NEAR-OPTIMALITY OF GREEDY SET SELECTION IN THE SAMPLING OF GRAPH SIGNALS

Luiz F. O. Chamon and Alejandro Ribeiro

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

ABSTRACT

Sampling has been extensively studied in graph signal processing, having found applications in estimation, clustering, and video compression. Still, sampling set selection remains an open issue. Indeed, although conditions for graph signal reconstruction from noiseless samples were derived, the presence of noise makes sampling set selection combinatorial and NP-hard in general. Performance bounds are available only for randomized sampling schemes, even though greedy search remains ubiquitous in practice. This work sets out to justify the success of greedy sampling by introducing the concept of approximate supermodularity and updating the classical greedy bound to account for this class of functions. Then, it quantifies the approximate supermodularity of two important reconstruction figures of merit, namely the $\log\det$ of the error covariance matrix and the mean-square error, showing that they can be optimized with worst-case guarantees using greedy sampling.

***Index Terms*—** Graph signal processing, sampling, submodularity, approximate submodularity, greedy algorithms

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 complex data structures and has found applications in problems from 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 [7–12]. This is not surprising given the fundamental role of sampling in signal processing and its place at the core of statistical methods, such as data subsetting and variable selection, that are crucial for *big data* applications [13, 14].

Sampling methods in GSP are broadly divided into two categories: *selection sampling*, in which the graph signal is observed on a subset of nodes [12], and *aggregation sampling*, in which the signal is observed on 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 and conditions under which it is possible to reconstruct (interpolate) a graph signal from noiseless samples were derived in [9–12]. However, because they do not necessarily lead to a unique sampling set, the presence of noise raises the issue of which sampling set to choose. In [7, 12], this issue was addressed using randomized sampling schemes, including uniform and leverage score sampling, for which optimal sampling distributions and performance bounds were derived. Nevertheless, greedy sampling remains ubiquitous and has proven successful in many applications, though no performance analysis is available [9–12, 15].

To be sure, this is not surprising given the attractive features of greedy algorithms for large-scale problems. First, their complexity is polynomial in the deterministic case and stochastic versions exist that are linear in the size of the ground set. Also, since they build the solution sequentially, they can be interrupted at any time if, for instance, a desired performance level is reached. More importantly, there is an upper bound on the suboptimality of the greedy solution to monotonic supermodular function minimization problems [16]. This is indeed why greedy algorithms are often used in sensor selection, experimental design, and machine learning [17–21].

In this work, we study the reconstruction (interpolation) performance of greedy sampling schemes in GSP. We show that the $\log\det$ of the reconstruction error covariance matrix is *supermodular and monotone decreasing with respect to the sampling set* and can therefore be minimized with near-optimal guarantees using greedy search. The MSE, on the other hand, is known to not be supermodular [16, 19, 19, 20, 20, 21]. We show, however, that it is *approximately supermodular* and update the greedy near-optimal bound for such functions. These results justify the use of greedy sampling set selection in GSP and explain their success.

Due to space constraints, all proofs are deferred to the extended version of this paper available at [22].

Notation: Lowercase boldface letters represent vectors ($\boldsymbol{x}$), uppercase boldface letters are matrices ($\boldsymbol{X}$), and calligraphic letters denote sets ($\mathcal{A}$). We write $2^{\mathcal{A}}$ for the power set of $\mathcal{A}$ and $|\mathcal{A}|$ for its cardinality. 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 ($\boldsymbol{X}_{\mathcal{A}}$). Finally, $\boldsymbol{X} \succeq 0$ is used to mean $\boldsymbol{X}$ is a positive semi-definite (PSD) matrix, 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 $\boldsymbol{b} \in \mathbb{R}^n$ [23].

