UNIVERSAL BOUNDS FOR THE SAMPLING OF GRAPH SIGNALS

Luiz F. O. Chamon and Alejandro Ribeiro

Electrical and Systems Engineering
University of Pennsylvania
e-mail: luizf@seas.upenn.edu, aribeiro@seas.upenn.edu

ABSTRACT

Sampling is a fundamental topic in graph signal processing with applications in estimation, clustering, and video compression. In contrast to traditional signal processing, however, the irregularity of the signal domain makes the selection of the sampling points non-trivial and hard to analyze. Indeed, although graph signal reconstruction is well-understood in the noiseless case, performance bounds for the interpolation of noisy samples exist mainly for randomized sampling schemes. This paper addresses this issue by deriving a lower bound on the mean-square interpolation error for graph signals. This bound is universal in the sense that it is not restricted to a specific sampling method and holds for all sampling sets. Simulations illustrate the tightness of the bound, which is then used to evaluate the performance of greedy sampling. Finally, a solution to the complexity issues of kernel principal component analysis is proposed using graph signal sampling.

Index Terms— Graph signal processing, sampling, interpolation, greedy algorithms, kernel principal component analysis

1. INTRODUCTION

Graph signal processing (GSP) is an emerging field that studies signals supported on irregular domains [1, 2]. It extends traditional signal processing techniques to more intricate data structures, finding applications in sensor networks, image processing, clustering, and neuroscience, to name a few [3–6]. Extensions of sampling, in particular, have attracted considerable interest from the GSP community [7–14]. This is not surprising given the fundamental role of sampling in signal processing [15].

Sampling methods in GSP are broadly divided into two categories: *selection sampling*, in which the graph signal is observed at a subset of nodes [13], and *aggregation sampling*, in which the signal is observed at a single node for many applications of the graph shift [8]. This work focuses on the former. As in classical signal processing, samples are only useful inasmuch as they represent the original signal. Conditions under which it is possible to recover a graph signal from noiseless samples can be found in [10–13]. For noisy observations, however, it not possible to exactly recover the original signal and it must therefore be approximated. Characterizing the approximation error is key, especially since selecting an optimal sampling set is in general NP-hard [16–19].

In this work, the graph signal is modeled as random and stationary process with respect to the given graph [20–22]. Using this model, we derive bounds on the interpolation mean-square error (MSE) that are universal in the sense that they hold for all sampling sets and any sampling scheme. These universal bounds are computationally tractable and provide a practical means of benchmarking the MSE performance of sampling techniques. Numerical analyses for small networks show that the bounds are tight. For large scale networks, we compare these bounds to the MSE of greedy selection sampling, showing that they are close to each other. This illustrates both the tightness of the bounds and how greedy selection sampling is close to optimal.

To illustrate the practical value of the bounds we recall that the concept of sampling is also at the core of statistical methods, such as data subsetting and variable selection, that are crucial for *big data* applications [23, 24]. Kernel methods, in particular, are prone to complexity issues for large data sets. For instance, performing kernel principal component analysis (kPCA) on a data set of size n needs n^2 kernel evaluations (KEs) and extracting projections for new data requires n KEs and $\Theta(nk)$ operations, where k is the number of principal components (PCs) [25, 26]. Solutions based on promoting sparsity or low-rank have been put forward [26–28]. We show here that this problem can be cast in the context of graph signal sampling. We then use greedy sampling to select a subset of kernels to be evaluated for kPCA and show that this complexity reduction comes at a small performance loss.

Notation: Lowercase boldface letters represent vectors ($\boldsymbol{x}$), uppercase boldface letters are matrices ($\boldsymbol{X}$), and calligraphic letters denote sets ($\mathcal{A}$). We write $|\mathcal{A}|$ for the cardinality of $\mathcal{A}$. Set subscripts refer either to the vector obtained by keeping only the elements with indices in the set ($\boldsymbol{x}_{\mathcal{A}}$) or to the submatrix whose columns have indices in the set ($\boldsymbol{X}_{\mathcal{A}}$). To say $\boldsymbol{X}$ is a positive semi-definite (PSD) matrix we write $\boldsymbol{X} \succeq 0$, so that for $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times n}$, $\boldsymbol{X} \preceq \boldsymbol{Y} \Leftrightarrow \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{b} \leq \boldsymbol{b}^T\boldsymbol{Y}\boldsymbol{b}$, for all $\boldsymbol{b} \in \mathbb{R}^n$. Finally, we take the derivative of a function f with respect to an $n \times 1$ vector $\boldsymbol{x}$ to yield the $1 \times n$ gradient vector, i.e., $\frac{\partial f}{\partial \boldsymbol{x}} = [\, \partial f/\partial x_1 \;\cdots\; \partial f/\partial x_n \,]$ [29].

