A PROBABILISTIC INTERPRETATION OF SAMPLING THEORY OF GRAPH SIGNALS

Akshay Gadde and Antonio Ortega

Department of Electrical Engineering
University of Southern California, Los Angeles
Email: agadde@usc.edu, ortega@sipi.usc.edu

ABSTRACT

We give a probabilistic interpretation of sampling theory of graph signals. To do this, we first define a generative model for the data using a pairwise Gaussian random field (GRF) which depends on the graph. We show that, under certain conditions, reconstructing a graph signal from a subset of its samples by least squares is equivalent to performing MAP inference on an approximation of this GRF which has a low rank covariance matrix. We then show that a sampling set of given size with the largest associated cut-off frequency, which is optimal from a sampling theoretic point of view, minimizes the worst case predictive covariance of the MAP estimate on the GRF. This interpretation also gives an intuitive explanation for the superior performance of the sampling theoretic approach to active semi-supervised classification.

Index Terms— Graph Signal Processing, Sampling theorem, Gaussian Markov random field, Semi-supervised learning, Active learning

1. INTRODUCTION

Graph signal processing aims to extend the tools for analysis, approximation, denoising and interpolation of traditional signals to signals defined on graphs. The advantage of this framework is that it allows us to process the given data while taking into consideration the underlying connectivity between the data points. The graph can be inherent to the data as is the case in application areas such as social networks and sensor networks or it can be constructed using the data to capture the underlying geometry. Examples of the latter are found in image processing and machine learning (see [1, 2]).

In this paper, we focus on the sampling theory of graph signals. The classical Nyquist-Shannon sampling theorem says that a signal with bandwidth f is uniquely determined by its (uniformly spaced) samples if the sampling rate is higher than $2f$. Intuitively, it tells us how "smooth" the signal has to be, for perfect recovery, given the sampling density, and vice versa. Moreover, the signal can be perfectly reconstructed from the samples by a simple low pass filter. Sampling theory of graph signals similarly deals with the problem of reconstructing an unknown graph signal from its samples on a subset of nodes. Frequency domain representation of graph signals is given by the eigenvectors and eigenvalues of the Laplacian matrix associated with the graph. In order to pose a sampling theorem analogous to the Nyquist-Shannon sampling theorem, we need to find the maximum bandwidth (in the graph spectral domain) that a graph signal can have so that it is uniquely determined by its samples on the given subset of nodes. Conversely, given the bandwidth, we need to find the smallest subset of nodes, so that recovery of any signal with that bandwidth, from its samples on that subset, is unique and stable.

Given that the signal is smooth enough to be uniquely represented by its samples on a subset of nodes, we need to give an efficient and stable algorithm to reconstruct the unknown samples. These questions have been answered to some extent in [3, 4, 5, 6]. We discuss some of these results in Section 2.2.

This sampling theoretic perspective has been shown to be very useful for graph based active semi-supervised learning [7]. In this context, label prediction is considered as a graph signal reconstruction problem. The characterization of a subset of nodes given by the sampling theory, namely the associated cutoff frequency is used as a criterion function to choose the optimal set nodes to be labelled for active learning.

Sampling theoretic approaches for active and semi-supervised learning [7] are purely deterministic. However, their probabilistic interpretation is desired for the following reasons: 1. It allows us to understand them as model based methods and thus, makes it easier to include them as components of a larger probabilistic model. 2. It can also suggest a principled way to refine the model parameters (which are given by the underlying graph) as more data is observed (see [8] for an example). 3. The interpretation presented in this paper assumes a Gaussian random field model for the data. This may lead to generalizations of the sampling theory to data with non-Gaussian distributions which might be more realistic for a classification problem. 4. This interpretation also makes the relationship between the sampling theoretic approach and previously proposed semi-supervised [9] and active learning [10, 11] methods more apparent as discussed in Section 5.

The main contributions of this paper are the following. We define a generative model for graph signals using a pairwise Gaussian random field (GRF) with a covariance matrix that depends on the graph. We show that, when conditions of the graph signal sampling theorem are satisfied, bandlimited reconstruction of a graph signal from a subset of its samples is equivalent to performing MAP inference on a low rank approximation of the above GRF. This learning model performs very well in classification problems, as demonstrated in the experiments, since the true data covariance matrix is expected to be close to low rank. We then show that a sampling set of given size with the largest associated cut-off frequency, which is optimal from a sampling theoretic point of view, minimizes the worst case predictive covariance of the MAP estimate on the GRF.