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 values at each node:

$$\boldsymbol{x} = \begin{bmatrix} \sigma(v_1) & \cdots & \sigma(v_n) \end{bmatrix}^T, \quad v_i \in \mathcal{V}. \tag{1}$$

Of interest to GSP is the spectral representation of (1), which depends on the graph on which $\boldsymbol{x}$ is supported. 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., that there exist $\boldsymbol{V}$ orthogonal and $\boldsymbol{\Sigma}$ diagonal such that $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T$ [24]. Then, the *graph Fourier transform* of $\boldsymbol{x}$ is given by $\bar{\boldsymbol{x}} = \boldsymbol{V}^T\boldsymbol{x}$ [1, 2]. Note that if $\boldsymbol{A}$ is not normal, spectral energy conservation properties analog to Parseval's theorem in classical signal processing no longer hold.

Similar traditional signal processing, a graph signal $\boldsymbol{x}$ is said to be bandlimited when its spectral representation is sparse. Explicitly, $\boldsymbol{x}$ is *$\mathcal{K}$-bandlimited* if $\bar{\boldsymbol{x}}$ is $\mathcal{K}$-sparse, i.e., $\bar{\boldsymbol{x}}_{\mathcal{V}\setminus\mathcal{K}}$ is a zero vector. Then,

$$\boldsymbol{x} = \boldsymbol{V}_{\mathcal{K}}\bar{\boldsymbol{x}}_{\mathcal{K}}. \tag{2}$$

Note that in classical signal processing, a signal is considered bandlimited when its spectral representation has finite support ("low-pass"). Although there have been works exploiting graph frequency orderings based on the eigenvalues of $\boldsymbol{A}$ [7, 11, 12], we abuse the term bandlimited since we do not rely on any such orderings.

In the sequel, we consider the class of bandlimited graph signals that are *stationary* random processes with respect to $\mathbb{G}$, i.e., for which $\bar{\boldsymbol{x}}_{\mathcal{K}}$ in (2) is a zero-mean random vector with $\boldsymbol{\Lambda} = \mathbb{E}\,\bar{\boldsymbol{x}}_{\mathcal{K}}\bar{\boldsymbol{x}}_{\mathcal{K}}^T = \operatorname{diag}\{\lambda_i\}$ [25–27]. Without loss of generality, we assume $\boldsymbol{\Lambda}$ is full-rank. Otherwise, remove from $\mathcal{K}$ any element for which $\lambda_i = 0$. Moreover, we assume that we only have access to a noisy version of the graph signal of interest:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}, \tag{3}$$

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 (3) is related to the concept of "approximately bandlimited" graph signal introduced in [7].

3. SAMPLING AND RECONSTRUCTION OF BANDLIMITED GRAPH SIGNALS

Consider sampling to be the operation of observing the value of a graph signal on a subset of $\mathcal{V}$. Formally, let $\mathcal{S} \subseteq \mathcal{V}$ denote the sampling set and $\boldsymbol{y}_{\mathcal{S}}$ be the vector that collects the samples of $\boldsymbol{y}$. To clarify the derivations, define the selection matrix $\boldsymbol{C} \in \{0,1\}^{|\mathcal{S}|\times N}$ composed by the rows of the identity matrix with indices in $\mathcal{S}$, so that

$$\boldsymbol{y}_{\mathcal{S}} = \boldsymbol{C}\boldsymbol{y}. \tag{4}$$

The samples $\boldsymbol{y}_{\mathcal{S}}$ from (4) are only useful inasmuch as they represent $\boldsymbol{x}$. A straightforward way to evaluate a sampling set is therefore through our ability to estimate $\boldsymbol{x}$ from it. Indeed, if $\boldsymbol{x}$ can be perfectly reconstructed from its samples, then no "information" is lost during sampling. Recall that we consider $\boldsymbol{x}$ to be $\mathcal{K}$-bandlimited as in (2) and assume that $\mathcal{K}$ is known.

