On the Supermodularity of Active Graph-based Semi-Supervised Learning with Stieltjes Matrix Regularization

Pin-Yu Chen* Dennis Wei*

IBM Thomas J. Watson Research Center, USA

Abstract

- Active Graph-based Semi-Supervised Learning (AG-SSL): actively select a small set of labeled examples and utilize their graph-based relation to other unlabeled examples to aid in machine learning / signal processing tasks
- A revisit to graph-based SSL formulation [1-3] with strategic node information querying and a performance guaranteed greedy algorithm
- Different from the perspective of graph signal processing and sampling, we dispense with any assumptions on the bandlimitedness of "graph signals"
- We formulate the problems of inferring a graph signal (labels) based on an incomplete set of noisy linear observations as well as selecting the set of observations to maximize the precision of this inference
- We prove that under a broad class of regularization functions parameterized by the family of **Stieltjes matrices**, the objective function is **supermodular** in the set $S \subset V$ of labeled examples

Background and Problem Formulation

Stieltjes matrix

- A real symmetric matrix $\mathbf{X}$ is said to be a (possibly singular) Stieltjes matrix if it is positive semidefinite and its off-diagonal entries are non-positive
- Examples: (unnormalized) graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, normalized graph Laplacian matrix $\mathbf{L}_N = \mathbf{D}^{-1/2}\mathbf{L}\,\mathbf{D}^{-1/2}$. $\mathbf{A}$: weight matrix. $\mathbf{D} = \mathbf{diag}(\mathbf{A1})$
- Inverse-positivity property: the inverse of a Stieltjes matrix is element-wise non-negative

Supermodularity

- A set function $f(S)$ is supermodular if $\delta_v(S) \geq \delta_v(T)$ for any $S, T = S \cup \{u\}, u \notin S$, and $v \notin T$, where $\delta_v(S) = f(S) - f(S \cup \{v\})$ is the decrease due to adding v to S – i.e., diminishing decrease in $f(S)$

Our model

- $\boldsymbol{x}$: the signal (i.e., labels) to be recovered on the graph $G = (V, E)$. $|V| = n$.
- Model $\boldsymbol{x}$ as a random signal with a multivariate Gaussian prior distribution

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Omega}_0\mathbf{x}\right)$$

- Ω_0: precision matrix with $Rank(\Omega_0) = n$ or $n-1$ (improper Gaussian)
- A total of m noisy linear observations of the form $\boldsymbol{y} = \mathbf{C}\boldsymbol{x} + \boldsymbol{n}$ are available to be taken. $\mathbf{C}$: fixed $m \times n$ measurement matrix. $\boldsymbol{n} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

Active Graph-based Semi-Supervised Learning (AG-SSL)

- Subset selection: select S with $|S| = s \leq m$ s.t. $\boldsymbol{y}_S = \mathbf{C}_S\boldsymbol{x} + \boldsymbol{n}_S$. $\mathbf{C}_S$: submatrix of $\mathbf{C}$ with rows indexed by S
- Posterior distribution $p(\mathbf{x} \mid \mathbf{y}_S) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\|\mathbf{y}_S - \mathbf{C}_S\mathbf{x}\|^2 + \mathbf{x}^T\boldsymbol{\Omega}_0\mathbf{x}\right)\right]$ is also Gaussian with precision matrix $\boldsymbol{\Omega}(S) = \boldsymbol{\Omega}_0 + \frac{1}{\sigma^2}\mathbf{C}_S^T\mathbf{C}_S$. Assume $\boldsymbol{\Omega}(S)$ to be non-singular for $S \neq \{\emptyset\}$
- MSE estimator: $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x} \mid \mathbf{y}_S] = \frac{1}{\sigma^2}\boldsymbol{\Sigma}(S)\mathbf{C}_S^T\mathbf{y}_S$. $\boldsymbol{\Sigma}(S) = \boldsymbol{\Omega}^{-1}(S)$: posterior covariance matrix
- The corresponding MSE is $f(S) = \mathbb{E}\left[\|\hat{\mathbf{x}} - \mathbf{x}\|^2 \mid \mathbf{y}_S\right] = \mathrm{tr}\,\boldsymbol{\Omega}(S)^{-1} = \mathrm{tr}\left(\boldsymbol{\Omega}_0 + \frac{1}{\sigma^2}\mathbf{C}_S^T\mathbf{C}_S\right)^{-1}$
- Goal of AG-SSL: select a set $S \subset V$ with $|S| = s$ to minimize $f(S)$