2. SAMPLING THEORY OF GRAPH SIGNALS

2.1. Preliminaries and Notation

We consider a connected, undirected and weighted graph $G = (\mathcal{V}, \mathcal{E})$. The nodes $\mathcal{V}$ in the graph are indexed by $\{1, 2, \ldots, N\}$. $\mathcal{S}^c$ denotes the complement of $\mathcal{S}$ in $\mathcal{V}$, i.e., $\mathcal{S}^c = \mathcal{V} \setminus \mathcal{S}$. The edge set $\mathcal{E}$ is given by $\{(i, j, w_{ij})\}$, where $i, j \in \mathcal{V}$ and $w_{ij} \in \mathbb{R}^+$.

(i, j, w_{ij}) denotes an edge with weight w_{ij} connecting nodes i and j. The connectivity information given by $\mathcal{E}$ is encoded by the adjacency matrix $\mathbf{W}$ of size $N \times N$ with $\mathbf{W}(i,j) = w_{ij}$. The degree matrix $\mathbf{D}$ is a diagonal matrix $\text{diag}\{d_1, \ldots d_N\}$, where $d_i = \sum_j w_{ij}$ is the degree of node i. The Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The symmetric normalized form of the Laplacian is given by $\mathcal{L} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$. A graph signal $f : \mathcal{V} \to \mathbb{R}$ is a mapping which takes a real value on each node of the graph. It can be represented as $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^\top \in \mathbb{R}^N$. For $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{x}_\mathcal{S}$ denotes a sub-vector of $\mathbf{x}$ consisting of its components indexed by $\mathcal{S}$. Similarly, for $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{A}_{\mathcal{S}_1 \mathcal{S}_2}$ is the sub-matrix of $\mathbf{A}$ with rows indexed by $\mathcal{S}_1$ and columns indexed by $\mathcal{S}_2$. For simplicity, we denote $\mathbf{A}_{\mathcal{S}\mathcal{S}}$ by $\mathbf{A}_\mathcal{S}$. We use $\lambda_{\max}[.]$ and $\lambda_{\min}[.]$ to denote the largest and the smallest eigenvalue of a matrix, respectively. $\text{tr}(.)$ denotes the trace of a matrix. $\mathbf{A}^+$ is used to denote the pseudo-inverse of $\mathbf{A}$. $\mathbf{1}$ and $\mathbf{0}$ denote vectors or matrices of ones and zeros, respectively.

It can be shown that $\mathbf{L}$ and $\mathcal{L}$ are positive semi-definite. Hence, $\mathbf{L}$ has real eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_N$ and a corresponding orthogonal set of eigenvectors $\{\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^N\}$. It can be diagonalized as $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U} = (\mathbf{u}^1, \ldots, \mathbf{u}^N)$ and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_N\}$. Variation in the eigenvectors of $\mathbf{L}$ over the graph (as captured by $\mathbf{u}^\top \mathbf{L} \mathbf{u} = \sum_{i,j} w_{ij}(\mathbf{u}_i - \mathbf{u}_j)^2$) increases as the corresponding eigenvalues increase. Thus, these eigenvectors allow us a to define a graph dependent notion of frequency for the graph signals. The so-called Graph Fourier Transform (GFT)[1] is defined as $\tilde{\mathbf{f}}_i = \langle \mathbf{f}, \mathbf{u}^i \rangle$ (or in an equivalent matrix form $\tilde{\mathbf{f}} = \mathbf{U}^\top \mathbf{f}$), where $\tilde{\mathbf{f}}_i$ is the GFT coefficient corresponding to frequency λ_i. An ω-bandlimited signal has its GFT supported on $[0, \omega]$, i.e., $\tilde{\mathbf{f}}_i = 0$ for $\lambda_i > \omega$. Conversely, such a signal is said to have a bandwidth equal to ω. If $\{\lambda_1, \ldots, \lambda_r\}$ are the eigenvalues less than ω, then any ω-bandlimited signal can be written as a linear combination of corresponding eigenvectors

$$\mathbf{f} = \sum_{i=1}^r \mathbf{a}_i \mathbf{u}^i = \mathbf{U}_{\mathcal{V}\mathcal{R}} \mathbf{a}, \tag{1}$$

where $\mathbf{a}$ is the coefficient vector. The space of ω-bandlimited signals is called a Paley-Wiener space $PW_\omega(G)$.

2.2. Sampling Theorem and Bandlimited Reconstruction

Sampling theory deals with the problem reconstructing an ω-bandlimited signal $\mathbf{f}$ from its samples $\mathbf{f}_\mathcal{S}$ on the nodes in $\mathcal{S} \subseteq \mathcal{V}$. There are three important questions that need to be answered in this context: 1. Given $\mathcal{S}$, what is the maximum bandwidth ω that $\mathbf{f}$ can have so that it is uniquely determined by $\mathbf{f}_\mathcal{S}$? 2. Which is the best sampling set $\mathcal{S}_{\text{opt}}$ of a given size m? 3. Given that $\mathbf{f}$ is uniquely determined by $\mathbf{f}_\mathcal{S}$, how to find the unknown samples $\mathbf{f}_{\mathcal{S}^c}$? We briefly review some of the results related to each of the above problems.