Conditions under which exact reconstruction is feasible in the noiseless case [$\boldsymbol{w} = \boldsymbol{0}$ in (3)] were derived in [9–12]. Due to the noise in (3), however, $\boldsymbol{x}$ cannot be perfectly reconstructed, so we turn to the problem of approximating it. Explicitly, for $\boldsymbol{L} \in \mathbb{R}^{N\times|\mathcal{S}|}$, let

$$\hat{\boldsymbol{x}} = \boldsymbol{L}\boldsymbol{y}_{\mathcal{S}} \tag{5}$$

be an estimate of $\boldsymbol{x}$ based on $\boldsymbol{y}_{\mathcal{S}}$ whose error covariance matrix is

$$\boldsymbol{K}(\hat{\boldsymbol{x}}) = \mathbb{E}(\boldsymbol{x} - \hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}})^T. \tag{6}$$

The operator $\boldsymbol{L}$ is sometimes called a *linear interpolation operator* [7, 11, 12]. We then seek $\hat{\boldsymbol{x}}^\star = \boldsymbol{L}^\star\boldsymbol{y}_{\mathcal{S}}$ such that $\boldsymbol{K}(\hat{\boldsymbol{x}}^\star) \preceq \boldsymbol{K}(\hat{\boldsymbol{x}})$ for all $\hat{\boldsymbol{x}}$ as in (5). Note that this problem is more general than the typical least-squares estimation, since the former criterion subsumes the MSE. Indeed, $\operatorname{MSE}(\hat{\boldsymbol{x}}) = \mathbb{E}\,\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2 = \operatorname{Tr}[\boldsymbol{K}(\hat{\boldsymbol{x}})]$ and $\boldsymbol{K}(\hat{\boldsymbol{x}}^\star) \preceq \boldsymbol{K}(\hat{\boldsymbol{x}}) \Rightarrow \operatorname{MSE}(\hat{\boldsymbol{x}}^\star) \leq \operatorname{MSE}(\hat{\boldsymbol{x}})$ [23]. This generality allows us to study the reconstruction properties of sampling sets for different *scalarizations* of (6) in Section 4.

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 $\boldsymbol{b} \in \mathbb{R}^n$ [23], where we replaced $\hat{\boldsymbol{x}}$ by its expression in (5). Then, since $\boldsymbol{x}$ is a bandlimited stationary process on $\mathbb{G}$,

$$\begin{aligned} J(\boldsymbol{L}) &= \boldsymbol{b}^T\,\mathbb{E}(\boldsymbol{x} - \hat{\boldsymbol{x}})(\boldsymbol{x} - \hat{\boldsymbol{x}})^T\boldsymbol{b} \\ &= \boldsymbol{b}^T\,\mathbb{E}(\boldsymbol{V}_{\mathcal{K}}\bar{\boldsymbol{x}}_{\mathcal{K}} - \boldsymbol{L}\boldsymbol{y}_{\mathcal{S}})(\boldsymbol{V}_{\mathcal{K}}\bar{\boldsymbol{x}}_{\mathcal{K}} - \boldsymbol{L}\boldsymbol{y}_{\mathcal{S}})^T\boldsymbol{b} \\ &= \boldsymbol{b}^T\Big[\boldsymbol{V}_{\mathcal{K}}\boldsymbol{\Lambda}\boldsymbol{V}_{\mathcal{K}}^T - \boldsymbol{L}\boldsymbol{C}\boldsymbol{V}_{\mathcal{K}}\boldsymbol{\Lambda}\boldsymbol{V}_{\mathcal{K}}^T - \boldsymbol{V}_{\mathcal{K}}\boldsymbol{\Lambda}\boldsymbol{V}_{\mathcal{K}}^T\boldsymbol{C}^T\boldsymbol{L}^T \\ &\qquad + \boldsymbol{L}\boldsymbol{C}(\boldsymbol{V}_{\mathcal{K}}\boldsymbol{\Lambda}\boldsymbol{V}_{\mathcal{K}}^T + \boldsymbol{\Lambda}_w)\boldsymbol{C}^T\boldsymbol{L}^T\Big]\,\boldsymbol{b}. \end{aligned} \tag{7}$$

Setting the derivative of (7) 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

$$\boldsymbol{L}^\star\boldsymbol{C}\left(\boldsymbol{V}_{\mathcal{K}}\boldsymbol{\Lambda}\boldsymbol{V}_{\mathcal{K}}^T + \boldsymbol{\Lambda}_w\right)\boldsymbol{C}^T = \boldsymbol{V}_{\mathcal{K}}\boldsymbol{\Lambda}\boldsymbol{V}_{\mathcal{K}}^T\boldsymbol{C}^T. \tag{8}$$

Note that due to the presence of noise, the matrix in parenthesis in (8) is always invertible. This is similar to the well-known regularization effect of noise in Kalman filtering [23]. Nevertheless, (8) holds even when $\boldsymbol{\Lambda}_w = \boldsymbol{0}$, 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 [9–12].