2. GRAPH SIGNALS

A *graph-supported signal* (*graph signal* for short) is an assignment of values to the nodes of a graph. Formally, let $\mathbb{G}$ be a weighted graph with node set $\mathcal{V}$, $|\mathcal{V}| = n$, and define a graph signal to be an injective mapping $\sigma : \mathcal{V} \to \mathbb{R}$. This signal can be represented by an $n \times 1$ vector that captures its value at each node of the graph:

$$\boldsymbol{x} = [\;\sigma(v_1) \quad \cdots \quad \sigma(v_n)\;]^T, \quad v_i \in \mathcal{V}. \tag{1}$$

As in traditional signal processing, GSP is interested in spectral representations of (1), which depend on the graph that supports $\boldsymbol{x}$. Indeed, let $\boldsymbol{A}$ be a matrix representation of $\mathbb{G}$. For instance, $\boldsymbol{A}$ can be its adjacency matrix or some choice of discrete Laplacian [1, 2]. Assume that $\boldsymbol{A}$ is consistent with the vector signal (1) in the sense that they employ the same ordering of the nodes in $\mathcal{V}$. Furthermore, assume that $\boldsymbol{A}$ is normal, i.e., there exist $\boldsymbol{V}$ orthonormal and $\boldsymbol{\Sigma}$ diagonal such that $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T$ [30]. Note that if $\boldsymbol{A}$ is not normal, spectral energy conservation properties analog to Parseval's theorem

in classical signal processing no longer hold. Then, the *graph Fourier transform* of $\boldsymbol{x}$ is given by [1, 2]

$$\bar{\boldsymbol{x}} = \boldsymbol{V}^T \boldsymbol{x}. \tag{2}$$

A case of particular interest is when the graph signal is related to the underlying graph in the sense that it lies in a subspace induced by $\mathbb{G}$ (more specifically, by $\boldsymbol{A}$). In this case, $\bar{\boldsymbol{x}}$ is $\mathcal{K}$-*sparse*, i.e., all elements of $\bar{\boldsymbol{x}}$ vanish except those with index in $\mathcal{K}$, and

$$\boldsymbol{x} = \boldsymbol{V}_{\mathcal{K}} \bar{\boldsymbol{x}}_{\mathcal{K}}. \tag{3}$$

These graph signals are called $\mathcal{K}$-*bandlimited*. Note that since we do not rely on graph frequency orderings as in [7, 12–14], we take bandlimitedness to be synonymous to spectral-sparsity instead of the typical "low-pass" definition.

The reason why bandlimited graphs signals are useful is similar to traditional signal processing: these signals can be sampled and interpolated without loss of information. Indeed, take sampling to be the operation of observing the value of a graph signal on $\mathcal{S} \subseteq \mathcal{V}$, the *sampling set*. Then, there exists a set $\mathcal{S}$ of size $|\mathcal{K}|$ such that exact interpolation is feasible in the noiseless case [10–13]. In the presence of noise, however, $\boldsymbol{x}$ can only be approximated. To do so, the next section poses the noisy interpolation problem as a stochastic estimation problem, from which the minimum MSE interpolation operator can be derived. This allows us to provide bounds on the reconstruction error that can be used to inform the choice of the sampling set.

3. INTERPOLATION OF SAMPLED GRAPH SIGNALS

To formulate graph signal interpolation from corrupted samples as a stochastic estimation problem, let $\boldsymbol{x}$ be a $\mathcal{K}$-bandlimited random vector, i.e., take $\bar{\boldsymbol{x}}_{\mathcal{K}}$ in (3) to be a zero-mean random vector with $\boldsymbol{\Lambda} = \mathbb{E}\, \bar{\boldsymbol{x}}_{\mathcal{K}} \bar{\boldsymbol{x}}_{\mathcal{K}}^T = \operatorname{diag}\{\lambda_i\}$. Without loss of generality, assume $\boldsymbol{\Lambda}$ is full-rank. Otherwise, remove from $\mathcal{K}$ any element i for which $\lambda_i = 0$. This class of random processes is referred to as *wide-sense stationary* with respect to $\mathbb{G}$ in [20–22]. Also, assume that we observe a noisy version of $\boldsymbol{x}$:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}, \tag{4}$$

where $\boldsymbol{w}$ is an $n \times 1$ zero-mean noise vector with diagonal covariance matrix $\boldsymbol{\Lambda}_w = \mathbb{E}\, \boldsymbol{w}\boldsymbol{w}^T = \operatorname{diag}\{\lambda_{w,i}\}$, $\lambda_{w,i} > 0$. Note that (4) is related to the class of *approximately bandlimited* graph signals from [7] and that the noiseless case is recovered for $\boldsymbol{w} = \boldsymbol{0}$.

The signal in (4) is now sampled by observing only the elements of $\boldsymbol{y}$ whose index are in the sampling set $\mathcal{S}$. To clarify the derivations, define the selection matrix $\boldsymbol{C} \in \{0, 1\}^{|\mathcal{S}| \times n}$ composed of the identity matrix rows with indices in $\mathcal{S}$, so that the samples of (4) can be written as

$$\boldsymbol{y}_{\mathcal{S}} = \boldsymbol{C}\boldsymbol{y}. \tag{5}$$

Using the samples in (5), $\boldsymbol{x}$ can be estimated as

$$\hat{\boldsymbol{x}} = \boldsymbol{L}\boldsymbol{y}_{\mathcal{S}} = \boldsymbol{L}\boldsymbol{C}\boldsymbol{y}, \tag{6}$$

for some $\boldsymbol{L} \in \mathbb{R}^{n \times |\mathcal{S}|}$, leading to an interpolation error whose covariance matrix is defined as

$$\boldsymbol{K}(\hat{\boldsymbol{x}}) = \mathbb{E}(\boldsymbol{x} - \hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}})^T. \tag{7}$$

Since $\boldsymbol{L}$ recovers (approximates) the original graph signal $\boldsymbol{x}$ from its samples $\boldsymbol{y}_{\mathcal{S}}$ it is called a *linear interpolation operator* [7, 12, 13].