❖ Note: maximizing the posterior distribution is equivalent to solving the following minimization problem with $\alpha = \sigma^2$: $\text{minimize}_{\mathbf{x}} \|\mathbf{y}_S - \mathbf{C}_S\mathbf{x}\|^2 + \alpha\mathbf{x}^T\boldsymbol{\Omega}_0\mathbf{x}$

❖ This formulation coincides with the formulation of graph-based SSL [1-3]

❖ And its solution is also given by the MSE estimator (posterior mean)

Main Results and AG-SSL Algorithm

➢ **Lemma:** Assume that $\boldsymbol{\Omega}(S)$ is invertible. Then the decrease in f due to adding node v to S is given by

$$\delta_v(\mathcal{S}) = f(\mathcal{S}) - f(\mathcal{S} \cup \{v\}) = \frac{\left\|\boldsymbol{\Sigma}(\mathcal{S})\mathbf{C}_v^T\right\|^2}{\sigma^2 + \mathbf{C}_v\boldsymbol{\Sigma}(\mathcal{S})\mathbf{C}_v^T}.$$

➢ **Corollary:** The objective function $f(S)$ is a monotonically nonincreasing set function

- Note: for general $\mathbf{C}$ and $\boldsymbol{\Omega}_0$, it is possible to construct counterexamples where supermodularity does not hold
- In the remainder of the main results, we specialize to the case $\mathbf{C} = \mathbf{I}$ and assume that $\boldsymbol{\Omega}_0$ is a (possibly singular) Stieltjes matrix

➢ **Theorem:** The objective function $f(S)$ is supermodular if $\mathbf{C} = \mathbf{I}$ and $\boldsymbol{\Omega}_0$ is a (possibly singular) Stieltjes matrix

- **Greedy AG-SLL algorithm with performance guarantee:**

Algorithm 1 Greedy AG-SSL Algorithm

Input: Graph G, regularization $\boldsymbol{\Omega}_0$, σ, # of samples s
Output: sampling set $\mathcal{S}$ and predictor $\widehat{\mathbf{x}}(\mathcal{S})$
Initialization: $\mathcal{S} = \{\varnothing\}$
for $k = 1$ to s **do**
  Find $v^* = \arg\min_{v \in \mathcal{V}/\mathcal{S}} f(\mathcal{S} \cup v)$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{v^*\}$
end for
$\widehat{\mathbf{x}}(\mathcal{S}) = \sigma^{-2}\boldsymbol{\Sigma}(\mathcal{S})\mathbf{I}_{\mathcal{S}}^T\mathbf{y}_{\mathcal{S}}$

- **Corollary:** Let S be the sampling set obtained from Algorithm 1, S^* a minimizer of $f(S)$ over sets of size $|S| = s$, and S_1^* a minimizer of $f(S)$ over singletons (e.g. from first iteration of Algorithm 1). Then

$$\frac{f(\mathcal{S}) - f(\mathcal{S}^*)}{f(\mathcal{S}_1^*) - f(\mathcal{S}^*)} \leq \left(1 - \frac{1}{s}\right)^{s-1} \to \frac{1}{e}\ as\ s \to \infty.$$

AG-SSL on Community Detection and Comparative GSP Methods

- **Community detection:** partition the nodes into well-connected groups based on the connectivity structure of a graph
- **AG-SSL:** node label (membership) querying during community detection process
- **Datasets:** (1) Karate club network (34 nodes); (2) Dolphin network (62 nodes)

