbf{x} : \mathbf{x}(\mathcal{S}) = \mathbf{f}(\mathcal{S})\}$ and $\mathcal{C}_2 = PW_\omega(G)$.
- We need to find a unique $\mathbf{f} \in \mathcal{C}_1 \cap \mathcal{C}_2 \Rightarrow$ sampling theorem guarantees uniqueness.

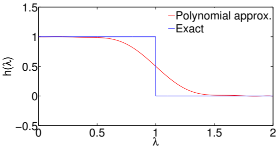


Projection onto convex sets

$\mathbf{f}_{i+1} = \mathbf{P}_{\mathcal{C}_2}\mathbf{P}_{\mathcal{C}_1}\mathbf{f}_i$, where $\mathbf{f}_0 = [\mathbf{f}(\mathcal{S})^\top, \mathbf{0}]^\top$.

- $\mathbf{P}_{\mathcal{C}_1}$ resets the samples on $\mathcal{S}$ to $\mathbf{f}(\mathcal{S})$.
- $\mathbf{P}_{\mathcal{C}_2} = \mathbf{U}h(\mathbf{\Lambda})\mathbf{U}^\top$ sets $\tilde{\mathbf{f}}(\lambda) = 0$ if $\lambda > \omega$.

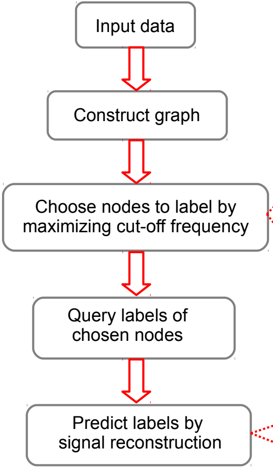
$$h(\lambda) = \begin{cases} 1, & \text{if } \lambda < \omega \\ 0, & \text{if } \lambda \geq \omega \end{cases}$$



- $\mathbf{P}_{\mathcal{C}_2} \approx \sum_{i=1}^{n} \left(\sum_{j=0}^{p} a_j \lambda_i^j\right) \mathbf{u}_i\mathbf{u}_i^\top = \sum_{j=0}^{p} a_j \boldsymbol{\mathcal{L}}^j \rightarrow p$-hop localized

Predicted class of node $n = \arg\max_c \mathbf{f}^c(n)$.

Summary of the Algorithm



Input: $G = \{\mathcal{V}, E\}, \mathbf{L}$, target size m, parameter $k \in \mathbb{Z}^+$

Initialize: $\mathcal{S} = \{\emptyset\}$

while $|\mathcal{S}| \leq m$

  For $\mathcal{S}$, compute the smoothest signal $\phi_k^{\text{opt}} \in L_2(\mathcal{S}^c)$

  $v \leftarrow \arg\max_i \left[(\phi_k^{\text{opt}}(i))^2\right]$

  $\mathcal{S} \leftarrow \mathcal{S} \cup v$

end while

POCS iteration: $\mathbf{f}_{i+1} = \mathbf{P}_{\mathcal{C}_1}\mathbf{P}_{\mathcal{C}_2}\mathbf{f}_i$

Label of node $n = \arg\max_c \mathbf{f}^c(n)$

Related Work

Submodular optimization:

- Optimizing "strength" of a network (Ψ-max) [Guillory and Bilmes, 2011]
  - computationally complex
- Graph partitioning based heuristic (METIS) [Guillory and Bilmes, 2009]

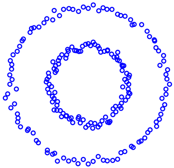
Generalization error bound minimization:

- 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

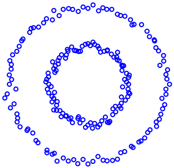
Results: Toy Example



Task

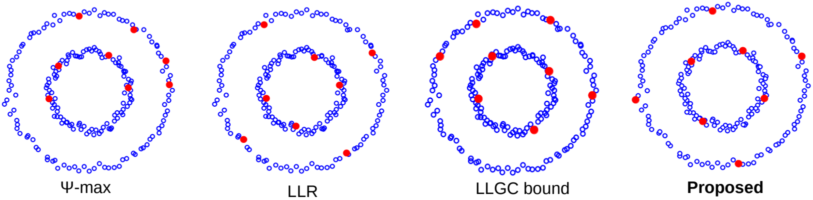
Pick 8 data points for labeling.