Let $L_2(\mathcal{S}^c)$ be the space of signals which are identically zero on $\mathcal{S}$ but can have non-zero samples on $\mathcal{S}^c$, i.e., $\mathbf{g}_\mathcal{S} = \mathbf{0} \; \forall \mathbf{g} \in L_2(\mathcal{S}^c)$. It is easy to see that for all signals in $PW_\omega(G)$ to be uniquely determined by their samples on $\mathcal{S}$, we need $PW_\omega(G) \cap L_2(\mathcal{S}^c) = \{\mathbf{0}\}$. This observation leads to the following theorem.

Theorem 1 (Sampling Theorem [6]). *Any signal in $PW_\omega(G)$ can be uniquely reconstructed from its samples on a subset of nodes $\mathcal{S}$ if and only if*

$$\omega < \inf_{\mathbf{g} \in L_2(\mathcal{S}^c)} \omega(\mathbf{g}), \tag{2}$$

---

[1]The GFT is usually defined using the normalized Laplacian $\mathcal{L}$. We define it using $\mathbf{L}$ for the sake of notational simplicity. However, most of the discussion in the paper can be easily generalized to $\mathcal{L}$.

where $\omega(.)$ denotes the bandwidth of a signal. If the above condition is satisfied, then $\mathcal{S}$ is said to be a uniqueness set for $PW_\omega(G)$.

To ensure unique recovery of a signal from its samples on $\mathcal{S}$, its bandwidth has to be less than $\inf_{\mathbf{g} \in L_2(\mathcal{S}^c)} \omega(\mathbf{g})$. This is called the cut-off frequency associated with the subset $\mathcal{S}$ and is denoted by $\omega(\mathcal{S})$. An estimate of the cut-off frequency is given by [6]

$$\Omega_k(\mathcal{S}) = \left(\lambda_{\min} \left[(\mathbf{L}^k)_{\mathcal{S}^c} \right] \right)^{1/k}. \tag{3}$$

It can be shown that $\Omega_k(\mathcal{S}) \leq \omega(\mathcal{S})$ and we get closer to $\omega(\mathcal{S})$ as k increases.

A larger cut-off frequency estimate $\Omega_k(\mathcal{S})$ implies that a bigger space of signals can be perfectly recovered from their samples on $\mathcal{S}$. Therefore, $\Omega_k(\mathcal{S})$ can be used as a criterion function to be maximized for choosing the optimal sampling set $\mathcal{S}_{\text{opt}}$ of given size m, i.e.,

$$\mathcal{S}_{\text{opt}} = \underset{|\mathcal{S}|=m}{\arg\max} \, \Omega_k(\mathcal{S}). \tag{4}$$

The above problem is combinatorial and NP-hard. A greedy algorithm for finding an approximate solution is proposed in [6].

Consider a signal $\mathbf{f} \in PW_\omega(G)$ with $\omega < \omega(\mathcal{S})$. Using the representation of a bandlimited signal in (1), we get that $\mathbf{f}_\mathcal{S} = \mathbf{U}_{\mathcal{S}\mathcal{R}} \mathbf{a}$. Since $\mathbf{f}$ is uniquely sampled on $\mathcal{S}$, $\mathbf{U}_{\mathcal{S}\mathcal{R}}$ must have full column rank so that the least squares solution $\mathbf{a}$ of the above system of equations is unique. The unknown samples can then be reconstructed by:

$$\mathbf{f}_{\mathcal{S}^c} = \mathbf{U}_{\mathcal{S}^c \mathcal{R}} (\mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \mathbf{U}_{\mathcal{S}\mathcal{R}})^{-1} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \mathbf{f}_\mathcal{S}. \tag{5}$$

A faster, iterative method for bandlimited reconstruction is proposed in [5], which does not need the computation of eigenvectors.

These sampling theory based algorithms for subset selection and signal reconstruction have been applied to graph based active semi-supervised learning and are shown to perform better than many state of the art approaches [7].