Given the above, the optimal interpolation problem becomes that of finding $\hat{\boldsymbol{x}}^\star = \boldsymbol{L}^\star \boldsymbol{y}_{\mathcal{S}}$ such as to minimize the covariance matrix in (7) in the sense that $\boldsymbol{K}(\hat{\boldsymbol{x}}^\star) \preceq \boldsymbol{K}(\hat{\boldsymbol{x}})$ for all $\hat{\boldsymbol{x}}$ as in (6). Note that this problem is more general than the typical least-squares estimation since $\text{MSE}(\hat{\boldsymbol{x}}) = \mathbb{E}\, \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 = \operatorname{Tr}[\boldsymbol{K}(\hat{\boldsymbol{x}})]$ and $\boldsymbol{K}(\hat{\boldsymbol{x}}^\star) \preceq \boldsymbol{K}(\hat{\boldsymbol{x}}) \Rightarrow \text{MSE}(\hat{\boldsymbol{x}}^\star) \leq \text{MSE}(\hat{\boldsymbol{x}})$ [29].

From the partial ordering of the PSD cone, $\boldsymbol{L}^\star$ can be obtained by minimizing the scalar cost function $J(\boldsymbol{L}) = \boldsymbol{b}^T \boldsymbol{K}(\boldsymbol{L}\boldsymbol{y}_{\mathcal{S}})\boldsymbol{b}$ simultaneously for all $\boldsymbol{b}$ [29], where we replaced $\hat{\boldsymbol{x}}$ by its expression (6). Then, since $\boldsymbol{x}$ is bandlimited [see (3)] and $\bar{\boldsymbol{x}}_{\mathcal{K}}$ is uncorrelated,

$$\begin{aligned} J(\boldsymbol{L}) &= \boldsymbol{b}^T\, \mathbb{E}(\boldsymbol{x} - \hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}})^T \boldsymbol{b} \\ &= \boldsymbol{b}^T\, \mathbb{E}(\boldsymbol{V}_{\mathcal{K}} \bar{\boldsymbol{x}}_{\mathcal{K}} - \boldsymbol{L}\boldsymbol{y}_{\mathcal{S}})(\boldsymbol{V}_{\mathcal{K}} \bar{\boldsymbol{x}}_{\mathcal{K}} - \boldsymbol{L}\boldsymbol{y}_{\mathcal{S}})^T \boldsymbol{b} \\ &= \boldsymbol{b}^T \Big[\boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T - \boldsymbol{L}\boldsymbol{C}\boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T - \boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T \boldsymbol{C}^T \boldsymbol{L}^T \\ &\qquad + \boldsymbol{L}\boldsymbol{C}(\boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T + \boldsymbol{\Lambda}_w) \boldsymbol{C}^T \boldsymbol{L}^T \Big] \boldsymbol{b}. \end{aligned} \tag{8}$$

Setting the derivative of (8) with respect to $\boldsymbol{b}^T \boldsymbol{L}$ to zero yields

$$\frac{\partial J(\boldsymbol{L})}{\partial \boldsymbol{b}^T \boldsymbol{L}} = \boldsymbol{0} \Leftrightarrow \boldsymbol{C} \left(\boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T + \boldsymbol{\Lambda}_w \right) \boldsymbol{C}^T \boldsymbol{L}^T \boldsymbol{b} = \boldsymbol{C}\boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T \boldsymbol{b},$$

which must hold for all $\boldsymbol{b}$. Therefore, $\boldsymbol{L}^\star$ is any solution of

$$\boldsymbol{L}^\star \boldsymbol{C} \left(\boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T + \boldsymbol{\Lambda}_w \right) \boldsymbol{C}^T = \boldsymbol{V}_{\mathcal{K}} \boldsymbol{\Lambda} \boldsymbol{V}_{\mathcal{K}}^T \boldsymbol{C}^T. \tag{9}$$

Given a sampling set $\mathcal{S}$, (6) and (9) can now be used to optimally estimate a graph signal from its samples. Note that (9) also holds in the noiseless case ($\boldsymbol{\Lambda}_w = \boldsymbol{0}$), although its solution may not be unique. This happens if the sampling set is not sufficient to determine $\boldsymbol{x}$, i.e., if $\boldsymbol{C}\boldsymbol{V}_{\mathcal{K}}$ is rank-deficient [10–13]. In contrast, since $\boldsymbol{\Lambda}_w \succ \boldsymbol{0}$, the matrix on the left-hand side of (9) is always invertible and $\boldsymbol{L}^\star$ is unique for all $\mathcal{S}$. This is similar to the well-known regularization effect of noise in Kalman filtering [29]. The interpolation performance, however, is not the same for all sampling sets and is examined in the sequel.