Results: Toy Example



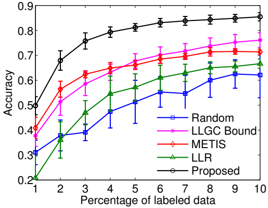
Task

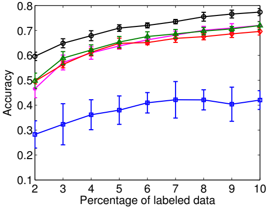
Pick 8 data points for labeling.

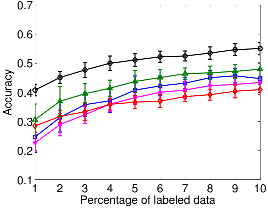


- 4 data points picked from each circle.
- Maximally separated points within one circle.
- Maximal spacing between selected data points in different circles.

Results: Real Datasets







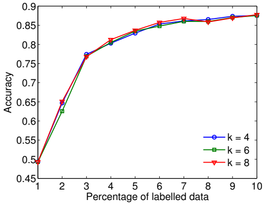
- USPS: handwritten digits
- $\mathbf{x}_i = 16 \times 16$ image
- number of classes $= 10$
- K-NN graph with $K = 10$
- $w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$

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

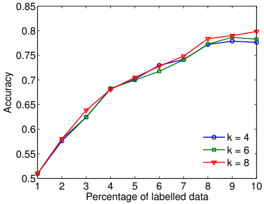
- Newsgroups: documents
- $\mathbf{x}_i \in \mathbb{R}^{3000}$ tf-idf of words
- number of classes $= 10$
- K-NN graph with $K = 10$
- $w_{ij} = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$

Results: Effect of k

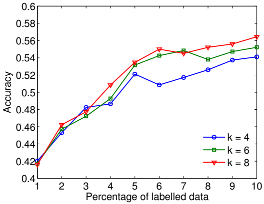
Larger $k \Rightarrow$ better estimate of cut-off frequency is optimized.



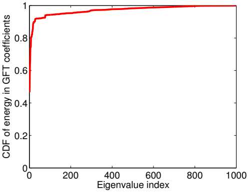
(a) USPS



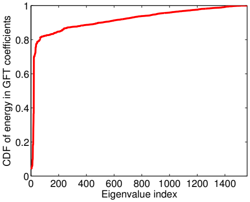
(b) Isolet



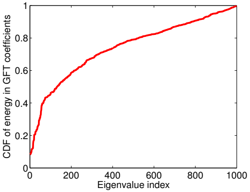
(c) 20 newsgroups



(a) USPS



(b) Isolet



(c) 20 newsgroups

Conclusion and Future Work

- Application of graph signal sampling theory to active SSL
  - Class labels ⇒ bandlimited graph signals
  - Choosing nodes ⇒ Best sampling set selection
  - Predicting unknown labels ⇒ Signal reconstruction from samples
- Proposed approach gives significantly better results.
- Future work:
  - Approximate optimality of proposed sampling set selection.
  - Robustness against noise

References

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.
In *IEEE GlobalSIP*, 2013.

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.

L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. Huang.
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 $\hat{\mathbf{f}}$ be the reconstruction of $\mathbf{f}$ obtained from its samples on $\mathcal{S}$.
- What is the minimum number of labels required so that $\|\mathbf{f} - \hat{\mathbf{f}}\| \leq \delta$?

Smoothness of a signal

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

Theorem

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