For $\boldsymbol{L}^\star$ satisfying (8), the error covariance matrix (6) becomes

$$\boldsymbol{K}(\hat{\boldsymbol{x}}^\star) = \boldsymbol{V}_{\mathcal{K}}\left(\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}_{\mathcal{K}}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Lambda}_w^{-1}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{V}_{\mathcal{K}}\right)^{-1}\boldsymbol{V}_{\mathcal{K}}^T, \tag{9}$$

where we used the matrix inversion lemma [24] and the fact that $(\boldsymbol{C}\boldsymbol{\Lambda}_w\boldsymbol{C}^T)^{-1} = \boldsymbol{C}\boldsymbol{\Lambda}_w^{-1}\boldsymbol{C}^T$. Moreover, given that we assume that $\boldsymbol{x}$ lies in the column span of $\boldsymbol{V}_{\mathcal{K}}$, we can focus on statistics of the error along these directions, i.e.,

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

4. NEAR-OPTIMAL SAMPLING SET SELECTION

It is clear from Section 3 how the sampling set influences the graph signal reconstruction: $\mathcal{S}$ determines $\boldsymbol{C}$ in (8), which gives the optimal interpolator $\boldsymbol{L}^\star$, that is used in (5) to provide an optimal estimate of $\boldsymbol{x}$, whose error covariance matrix is given by (10). In fact, (10) can be written as a matrix-valued function of the sampling set:

$$\bar{\boldsymbol{K}}(\mathcal{S}) = \left[\boldsymbol{\Lambda}^{-1} + \sum_{i\in\mathcal{S}}\lambda_{w,i}^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^T\right]^{-1}, \tag{11}$$

where $\boldsymbol{v}_i^T$ is the i-th row of $\boldsymbol{V}_{\mathcal{K}}$.

However, (11) cannot be used directly to inform the choice of $\mathcal{S}$. Indeed, (11) is a matrix-valued cost function whose minimization may have several *minimal* solutions. This issue is straightforward to address using a scalarization of $\bar{\boldsymbol{K}}(\mathcal{S})$, such as its trace or determinant [28]. Still, minimizing (11) remains a combinatorial problem: selecting a sampling set of size $k \approx |\mathcal{K}| \ll n$ would require $\binom{n}{k}$ evaluations of (11), which is impractical even for moderately small n. Under some conditions, however, it is possible to find a guaranteed near-optimal minimizer of (11). These are explored in the sequel.

4.1. Approximate supermodularity and greedy minimization

Although set function optimization is in general NP-hard [16], there are cases in which the optimal solution can be approximated in polynomial (and even linear) time. In particular, a theorem due to Nemhauser, Wolsey, and Fisher shows that a greedy search yields guaranteed near-optimal results for the minimization of set functions that are monotonically decreasing and supermodular [29]. However, the reconstruction MSE, i.e., the trace of (11), is not supermodular in general. This can be seen from [19, Thm. 2.4] and the fact that $t \mapsto t^{-2}$ is not operator antitone [30]. Still, we can provide near-optimal guarantees by introducing the concept of *approximate supermodularity*.

Formally, a set function $f : 2^{\mathcal{V}} \to \mathbb{R}$ is *approximately supermodular* or *α-supermodular* if for all sets $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and all $u \notin \mathcal{B}$ it holds that

$$f\left(\mathcal{A} \cup \{u\}\right) - f\left(\mathcal{A}\right) \leq \alpha \left[f\left(\mathcal{B} \cup \{u\}\right) - f\left(\mathcal{B}\right)\right], \tag{12}$$

for $0 \leq \alpha \leq 1$. Notice that for $\alpha = 1$, (12) reduces to the traditional definition of supermodularity, in which case we refer to the function simply as *supermodular* [16,21]. Also, (12) always holds for $\alpha = 0$ if f is monotone decreasing. Indeed, a set function f is *monotone decreasing* if for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ it holds that $f(\mathcal{A}) \geq f(\mathcal{B})$. It is *monotone increasing* if $-f$ is monotone decreasing. Thus, α-supermodularity is only of interest when α takes the largest value for which (12) holds, i.e.,

$$\alpha = \min_{\substack{\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \\ v \notin \mathcal{B}}} \frac{f(\mathcal{A} \cup \{v\}) - f(\mathcal{A})}{f(\mathcal{B} \cup \{v\}) - f(\mathcal{B})}. \tag{13}$$