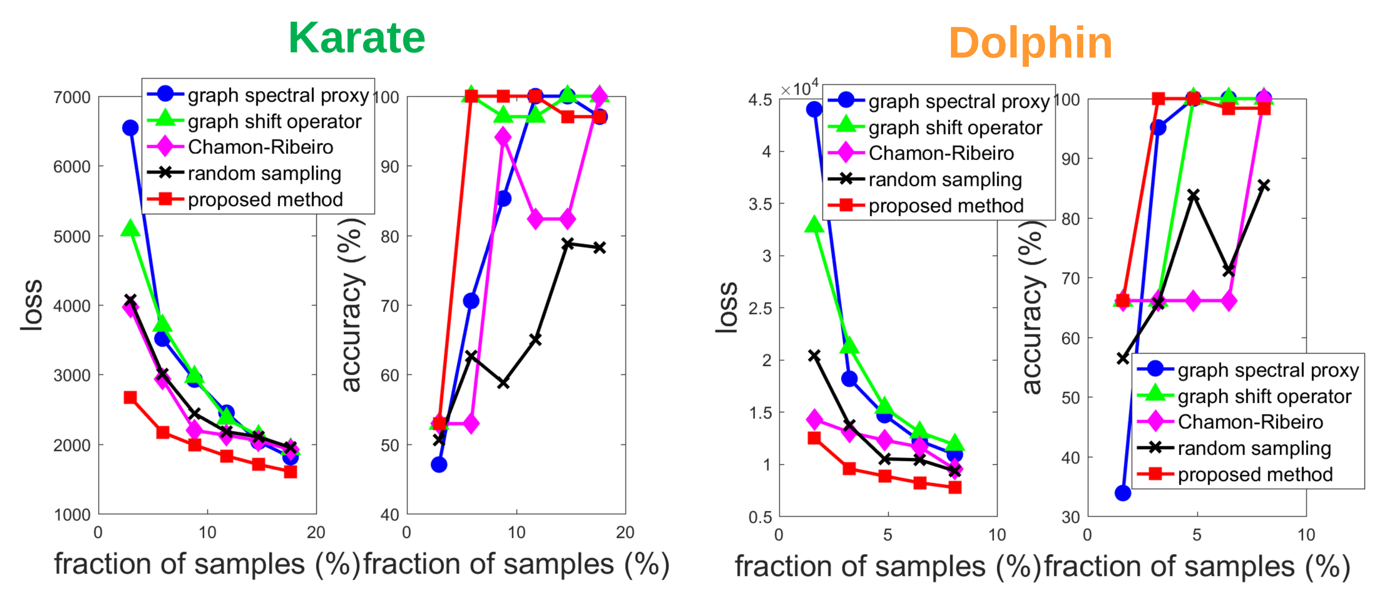
a) Only 2 community labels (+1 or -1) – control multi-label class representation issue
b) Network topology is given – control graph construction issue
c) Small network size – tracking similarity (overlapping sampled nodes) to GSP