4. BOUNDS ON INTERPOLATION PERFORMANCE

For any sampling set $\mathcal{S}$, using $\boldsymbol{L}^\star$ from (9) leads to the smallest error covariance matrix out of all possible linear interpolators. This does not guarantee, however, that there is no $\mathcal{S}'$, $|\mathcal{S}'| = |\mathcal{S}|$, for which the reconstruction error is smaller. Indeed, due to the irregularity of the domain of graph signals, selecting the "best" sampling set is a combinatorial problem that is NP-hard in general [16–19]. Therefore, finding bounds on the interpolation performance that hold for all $\mathcal{S}$ can inform the sampling set selection by (i) describing how different factors influence the reconstruction performance and (ii) gauging the quality of given sampling sets.

To bound the performance of the optimal interpolation operator (9), we start by determining its error covariance matrix $\boldsymbol{K}(\hat{\boldsymbol{x}}^\star)$. Substituting any $\boldsymbol{L}^\star$ satisfying (9) into (7) gives

$$\boldsymbol{K}(\hat{\boldsymbol{x}}^\star) = \boldsymbol{V}_{\mathcal{K}} \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}_{\mathcal{K}}^T \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{\Lambda}_w^{-1} \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{V}_{\mathcal{K}} \right)^{-1} \boldsymbol{V}_{\mathcal{K}}^T, \tag{10}$$

where we used the matrix inversion lemma [30] and the fact that $\boldsymbol{\Lambda}_w$ is diagonal so that $(\boldsymbol{C}\boldsymbol{\Lambda}_w \boldsymbol{C}^T)^{-1} = \boldsymbol{C}\boldsymbol{\Lambda}_w^{-1}\boldsymbol{C}^T$. Using (10), we get

$$\text{MSE}(\hat{\boldsymbol{x}}^\star) = \operatorname{Tr}[\boldsymbol{K}(\hat{\boldsymbol{x}}^\star)] = \operatorname{Tr}\left(\bar{\boldsymbol{K}}\right), \tag{11}$$

where $\bar{\boldsymbol{K}} = \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}_{\mathcal{K}}^T \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{\Lambda}_w^{-1} \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{V}_{\mathcal{K}} \right)^{-1}$. Finally, to simplify the derivations, we assume that both signal and noise are homoscedastic, i.e., $\boldsymbol{\Lambda} = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{\Lambda}_w = \sigma_w^2 \boldsymbol{I}$. Under these conditions, the following holds.

Proposition 1. *Let $\boldsymbol{x} = \boldsymbol{V}_{\mathcal{K}}\bar{\boldsymbol{x}}_{\mathcal{K}}$ be a bandlimited stationary graph signal, $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}$ be its noisy observations, and $\hat{\boldsymbol{x}}^\star = \boldsymbol{L}^\star \boldsymbol{y}_{\mathcal{S}}$ be its minimum MSE estimate based on a sampling set $\mathcal{S}$. Assuming $\mathbb{E}\,\bar{\boldsymbol{x}}_{\mathcal{K}}\bar{\boldsymbol{x}}_{\mathcal{K}}^T = \sigma^2\boldsymbol{I}$ and $\mathbb{E}\,\boldsymbol{w}\boldsymbol{w}^T = \sigma_w^2\boldsymbol{I}$, the reconstruction error* $\text{MSE}(\hat{\boldsymbol{x}}^\star) = \mathbb{E}\,\|\boldsymbol{x} - \hat{\boldsymbol{x}}^\star\|^2$ *is bounded by*

$$\frac{|\mathcal{K}|^2}{|\mathcal{K}|\sigma^{-2} + \sigma_w^{-2}\bar{\ell}_{|\mathcal{S}|}} \leq \text{MSE}(\hat{\boldsymbol{x}}^\star) \leq |\mathcal{K}|\sigma^2, \tag{12}$$

where $\bar{\ell}_m$ is the sum of the m largest leverage scores, i.e., $\bar{\ell}_m = \max_{\mathcal{X}:|\mathcal{X}|=m} \sum_{j\in\mathcal{X}} \|\boldsymbol{v}_j\|^2$ with $\boldsymbol{v}_j^T$ the j-th row of $\boldsymbol{V}_{\mathcal{K}}$.

Proof. Start with the upper bound that is achieved for an empty sampling set, i.e., for $\mathcal{S} = \{\} \Rightarrow \boldsymbol{C} = \boldsymbol{0}$. Since $\boldsymbol{V}_{\mathcal{K}}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Lambda}_w^{-1}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{V}_{\mathcal{K}} \succeq \boldsymbol{0}$ and that matrix inversion is operator antitone [31], one gets $\bar{\boldsymbol{K}}(\hat{\boldsymbol{x}}^\star) \preceq \boldsymbol{\Lambda}$, from which the upper bound in (12) follows.