It is worth noting that α is related to the *submodularity ratio* defined in [20]. Given these definitions, the following holds:

Theorem 1. *Let $f^\star = f(\mathcal{S}^\star)$ be the optimal value of the problem*

$$\underset{\mathcal{S} \subseteq \mathcal{V},\ |\mathcal{S}| = k}{\text{minimize}} \quad f\left(\mathcal{S}\right) \tag{14}$$

and $\mathcal{G}_\ell$ be the ℓ-th iteration of its greedy solution, obtained by taking $\mathcal{G}_0 = \{\}$ and repeating ℓ times

$$u = \text{argmin}_{s \in \mathcal{V}} f\left(\mathcal{G}_{j-1} \cup \{s\}\right) \tag{15a}$$

$$\mathcal{G}_j = \mathcal{G}_{j-1} \cup \{u\} \text{ and } \mathcal{V} = \mathcal{V} \setminus \{u\} \tag{15b}$$

If f is (i) monotone decreasing and (ii) α-supermodular, then

$$\frac{f(\mathcal{G}_\ell) - f^\star}{f(\{\}) - f^\star} \leq \left(1 - \frac{\alpha}{k}\right)^{\ell} \leq e^{-\alpha \ell / k}. \tag{16}$$

Proof. Using the fact that f is monotone decreasing, it holds for every set $\mathcal{G}_j$ that

$$\begin{aligned} f(\mathcal{S}^\star) &\geq f(\mathcal{S}^\star \cup \mathcal{G}_j) \\ &= f(\mathcal{G}_j) + \sum_{i=1}^{k} f(v_i^\star \cup \mathcal{T}_{i-1}) - f(\mathcal{T}_{i-1}), \end{aligned} \tag{17}$$

where $\mathcal{T}_i = \mathcal{G}_j \cup \{v_1^\star, \dots, v_i^\star\}$ and $v_i^\star$ is the i-th element of $\mathcal{S}^\star$. Since f is α-supermodular and $\mathcal{G}_j \subset \mathcal{T}_i$ for all i, the incremental gains in the summation in (17) can be bounded using (12) to get

$$f(\mathcal{S}^\star) \geq f(\mathcal{G}_j) + \alpha^{-1} \sum_{i=1}^{k} f(v_i^\star \cup \mathcal{G}_j) - f(\mathcal{G}_j).$$

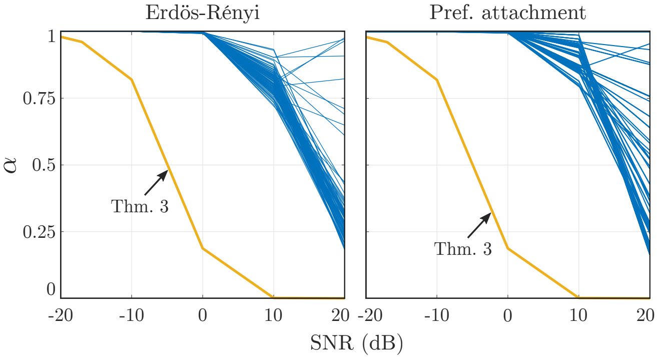

Fig. 1. Comparison between the bound in Theorem 3 and α

Finally, given that $\mathcal{G}_{j+1}$ is chosen as to minimize the incremental gain in (15),

$$f(\mathcal{S}^\star) \geq f(\mathcal{G}_j) + \alpha^{-1} k \left[f(\mathcal{G}_{j+1}) - f(\mathcal{G}_j)\right]. \tag{18}$$