- **Comparative Methods:**

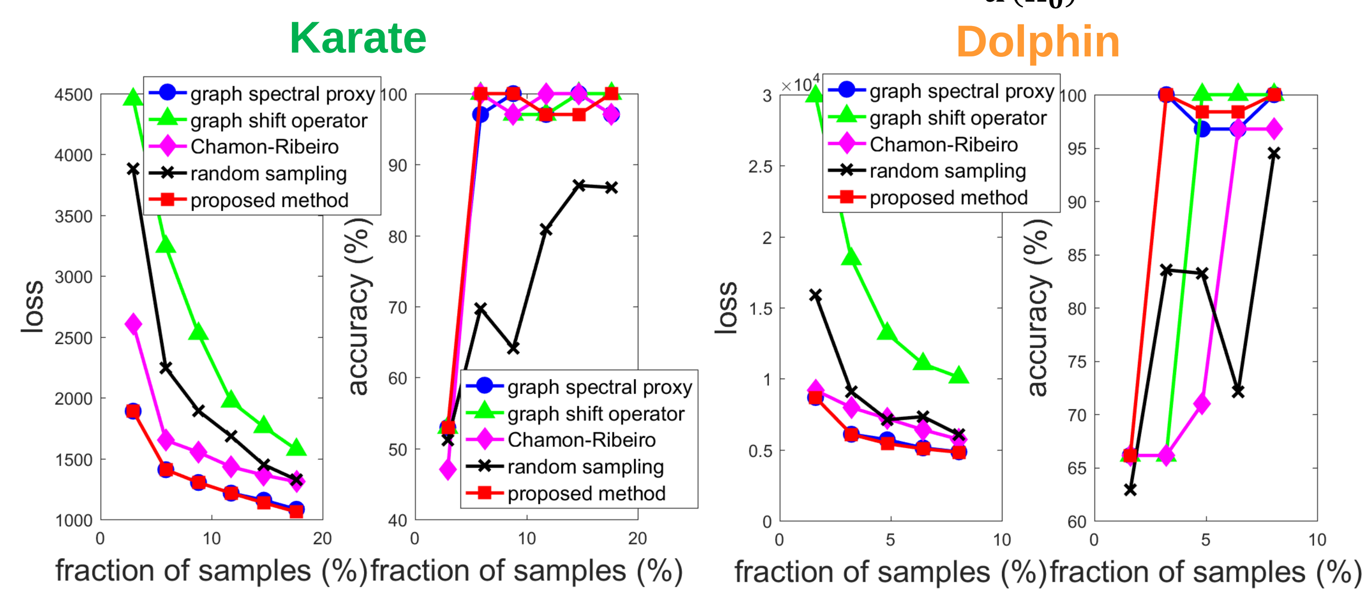
1) Random sampling (Rand)
2) Graph spectral proxy (GSP) [4]
3) Graph shift operator (GSO) [5]
4) Chamon-Ribeiro's method (CRM) [6]

Experimental Results

- $\boldsymbol{\Omega}_0 = \mathbf{L}$ (unnormalized graph Laplacian) and $\boldsymbol{\sigma}^2 = \frac{1}{\mathrm{tr}(\boldsymbol{\Omega}_0)}$



- $\boldsymbol{\Omega}_0 = \mathbf{L}_N$ (normalized graph Laplacian) and $\boldsymbol{\sigma}^2 = \frac{1}{\mathrm{tr}(\boldsymbol{\Omega}_0)}$



✓ **The proposed AG-SSL algorithm yields perfect community detection by only sampling 2 nodes in each dataset**

- **Fraction of overlapping samples to other GSP methods**

Karate, # samples $s = 6$, $\Omega_0 = \mathrm{L}$					Dolphin, # samples $s = 5$, $\Omega_0 = \mathrm{L}$				
	Proposed	GSP	GSO	CRM		Proposed	GSP	GSO	CRM
Proposed	100	66.67	16.67	16.67	Proposed	100	40	20	0
GSP	66.67	100	33.33	0	GSP	40	100	0	0
GSO	16.67	33.33	100	0	GSO	0	0	100	0
CRM	16.67	0	0	100	CRM	20	0	0	100

✓ **Relaxed (approximate) bandwidth in GSP → higher similarity**

References

[1] Wei Liu, Jun Wang, and Shih-Fu Chang, "Robust and scalable graph-based semisupervised learning," *Proceedings of the IEEE*, 2012
[2] Mikhail Belkin, Irina Matveeva, and Partha Niyogi, "Regularization and semi-supervised learning on large graphs," in *COLT*, 2004
[3] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, 2006
[4] Aamir Anis, Akshay Gadde, and Antonio Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Trans. Signal Process.*, 2016
[5] Siheng Chen, Rohan Varma, Aliaksei Sandryhaila, and Jelena Kovacevic, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, 2015
[6] Luiz FO Chamon and Alejandro Ribeiro, "Greedy sampling of graph signals," *IEEE Trans. Signal Process.*, 2018