3. GRF MODEL FOR GRAPH SIGNALS

In order to give a probabilistic interpretation of the graph signal processing framework, we define a generative model for the signal using a pairwise Gaussian Random Field (GRF) based on the graph G. A random signal $\mathbf{f} = (\mathbf{f}_1, \ldots, \mathbf{f}_N)^\top$ is assumed to be drawn from the following distribution:

$$p(\mathbf{f}) \propto \exp\left(-\sum_{i,j} w_{ij}(\mathbf{f}_i - \mathbf{f}_j)^2 - \delta \sum_i \mathbf{f}_i^2 \right)$$

$$= \exp\left(-\mathbf{f}^\top (\mathbf{L} + \delta \mathbf{I}) \mathbf{f} \right), \tag{6}$$

where $\mathbf{I}$ denotes an identity matrix of size $N \times N$. Let $\mathbf{K}$ be the covariance matrix of the the GRF. Then, from the above equation, the inverse covariance matrix (also known as the precision matrix) can be written as:

$$\mathbf{K}^{-1} = \mathbf{L} + \delta \mathbf{I}. \tag{7}$$

Note that $\mathbf{K}$ has the same eigenvectors as $\mathbf{L}$, while the corresponding eigenvalues are $\sigma_i = \frac{1}{\lambda_i + \delta}$. Thus, $\mathbf{K}$ can be diagonalized as

$$\mathbf{K} = \sum_{i=1}^N \frac{1}{\lambda_i + \delta} \mathbf{u}^i \mathbf{u}^{i\top} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top, \tag{8}$$

where $\mathbf{\Sigma} = \text{diag}\{\sigma_1, \ldots, \sigma_N\}$. The advantage of introducing the parameter δ is that it leads to a non-singular precision matrix and thus, allows us to have a proper covariance matrix. $\sigma_1 = 1/\delta$ can be thought of as the variance of the DC component of $\mathbf{f}$ since $\mathbf{u}^1 = \mathbf{1}$.

4. SAMPLING THEORY AND INFERENCE OVER GRF

Consider a signal $\mathbf{f}$ generated using the GRF defined in (6) with covariance matrix $\mathbf{K} = (\mathbf{L} + \delta\mathbf{I})^{-1}$. As in the sampling problem, we observe the samples of $\mathbf{f}$ on a subset $\mathcal{S}$ of nodes. Our goal is to estimate the unknown samples. It is well known that the conditional distribution of $\mathbf{f}_{\mathcal{S}^c}$ given $\mathbf{f}_{\mathcal{S}}$ equals $\mathcal{N}(\boldsymbol{\mu}_{\mathcal{S}^c|\mathcal{S}}, \mathbf{K}_{\mathcal{S}^c|\mathcal{S}})$, where

$$\boldsymbol{\mu}_{\mathcal{S}^c|\mathcal{S}} = \mathbf{K}_{\mathcal{S}^c\mathcal{S}}(\mathbf{K}_{\mathcal{S}})^+ \mathbf{f}_{\mathcal{S}} \text{ and} \tag{9}$$

$$\mathbf{K}_{\mathcal{S}^c|\mathcal{S}} = \mathbf{K}_{\mathcal{S}^c} - \mathbf{K}_{\mathcal{S}^c\mathcal{S}}(\mathbf{K}_{\mathcal{S}})^+ \mathbf{K}_{\mathcal{S}\mathcal{S}^c} \tag{10}$$

are the MAP estimate and the predictive covariance matrix of $\mathbf{f}_{\mathcal{S}^c}$ given $\mathbf{f}_{\mathcal{S}}$, respectively [9, 12].

4.1. Bandlimited Reconstruction as MAP Inference

Let λ_r be the largest eigenvalue of $\mathbf{L}$ which is less than ω. We define $\hat{\mathbf{K}}$ to be a low rank approximation of $\mathbf{K}$ which only contains the spectral components corresponding to $\{\lambda_1, \ldots, \lambda_r\}$, i.e.,

$$\hat{\mathbf{K}} = \sum_{i=1}^{r} \frac{1}{\lambda_i + \delta} \mathbf{u}^i \mathbf{u}^{i\top} = \mathbf{U}_{\mathcal{VR}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{VR}}^\top. \tag{11}$$

Consider the problem of reconstructing a random signal generated using a GRF with covariance $\hat{\mathbf{K}}$, from its samples on $\mathcal{S}$. The following theorem shows that, if conditions of the sampling theorem are satisfied, then the error of bandlimited reconstruction is zero.

Theorem 2. *Let $\mathbf{f}$ be a random graph signal generated using the GRF with covariance $\hat{\mathbf{K}}$ given by (11). Let $\hat{\mathbf{f}}_{\mathcal{S}^c}$ be the bandlimited reconstruction of $\mathbf{f}_{\mathcal{S}^c}$ obtained from its samples on $\mathcal{S}$, where $\mathcal{S}$ is a uniqueness set for $PW_\omega(G)$. Then, $\|\mathbf{f}_{\mathcal{S}^c} - \hat{\mathbf{f}}_{\mathcal{S}^c}\| = 0$.*

Before proving the above theorem, we show, in the following lemma, that bandlimited reconstruction is equivalent to MAP inference on the GRF with covariance $\hat{\mathbf{K}}$.