To obtain a recursion, let $\delta_j = f(\mathcal{G}_j) - f(\mathcal{S}^\star)$ so that (18) becomes

$$\delta_j \leq \alpha^{-1} k \left[\delta_j - \delta_{j+1}\right] \Rightarrow \delta_{j+1} \leq \left(1 - \frac{1}{\alpha^{-1} k}\right) \delta_j.$$

Noting that $\delta_0 = f(\{\}) - f(\mathcal{S}^\star)$, we can solve this recursion to get

$$\frac{f(\mathcal{G}_\ell) - f(\mathcal{S}^\star)}{f(\{\}) - f(\mathcal{S}^\star)} \leq \left(1 - \frac{\alpha}{k}\right)^{\ell}.$$

Using the fact that $1 - x \leq e^{-x}$ yields (16). ■

Theorem 1 bounds the relative suboptimality of the greedy solution to problem (14) when f is decreasing and α-supermodular. Under these conditions, it guarantees a minimum improvement of the greedy solution over the empty set. What is more, it quantifies the effect of relaxing the supermodularity hypothesis in (12). Indeed, when f is supermodular ($\alpha = 1$) and the greedy search in (15) is repeated k times ($\ell = k$), we recover the $e^{-1} \approx 0.37$ guarantee from [29]. On the other hand, if f is not supermodular ($\alpha < 1$), (16) shows that the same 37% guarantee can be obtained by greedily selecting a set of size k/α. Thus, α not only quantifies how much f violates supermodularity, but also gives a factor by which a solution set must grow to maintain supermodular near-optimality. It is worth noting that, as with the original bound in [29], (16) is not tight and that better results are common in practice (see Section 5).

In the sequel, we show that two important scalarizations of (11), namely $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ and $\text{MSE}(\mathcal{S}) = \text{Tr}[\bar{\boldsymbol{K}}(\mathcal{S})]$, are monotone decreasing and α-supermodular functions of the sampling set $\mathcal{S}$. For the $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$, we prove that it is supermodular, i.e., that (12) holds with $\alpha = 1$. For the MSE, we provide an explicit lower bound on α as a function of the problem signal-to-noise ratio (SNR). These results simultaneously provide near-optimal performance guarantees based on Theorem 1 and shed light on why greedy algorithms have been so successful in GSP applications.

4.2. α-supermodular scalarizations for sampling set selection

To make the following derivations more tractable, we assume that signal and noise are homoscedastic, i.e., $\boldsymbol{\Lambda} = \sigma_x^2 \boldsymbol{I}$ and $\boldsymbol{\Lambda}_w = \sigma_w^2 \boldsymbol{I}$. Then, (11) can be rewritten as

$$\bar{\boldsymbol{K}}(\mathcal{S}) = \sigma_x^2 \left[\boldsymbol{I} + \gamma \boldsymbol{W}(\mathcal{S})\right]^{-1}, \tag{19}$$

where $\gamma = \sigma_x^2 / \sigma_w^2$ is the SNR and $\boldsymbol{W}(\mathcal{S}) = \sum_{i \in \mathcal{S}} \boldsymbol{v}_i \boldsymbol{v}_i^T$. Then, we can state our results as:

Theorem 2. *The scalar set functions* $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ *is (i) monotone decreasing and (ii) supermodular.*

Theorem 3. *The scalar set functions* $\mathrm{MSE}(\mathcal{S}) = \mathrm{Tr}[\bar{\boldsymbol{K}}(\mathcal{S})]$ *is (i) monotone decreasing and (ii)* α*-supermodular with*

$$\alpha \geq \frac{1+2\gamma}{(1+\gamma)^4}, \quad \textit{for } \gamma = \frac{\sigma_x^2}{\sigma_w^2}. \tag{20}$$

Theorems 2 and 3 establish that a near-optimal solution to the sampling set selection problem can be obtained efficiently when the figure of merit is the $\log\det$ or the MSE. To be sure, the $\log\det$ function is typically used in statistics and experimental design as an alternative to the MSE due to its well-known supermodular characteristics [16, 18–21]. It is also common in the sensor placement literature due to its relation to information theoretic measures, such as entropy and mutual information [16–18]. Furthermore, when $\bar{\boldsymbol{x}}$ and $\boldsymbol{w}$ in (2) and (3) are Gaussian, $\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ is proportional to the volume of the confidence ellipsoids of $\hat{\bar{\boldsymbol{x}}}^\star$ [31].