The lower bound is obtained by using the fact that the trace is the sum of the eigenvalues and the arithmetic/harmonic means inequality. Indeed, for any $n \times n$ positive definite matrix $\boldsymbol{X}$, it holds that [30]:

$$\text{Tr}\,(\boldsymbol{X}) \geq \frac{n^2}{\text{Tr}\,(\boldsymbol{X}^{-1})},$$

with equality if and only if $\boldsymbol{X} = \gamma\boldsymbol{I}$, $\gamma > 0$. Applying this bound to (11) yields

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{x}}^\star) &\geq \frac{|\mathcal{K}|^2}{\text{Tr}\left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}_{\mathcal{K}}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Lambda}_w^{-1}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{V}_{\mathcal{K}}\right)} \\ &\geq \frac{|\mathcal{K}|^2}{|\mathcal{K}|\sigma^{-2} + \sigma_w^{-2}\bar{\ell}_{|\mathcal{S}|}}.\end{aligned}$$

■

The bound in (12) was derived by taking the graph signal to be stochastic, so that the expectation is taken over realizations of the signal and the bound holds for all sampling sets $\mathcal{S} \subseteq \mathcal{V}$. It is worth noting that (12) depends only on statistics of the graph signal, structural properties of the underlying graph, and the sampling set size. As expected, (12) decreases with the sampling set size. The rate of decay, however, depends on the value of the leverage scores $\|\boldsymbol{v}_i\|^2$. In a sequential sampling scheme, their value can therefore be used to inform whether it is worth acquiring a new sample. A rate bound can be obtained using the fact that $\bar{\ell}_{|\mathcal{K}|} \leq |\mathcal{K}|\ell_{\text{max}}$ for $\ell_{\text{max}} = \max_j \|\boldsymbol{v}_j\|^2$. Then, if the sample set is chosen so as to uniquely determine the graph signal [10–13], i.e., $|\mathcal{S}| \geq |\mathcal{K}|$, (12) reduces to

$$\text{MSE}(\hat{\boldsymbol{x}}^\star) \geq \frac{|\mathcal{K}|}{\sigma^{-2} + \sigma_w^{-2}\ell_{\text{max}}}. \tag{13}$$

It is clear from (13) that the reconstruction error increases linearly with the bandwidth of the graph signal, which is a fundamental limitation for large dimensional signals. It therefore shows the importance of working with low bandwidth signals and, consequently, of appropriately identifying the signal's underlying graph.

Although these observations give insights into graph signal interpolation, one of the main motivation behind Proposition 1 was addressing the issue of sampling set selection. Towards this end, we propose the following corollary:

Corollary 1. *For any graph signal and its interpolation as in Proposition 1, any sampling set $\mathcal{S}$ for which* $\text{MSE}(\hat{\boldsymbol{x}}^\star) = \eta$ *satisfies*

$$\bar{\ell}_{|\mathcal{S}|} \geq \frac{|\mathcal{K}|^2 - \eta|\mathcal{K}|\sigma^{-2}}{\eta\sigma_w^{-2}}. \tag{14}$$

Given that $\bar{\ell}_{|\mathcal{S}|} \leq |\mathcal{S}|\ell_{max}$, it also holds that

$$|\mathcal{S}| \geq \frac{|\mathcal{K}|^2 - \eta|\mathcal{K}|\sigma^{-2}}{\eta\ell_{max}\sigma_w^{-2}}. \tag{15}$$

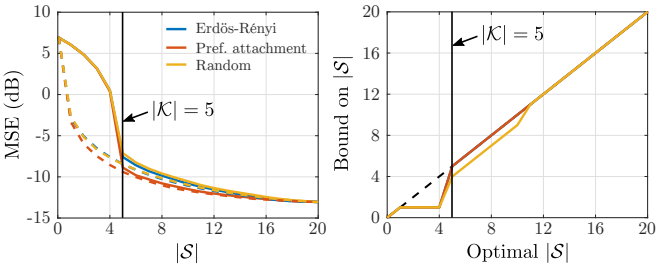


Fig. 1. Comparison between (12) and (14) (dashed lines) and optimal values (solid lines) for three random graphs ($n = 20$)

Corollary 1 allows us to lower bound the number of samples needed to achieve a desired MSE. From (15), note that the number of samples is inversely proportional to the MSE. Moreover, although (15) suggest that the sample set size required to achieve a certain MSE grows with $\mathcal{O}(|\mathcal{K}|^2)$, it is not necessarily the case. Indeed, recall that ℓ_{max} is a function of $|\mathcal{K}|$ and in fact monotonically increases to 1 as $|\mathcal{K}| \to n$. Still, as in the noiseless case, the signal bandwidth is a dominating factor in the determination of the minimum sampling set size.