Lemma 1. *Let $\mathcal{S} \subseteq \mathcal{V}$ be a uniqueness set for $PW_\omega(G)$. Then the MAP estimate of $\mathbf{f}_{\mathcal{S}^c}$ given $\mathbf{f}_{\mathcal{S}}$ in a GRF with covariance matrix $\hat{\mathbf{K}}$ is equal to the bandlimited reconstruction given by (5).*

Proof. Under a permutation which groups together nodes in $\mathcal{S}^c$ and $\mathcal{S}$, we can write $\hat{\mathbf{K}}$ as the following block matrix

$$\begin{bmatrix} \hat{\mathbf{K}}_{\mathcal{S}^c} & \hat{\mathbf{K}}_{\mathcal{S}^c\mathcal{S}} \\ \hat{\mathbf{K}}_{\mathcal{S}\mathcal{S}^c} & \hat{\mathbf{K}}_{\mathcal{S}} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_{\mathcal{S}^c\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}^c\mathcal{R}}^\top & \mathbf{U}_{\mathcal{S}^c\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \\ \mathbf{U}_{\mathcal{S}\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}^c\mathcal{R}}^\top & \mathbf{U}_{\mathcal{S}\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \end{bmatrix} \tag{12}$$

Therefore, we can write the MAP estimate obtained with covariance $\hat{\mathbf{K}}$ as,

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}^c|\mathcal{S}} = \mathbf{U}_{\mathcal{S}^c\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top (\mathbf{U}_{\mathcal{S}\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top)^+ \mathbf{f}_{\mathcal{S}}. \tag{13}$$

Because $\omega < \omega(\mathcal{S})$, we have that $\mathbf{U}_{\mathcal{S}\mathcal{R}}$ has full column rank and equivalently, $\mathbf{U}_{\mathcal{S}\mathcal{R}}^\top$ has full row rank. Therefore, we can write $(\mathbf{U}_{\mathcal{S}\mathcal{R}} \boldsymbol{\Sigma}_{\mathcal{R}} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top)^+ = (\mathbf{U}_{\mathcal{S}\mathcal{R}}^\top)^+ \boldsymbol{\Sigma}_{\mathcal{R}}^+ \mathbf{U}_{\mathcal{S}\mathcal{R}}^+$ and $\mathbf{U}_{\mathcal{S}\mathcal{R}}^+ = (\mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \mathbf{U}_{\mathcal{S}\mathcal{R}})^{-1} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top$. Simplifying (13) using these equalities leads to

$$\hat{\mathbf{f}}_{\mathcal{S}^c} = \mathbf{U}_{\mathcal{S}^c\mathcal{R}} (\mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \mathbf{U}_{\mathcal{S}\mathcal{R}})^{-1} \mathbf{U}_{\mathcal{S}\mathcal{R}}^\top \mathbf{f}_{\mathcal{S}},$$

which is equal to the least squares solution given in (5). $\square$

Proof of Theorem 2. From Lemma 1, $\hat{\mathbf{f}}_{\mathcal{S}^c} = \hat{\boldsymbol{\mu}}_{\mathcal{S}^c|\mathcal{S}}$. Therefore, $\mathbb{E}(\|\mathbf{f}_{\mathcal{S}^c} - \hat{\mathbf{f}}_{\mathcal{S}^c}\|^2) = \text{tr}(\mathbb{E}(\mathbf{f}_{\mathcal{S}^c} - \hat{\boldsymbol{\mu}}_{\mathcal{S}^c|\mathcal{S}})(\mathbf{f}_{\mathcal{S}^c} - \hat{\boldsymbol{\mu}}_{\mathcal{S}^c|\mathcal{S}})^\top) = \text{tr}(\hat{\mathbf{K}}_{\mathcal{S}^c|\mathcal{S}})$. Now, $\hat{\mathbf{K}}_{\mathcal{S}^c|\mathcal{S}} = \hat{\mathbf{K}}_{\mathcal{S}^c} - \hat{\mathbf{K}}_{\mathcal{S}^c\mathcal{S}}(\hat{\mathbf{K}}_{\mathcal{S}})^+ \hat{\mathbf{K}}_{\mathcal{S}\mathcal{S}^c}$. Using the block form of $\hat{\mathbf{K}}$ in (12), and the fact that $\mathbf{U}_{\mathcal{S}\mathcal{R}}$ has full column rank, it is easy to show that $\hat{\mathbf{K}}_{\mathcal{S}^c|\mathcal{S}} = \mathbf{0}$, which implies $\mathbb{E}(\|\mathbf{f}_{\mathcal{S}^c} - \hat{\mathbf{f}}_{\mathcal{S}^c}\|^2) = 0$. But since, $\|\mathbf{f}_{\mathcal{S}^c} - \hat{\mathbf{f}}_{\mathcal{S}^c}\| \geq 0$, we get $\|\mathbf{f}_{\mathcal{S}^c} - \hat{\mathbf{f}}_{\mathcal{S}^c}\| = 0$. $\square$

4.2. Cut-off Frequency and Estimation Error

If the true covariance matrix is only approximately low rank, then MAP inference with $\hat{\mathbf{K}}$ gives a non-zero reconstruction error. The best sampling set in this case is the one which minimizes the predictive covariance. According to the sampling theory of graph signals, the optimal sampling set of given size is the one which has the largest associated cut-off frequency. We show that finding a sampling set $\mathcal{S}$ which maximizes a crude estimate of the cut-off frequency $\Omega_1(\mathcal{S})$ is equivalent to minimizing the maximum eigenvalue of the predictive covariance of $\mathbf{f}_{\mathcal{S}^c}$ given $\mathbf{f}_{\mathcal{S}}$.

Proposition 1. *Let $\mathcal{S}_{opt} = \arg\max_{|\mathcal{S}|=m} \Omega_1(\mathcal{S})$. Let $\mathbf{K} = (\mathbf{L} + \delta\mathbf{I})^{-1}$. Then, $\mathcal{S}_{opt} = \arg\min_{|\mathcal{S}|=m} \lambda_{max}[\mathbf{K}_{\mathcal{S}^c|\mathcal{S}}]$.*

Proof. Consider a block matrix representation of $\mathbf{K}$ similar to (12). Using the block matrix inversion formula, we can write $\mathbf{K}^{-1}$ as

$$\mathbf{K}^{-1} = \begin{bmatrix} \mathbf{S}_{\mathbf{K}_{\mathcal{S}}}^{-1} & -(\mathbf{K}_{\mathcal{S}^c})^{-1} \mathbf{K}_{\mathcal{S}^c\mathcal{S}} \mathbf{S}_{\mathbf{K}_{\mathcal{S}^c}}^{-1} \\ -(\mathbf{K}_{\mathcal{S}})^{-1} \mathbf{K}_{\mathcal{S}^c\mathcal{S}}^\top \mathbf{S}_{\mathbf{K}_{\mathcal{S}}}^{-1} & \mathbf{S}_{\mathbf{K}_{\mathcal{S}^c}}^{-1} \end{bmatrix},$$

$$\text{where} \quad \mathbf{S}_{\mathbf{K}_{\mathcal{S}}} = \mathbf{K}_{\mathcal{S}^c} - \mathbf{K}_{\mathcal{S}^c\mathcal{S}}(\mathbf{K}_{\mathcal{S}})^{-1} \mathbf{K}_{\mathcal{S}^c\mathcal{S}}^\top,$$

$$\mathbf{S}_{\mathbf{K}_{\mathcal{S}^c}} = \mathbf{K}_{\mathcal{S}} - \mathbf{K}_{\mathcal{S}^c\mathcal{S}}^\top (\mathbf{K}_{\mathcal{S}^c})^{-1} \mathbf{K}_{\mathcal{S}^c\mathcal{S}} \tag{14}$$

are the Schur complements of $\mathbf{K}_{\mathcal{S}}$ and $\mathbf{K}_{\mathcal{S}^c}$ respectively. $\mathbf{L}_{\mathcal{S}^c} = (\mathbf{K}^{-1})_{\mathcal{S}^c} - \delta\mathbf{I}_{\mathcal{S}^c} = \mathbf{S}_{\mathbf{K}_{\mathcal{S}}}^{-1} - \delta\mathbf{I}_{\mathcal{S}^c}$. Note that $\mathbf{S}_{\mathbf{K}_{\mathcal{S}}} = \mathbf{K}_{\mathcal{S}^c|\mathcal{S}}$. Thus, the estimated cut-off frequency corresponding to the subset $\mathcal{S}$ of nodes can be written in terms of the conditional covariance matrix

$$\Omega_1(\mathcal{S}) = \lambda_{\min}[\mathbf{L}_{\mathcal{S}^c}] = \frac{1}{\lambda_{\max}[\mathbf{K}_{\mathcal{S}^c|\mathcal{S}}]} - \delta. \tag{15}$$

The result readily follows from this. $\square$

A sampling set with the largest estimated cut-off frequency $\Omega_1(\mathcal{S})$ also minimizes the worst case prediction error of the MAP estimate on a GRF with $\mathbf{K} = (\mathbf{L} + \delta\mathbf{I})^{-1}$. However, as shown in Lemma 1, bandlimited signal reconstruction is equivalent to MAP estimation with a low rank approximation of $\mathbf{K}$. Intuitively, a better estimate of the predictive covariance, in this model of signal reconstruction, can be obtained with by $((\mathbf{K}^k)_{\mathcal{S}^c|\mathcal{S}})^{1/k}$ with larger values of k as it gives more weight to the principal components with larger variance. This justifies the use of $\Omega_k(\mathcal{S})$ with $k > 1$ as a criterion for active learning.