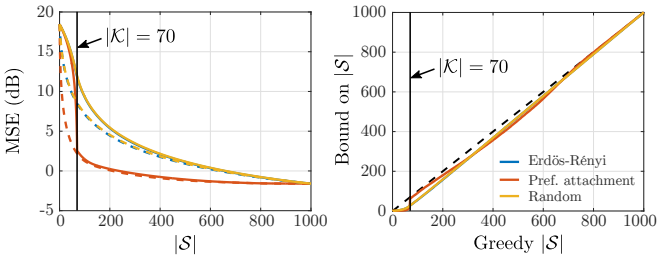
Although (15) characterizes the overall behavior of the sampling set size, it is not informative in practice because it largely underestimates $|\mathcal{S}|$. On the other hand, (14) yields a tighter bound which can be used, together with (12), to evaluate a sampling set or sampling technique. Indeed, Fig. 1 compares (12) and (14) to the minimum interpolation MSE and optimal set size, found by exhaustive search, for three graph models: Erdös-Rényi [32], preferential attachment [33], and a random undirected graph with weights uniformly distributed in $[0, 1]$. Note that the bounds are too conservative for $|\mathcal{S}| < |\mathcal{K}|$, but become tighter as $|\mathcal{S}|$ increases. This is because the inequality used to derive (12) becomes tighter as the eigenvalues of $\boldsymbol{K}(\hat{\boldsymbol{x}})$ become more similar.

Remark 1. Bounds on the MSE performance of graph sampling have also been derived in [7, 13]. These papers consider randomized sampling schemes, including uniform and leverage score sampling, and derive bounds on the optimal sampling distributions and interpolation error. The bounds in Proposition 1 and Corollary 1 differ from the bounds in [7, 13] in that in the latter the spectrum of the graph signal is deterministic and the sampling is random. Thus, these bounds hold in expectation over different sampling realizations for a specific randomized strategy. The bounds in (12) hold in expectation over realizations of the signal and apply to any sampling strategy.

4.1. Greedy MSE sampling

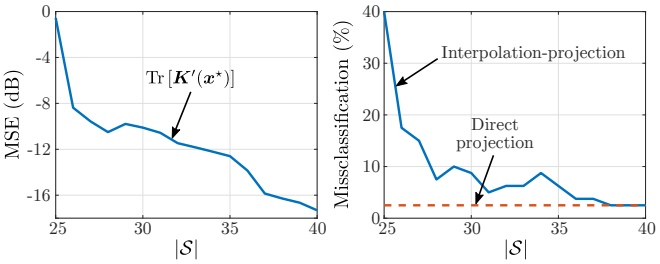
Greedy sampling remains ubiquitous in GSP and has proven successful in many applications [10–13, 34]. This is not surprising given the attractive features of greedy algorithms for large-scale problems. First, their complexity is polynomial. Also, since they build the solution sequentially, they can be interrupted at any time if, for instance, a desired performance level is reached. Finally, near-optimality results exist for the minimization of supermodular functions. This is indeed why greedy algorithms are often used in sensor selection, experimental design, and machine learning [16–19].

No performance analysis, however, is available to justify their success in GSP. In fact, the MSE in (11) is not a supermodular set function, which would provide a near-optimal guarantee for greedy search [35]. This can be seen from [18, Thm. 2.4] and the fact that $f(t) = t^{-2}$ is not operator antitone. Therefore, although greedily



(a) (b)

Fig. 2. Comparison between greedy MSE sampling (solid lines) and (12) and (14) (dashed lines) for random graphs ($n = 1000$)



(a) (b)

Fig. 3. Performance of kPCA for different sampling set sizes

minimizing the MSE appears to work in practice, there has yet to be a theoretical justification for it.

The bounds in Section 4 cannot be used to show optimality results for greedy sampling in general. Nevertheless, they can certify specific instances by comparing the obtained MSE or sampling set size to (12) and (14). Fig. 2 displays these results for the same three random graph models as Fig. 1. Fig. 2b, in particular, shows that the greedy sampling set size is within 10% of the lower bound. Even though these are not evidence for the optimality of greedy sampling, they allow us to gauge the quality of the sampling sets obtained using this scheme.

5. KERNEL PCA AND GSP

Kernel PCA is a nonlinear version of PCA [25] that identifies a data subspace by truncating the eigenvalue decomposition (EVD) of a Gram matrix constructed from a training dataset. Indeed, whereas PCA uses the empirical covariance matrix, kPCA evaluates inner products in a higher dimensional space $\mathbb{F}$ known as the *feature space*. Since the map φ between $\mathbb{R}^m$ and $\mathbb{F}$ can be nonlinear and $\mathbb{F}$ can have large or even infinite dimensionality, kPCA results in richer subspaces than linear PCA [25, 26, 36].

Naturally, constructing the Gram matrix by taking inner products in $\mathbb{F}$ can be challenging due to its dimensionality. This problem is addressed using the so called *kernel trick* [25,26,36]. A kernel is a function $\kappa : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ that evaluates an inner product in $\mathbb{F}$ directly from vectors in $\mathbb{R}^m$, i.e., $\kappa(\boldsymbol{r}, \boldsymbol{s}) = \langle \varphi(\boldsymbol{r}), \varphi(\boldsymbol{s}) \rangle_{\mathbb{F}}$. Then, given the training set $\{\boldsymbol{u}_i\}_{i=1,\dots,n}$, $\boldsymbol{u}_i \in \mathbb{R}^m$, the $n \times n$ kernel (Gram) matrix and its EVD is calculated as

$$\boldsymbol{\Phi} = \left[\kappa(\boldsymbol{u}_i, \boldsymbol{u}_j)\right]_{i,j=1,\dots,n} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T. \tag{16}$$