4.3. Justification for the Sampling Theoretic Approach to Active Semi-supervised Classification

MAP estimation is optimal for reconstructing signals generated using a GRF with a full rank covariance matrix, because it minimizes the mean squared error of estimation. Moreover, since the estimation error equals $\text{tr}(\mathbf{K}_{\mathcal{S}^c|\mathcal{S}})$, an optimal sampling set of size m is given by $\arg\min_{|\mathcal{S}|=m} \text{tr}(\mathbf{K}_{\mathcal{S}^c|\mathcal{S}})$. Indeed, this is the so-called *V-optimality criterion* for active learning proposed in [10].

However, in a classification problem, data points in the same class are highly correlated whereas data points in different classes

have very small correlation. Since the number of classes is typically very small compared to the number of data points, we expect the (unknown) "true" covariance matrix to be very well-approximated by a low rank matrix [13]. Thus, bandlimited interpolation is a better model for signal reconstruction in this context, since it is equivalent to MAP estimation with a low rank covariance matrix. Maximizing the cut-off frequency is a natural set selection criterion for this learning model.

5. RELATED WORK

Different criteria have been proposed for batch mode active learning on Gaussian random fields. The approach presented in [14] selects the points to label such the mutual information between the labelled and unlabelled data points is maximized. Our sampling theoretic approach (4) is more similar to the methods proposed in [10, 11]. These methods use MAP estimation on GRF [9] as their model for label prediction. As stated before, [10] chooses the sampling set $\mathcal{S}$ by minimizing $\text{tr}(\mathbf{K}_{\mathcal{S}^c|\mathcal{S}})$. The method in [11], on the other hand, tries to minimize $\sum_{ij}(\mathbf{K}_{\mathcal{S}^c|\mathcal{S}})_{ij}$ (also known as Σ-*optimality criterion*). This is equivalent to minimizing the risk of the surveying problem [15] (which is the problem of determining the proportion of nodes belonging to one class). All the above methods are closely related to the optimal design of experiments [16]. Experiment design deals with the problem of estimating a vector from a set of linear measurements. The goal is to choose the optimal set of m measurements so that the estimation error is minimized. Different error measures lead to different optimality criteria. For example, minimizing the trace of estimation covariance leads to A-optimal design whereas minimizing its determinant gives the D-optimal design. The sampling theoretic approach is closer to the so-called E-optimal design which minimizes the worst case prediction error given by the maximum eigenvalue of the predictive covariance matrix.

6. EXPERIMENTS

To demonstrate the effectiveness of the framework of sampling theory, we first apply it to the problem of graph based active semi-supervised classification. In our experiment, we use a subset of the USPS handwritten digit dataset containing 100 16×16 images each of digits 0 to 9. We construct a weighted K-NN graph of 1000 nodes with $K = 10$ and the similarities given by $w_{ij} = \exp\left(-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|^2}{\sigma^2}\right)$. The problem is to choose the nodes to be labelled and then predict the unknown labels from the queried labels. We consider different combinations of active learning criteria and learning models. As expected from the discussion in Section 4.3, selecting the sampling set by maximizing the cutoff frequency and then performing bandlimited reconstruction outperforms Σ and V-optimality criteria used in conjunction with MAP estimation (see Figure 1(a)). Even if the learning model is fixed to bandlimited interpolation, the sampling theoretic approach gives better results as seen in Figure 1(b)). This is because maximizing the cutoff frequency is a more suitable set selection criterion under this model.

On the other hand, if we consider the problem of regression of a random real valued graph signal generated using a covariance matrix that is not low rank, a V-optimal set is expected to give a better SNR of reconstruction. This is demonstrated in Figure 2 where we reconstruct a random real valued signal generated with the covariance matrix obtained using the graph from the previous example.