Using the representer's theorem, a new data vector $\boldsymbol{z}$ can be projected onto the first k eigenvectors of $\boldsymbol{\Phi}$ using $\mathcal{K} = 1, \dots, k$ in

$$\tilde{\boldsymbol{z}} = \boldsymbol{V}_{\mathcal{K}}^T \tilde{\boldsymbol{z}}, \quad \tilde{\boldsymbol{z}} = \left[\kappa(\boldsymbol{u}_i, \boldsymbol{z})\right]_{i=1,\dots,n}. \tag{17}$$

The projection in (17) takes $\Theta(kn)$ operations and n KEs, which makes this method impractical for large data sets even if the subspace of interest is small. In [27], this issue was addressed by using a Gaussian generative model for $\boldsymbol{\Phi}$ and showing that its maximum likelihood estimate depends only on a subset of the $\boldsymbol{u}_i$. Another approach is to impose sparsity on $\boldsymbol{V}$ *a priori* so that it depends only on a reduced number of training points [26]. Alternatively, one can find a representative subset of the training data and apply kPCA to that subset [28]. The issue with this last method is that finding a good data subset is known to be a hard problem [23, 24]. In fact, it is equivalent to the problem of sampling set selection in GSP.

Indeed, since we used the same notation as Section 2, it is straightforward to see that (17) has the form of the graph Fourier transform in (2) when $\boldsymbol{A} = \boldsymbol{\Phi}$. Certainly, kPCA can be seen as forcing the graph signal $\tilde{\boldsymbol{z}}$ to be bandlimited on the graph represented by $\boldsymbol{\Phi}$. Using the observations from Section 4.1, we then can find a data subset by sampling the training data greedily. This method is illustrated in a face recognition application based on the AT&T face data set [37]. First, we apply kPCA with a polynomial kernel of degree $d = 2$ [36] and retain $k = 25$ PCs. Then, we fit a one-against-one multiclass support vector machine (SVM) to the retained PCs (see [38] for details on this scheme).

It is worth noting that in this application we are no longer interested in $\tilde{\boldsymbol{z}}$ directly, i.e., the graph signal $\boldsymbol{x}$ from (4), but in the output of the SVM classifier. Indeed, our performance measure is now the classification error as opposed to the interpolation MSE. Explicitly, we wish to minimize the reconstruction MSE of the classification vector

$$\boldsymbol{c} = \boldsymbol{H}_{\text{SVM}} \bar{\boldsymbol{z}} = \boldsymbol{H}_{\text{SVM}} \boldsymbol{V}_{\mathcal{K}}^T \tilde{\boldsymbol{z}}, \tag{18}$$

where $\boldsymbol{H}_{\text{SVM}}$ is the $n(n-1)/2 \times k$ matrix that collects the SVM classifiers and $\bar{\boldsymbol{z}}$ is the projection of $\tilde{\boldsymbol{z}}$ onto the PCs [see (17)]. To do so, we replace $\hat{\boldsymbol{x}}$ in (7) by $\boldsymbol{H}_{\text{SVM}} \boldsymbol{V}_{\mathcal{K}}^T \hat{\boldsymbol{x}}$, so that (10) becomes

$$\boldsymbol{K}' = \boldsymbol{H}_{\text{SVM}} \left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}_{\mathcal{K}}^T \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{\Lambda}_w^{-1} \boldsymbol{C}^T \boldsymbol{C} \boldsymbol{V}_{\mathcal{K}} \right)^{-1} \boldsymbol{H}_{\text{SVM}}^T.$$

We can now sample to minimize $\mathrm{Tr}\left[\boldsymbol{K}'(\hat{\boldsymbol{x}}^\star)\right]$, the error in reconstructing $\boldsymbol{c}$.

Note that $\boldsymbol{K}'$ stems from a more general problem than the one used to derive Proposition 1. Finding a bound for this case is left for future works. Moreover, given that the images used to estimate $\boldsymbol{K}$ (train the kPCA) and those used as "graph signals" in (17) come from the same data set, there is no actual observation noise. Still, σ_w^2 can be used to regularize the inversions in (9) and (10), e.g., by choosing $\sigma_w^2 = 10^{-3}$ [29]. Fig. 3 shows the results of this procedure. Note from Fig. 3b that for $|\mathcal{S}| = 40$ we obtain the same result as implementing the kPCA projection without sampling (*direct projection*), but using only $\Theta(k|\mathcal{S}|)$ operations and $|\mathcal{S}|$ KEs, yielding an 8-fold reduction in complexity.

6. CONCLUSION

This work addressed the issue of sampling set selection in GSP by bounding the interpolation error of graph signals. First, the optimal linear interpolator was derived for noisy samples. Then, this result was used to obtain a bound on the MSE that holds for any sampling strategy. Simulations illustrated the tightness of this bound, which was used to assess the quality of greedy MSE sampling. Finally, data subsetting for kernel PCA was formulated as a graph signal sampling problem giving a considerable reduction in complexity at a negligible performance cost.