7. CONCLUSION AND FUTURE WORK

In this paper, we gave a probabilistic interpretation for the sampling theory of graph signals. We showed that if the data is generated using
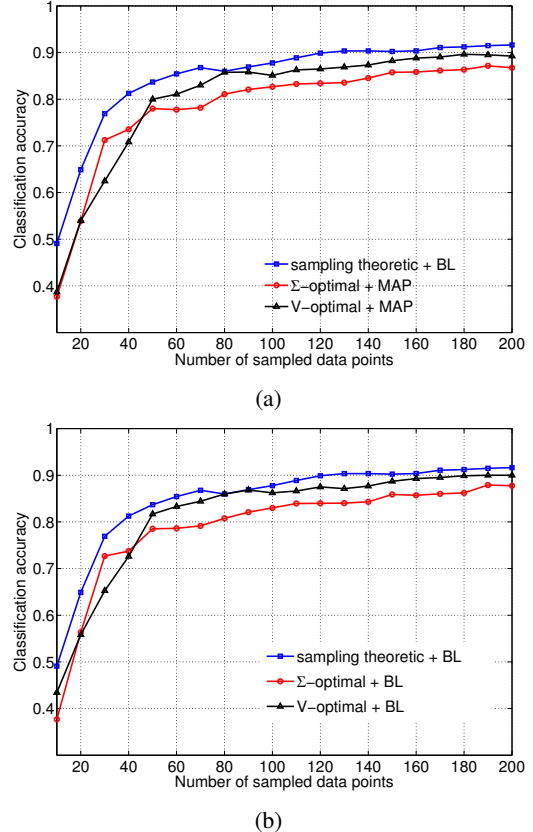
Fig. 1: Figure shows the performance of different active learning criteria in conjunction with two learning models, namely, (a) MAP [9] and (b) bandlimited reconstruction (BL)
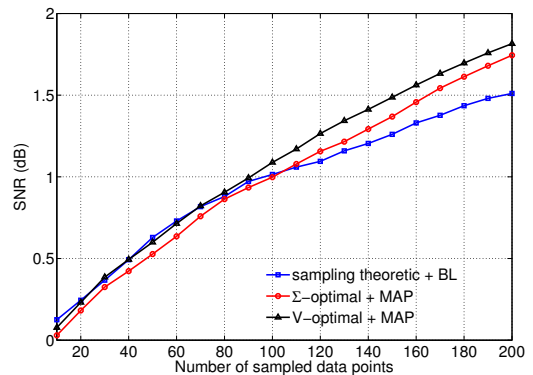
Fig. 2: Performance in the case of reconstruction of a random real valued signal (averaged over 100 trials)

a Gaussian random field whose precision matrix equals the graph Laplacian, then bandlimited reconstruction is equivalent to the MAP inference on an approximation of this GRF which has a low rank covariance matrix. Moreover, an optimal sampling set obtained via sampling theory minimizes the worst case predictive covariance of MAP estimation on the GRF.

A probabilistic interpretation allows us to view graph signal sampling theory as a model based method. It would be interesting to consider it as part of a larger probabilistic model which refines the covariance matrix as more data is observed. This interpretation also suggests a generalization of the sampling theory to non-Gaussian models which might be more realistic for some applications.

8. REFERENCES

[1] D.I. Shuman, S.K. Narang, P. Frossard, A Ortega, and P. Van-dergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 83–98, May 2013.

[2] A. Sandryhaila and J.M.F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *Signal Processing Magazine, IEEE*, vol. 31, no. 5, pp. 80–90, Sept 2014.

[3] I. Pesenson, "Sampling in Paley-Wiener spaces on combinatorial graphs," *Transactions of the American Mathematical Society*, vol. 360, no. 10, pp. 5603–5627, 2008.

[4] S. K. Narang, A. Gadde, and A. Ortega, "Signal processing techniques for interpolation in graph structured data," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2013, pp. 5445–5449.

[5] S. K. Narang, A. Gadde, E. Sanou, and A. Ortega, "Localized iterative methods for interpolation in graph structured data," in *Signal and Information Processing (GlobalSIP), IEEE Global Conference on*, 2013.

[6] A. Anis, A. Gadde, and A. Ortega, "Towards a sampling theorem for signals on arbitrary graphs," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2014.

[7] A. Gadde, A. Anis, and A. Ortega, "Active semi-supervised learning using sampling theory for graph signals," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 492–501.

[8] A. Kapoor, H. Ahn, Y. Qi, and R. Picard, "Hyperparameter and kernel learning for graph based semi-supervised classification," in *Advances in Neural Information Processing Systems*, 2005, pp. 627–634.

[9] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *International Conference on Machine Learning (ICML)*, 2003, vol. 3, pp. 912–919.

[10] M. Ji and J. Han, "A variance minimization criterion to active learning on graphs," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012, vol. 22, pp. 556–564.

[11] Y. Ma, R. Garnett, and J. Schneider, "Σ-optimality for active learning on Gaussian random fields," in *Advances in Neural Information Processing Systems*, 2013, pp. 2751–2759.

[12] R. Scholtz, *Supplemental Notes on Random Processes*, 2012.

[13] D. Kuang, H. Park, and C. Ding, "Symmetric nonnegative matrix factorization for graph clustering," in *SIAM International Conference on Data Mining*, 2012, vol. 12, pp. 106–117.

[14] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *The Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.

[15] R. Garnett, Y. Krishnamurthy, X. Xiong, R. Mann, and J. Schneider, "Bayesian optimal active search and surveying," in *International Conference on Machine Learning (ICML)*, 2012, pp. 1239–1246.

[16] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2009.