7. REFERENCES

[1] D.I. Shuman, S.K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30[3], pp. 83–98, 2013.

[2] A. Sandryhaila and J.M.F. Moura, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61[7], pp. 1644–1656, 2013.

[3] S.K. Narang and A. Ortega, "Perfect reconstruction two-channel wavelet filter banks for graph structured data," *IEEE Trans. Signal Process.*, vol. 60[6], pp. 2786–2799, 2012.

[4] N. Tremblay, G. Puy, R. Gribonval, and P. Vandergheynst, "Compressive spectral clustering," in *Int. Conf. on Mach. Learning*, 2016, pp. 1002—-1011.

[5] W. Huang, L. Goldsberry, N.F. Wymbs, S.T. Grafton, D.S. Bassett, and A. Ribeiro, "Graph frequency analysis of brain signals," 2016, arXiv:1512.00037v2.

[6] X. Zhu and M. Rabbat, "Approximating signals supported on graphs," in *Int. Conf. on Acoust., Speech and Signal Process.*, 2012, pp. 3921–3924.

[7] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Signal recovery on graphs: Fundamental limits of sampling strategies," *IEEE Trans. on Signal and Inf. Process. over Netw.*, vol. 2[4], pp. 539–554, 2016.

[8] A.G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Sampling of graph signals with successive local aggregations," *IEEE Trans. Signal Process.*, vol. 64[7], pp. 1832–1843, 2016.

[9] R. Varma, S. Chen, and J. Kovačević, "Spectrum-blind signal recovery on graphs," in *Int. Workshop on Comput. Advances in Multi-Sensor Adaptive Process.*, 2015, pp. 81–84.

[10] H. Shomorony and A.S. Avestimehr, "Sampling large data on graphs," in *Global Conf. on Signal and Inform. Process.*, 2014, pp. 933–936.

[11] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *IEEE Trans. Signal Process.*, vol. 64[18], pp. 4845–4860, 2016.

[12] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Trans. Signal Process.*, vol. 64[14], pp. 3775–3789, 2016.

[13] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Process.*, vol. 63[24], pp. 6510–6523, 2015.

[14] S. Chen, A. Sandryhaila, J.M.F. Moura, and J. Kovačević, "Signal recovery on graphs: Variation minimization," *IEEE Trans. Signal Process.*, vol. 63[17], pp. 4609–4624, 2015.

[15] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88[4], pp. 569–587, 2000.

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

[17] A. Das and D. Kempe, "Algorithms for subset selection in linear regression," in *ACM Symp. on Theory of Comput.*, 2008, pp. 45–54.

[18] G. Sagnol, "Approximation of a maximum-submodular-coverage problem involving spectral functions, with application to experimental designs," *Discrete Appl. Math.*, vol. 161[1-2], pp. 258–276, 2013.

[19] J. Ranieri, A. Chebira, and M. Vetterli, "Near-optimal sensor placement for linear inverse problems," *IEEE Trans. Signal Process.*, vol. 62[5], pp. 1135–1146, 2014.

[20] B. Girault, "Stationary graph signals using an isometric graph translation," in *European Signal Process. Conf.*, 2015, pp. 1516–1520.

[21] A.G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," 2016, arXiv:1603.04667v1.

[22] N. Perraudin and P. Vandergheynst, "Stationary signal processing on graphs," 2016, arXiv:1601.02522v3.

[23] D.P. Woodruff, "Sketching as a tool for numerical linear algebra," *Foundations and Trends in Theoretical Computer Science*, vol. 10[1-2], pp. 1–157, 2014.

[24] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for K-means, PCA and projective clustering," in *ACM-SIAM Symp. on Discrete Algorithms*, 2013, pp. 1434–1453.

[25] B. Schölkopf, , A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10[5], pp. 1299–1319, 1998.

[26] J. Arenas-Garcia, K.B. Petersen, G. Camps-Valls, and L.K. Hansen, "Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods," *IEEE Signal Process. Mag.*, vol. 30[4], pp. 16–29, 2013.

[27] M.E. Tipping, "Sparse kernel principal component analysis," in *NIPS*, 2000.

[28] Y. Washizawa, "Subset kernel principal component analysis," in *Int. Workshop on Mach. Learning for Signal Process.*, 2009.

[29] T. Kailath, A.H. Sayed, and B. Hassibi, *Linear estimation*, Prentice-Hall, 2000.

[30] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press, 2013.

[31] R. Bhatia, *Matrix analysis*, Springer, 1997.

[32] B. Bollobás, *Modern graph theory*, Springer, 1998.

[33] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286[5439], pp. 509–512, 1999.

[34] D. Thanou, D.I. Shuman, and P. Frossard, "Learning parametric dictionaries for signals on graphs," *IEEE Trans. Signal Process.*, vol. 62[15], pp. 3849–3862, 2014.

[35] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical Programming*, vol. 14[1], pp. 265–294, 1978.

[36] C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2007.

[37] AT&T Laboratories Cambridge, "The ORL database of faces," 1994, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[38] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13[2], pp. 415–425, 2002.