In contrast, the MSE is not supermodular in general. For instance, restrictive and often unrealistic conditions on data distribution are required to obtain supermodularity in the context of regression [20]. Nevertheless, there is strong empirical evidence that greedily optimizing the MSE yields good results in several contexts, such as regression, dictionary learning, and graph signal sampling [9–12, 15, 20, 20, 32]. Theorem 3 seeks to reconcile these observations by bounding the suboptimality of greedy sampling as a function of the SNR.

As in the case of the greedy bound in Theorem 1, (20) heavily underestimates the value of α. Nevertheless, it give insights into the behavior of the parameter. Indeed, as $\gamma \to \infty$ and we approach the noiseless case, $\alpha \to 0$. This is expected as in the noiseless case almost every set of size $|\mathcal{K}|$ achieves perfect reconstruction, so that the choice of sampling nodes is irrelevant. On the other hand, $\alpha \to 1$ as $\gamma \to 0$, i.e., the MSE becomes closer to supermodular as the noise increases. Given that reconstruction errors are small for high SNR, Theorem 1 guarantees that greedy sampling performs well when it is most needed. These observations are illustrated in Fig. 1 that compares the bound in Theorem 3 to the true value of α for the MSE (found by exhaustive search) in 100 realizations of Erdös-Rényi and preferential attachment graphs (see Section 5).

Before proceeding with the proof of Theorems 2 and 3, start with the following more general result:

Lemma 1. *The matrix* $\bar{\boldsymbol{K}}(\mathcal{S})$ *in* (11) *is a monotonically decreasing set function with respect to the PSD cone.*

Proof. First, note that $\bar{\boldsymbol{K}}(\mathcal{S})$ depends on $\mathcal{S}$ only through $\boldsymbol{W}(\mathcal{S})$ and that $\boldsymbol{W}(\cdot)$ is additive, i.e., $\boldsymbol{W}(\mathcal{A} \cup \mathcal{B}) = \boldsymbol{W}(\mathcal{A}) + \boldsymbol{W}(\mathcal{B})$. Then, since $\boldsymbol{W}(\cdot)$ is a sum of PSD matrices, it is straightforward from (19) that $\mathcal{A} \subseteq \mathcal{B} \Rightarrow \boldsymbol{W}(\mathcal{A}) \preceq \boldsymbol{W}(\mathcal{B})$. From the antitonicity of the matrix inverse [30] and the fact that $\gamma \geq 0$, it follows that $\bar{\boldsymbol{K}}(\mathcal{A}) \succeq \bar{\boldsymbol{K}}(\mathcal{B})$. ■

It is straightforward to see that part (i) of Theorems 2 and 3 are corollaries of Lemma 1. Also, Lemma 1 implies that the unconstrained minimization of either $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ and $\mathrm{Tr}[\bar{\boldsymbol{K}}(\mathcal{S})]$ yields the trivial solution $\mathcal{S}^\star = \mathcal{V}$. However, the optimization problem of interest (14) contains a cardinality constraint that makes it combinatorial.

Proof of Theorem 2. Having established part (i) using Lemma 1, we proceed to show that $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ is supermodular [part (ii)]. To do so, take $\alpha = 1$ in (12) and notice that it now implies that $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ is supermodular if and only if the induced set function $f_v : 2^{\mathcal{V}\setminus\{v\}} \to \mathbb{R}$ defined as

$$f_u(\mathcal{S}) = \log\det\left[\bar{\boldsymbol{K}}\left(\mathcal{S} \cup \{u\}\right)\right] - \log\det\left[\bar{\boldsymbol{K}}\left(\mathcal{S}\right)\right]$$

is monotone increasing [16, 21]. To show this is indeed the case, let $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V} \setminus \{u\}$ and use the additivity of $\boldsymbol{W}(\cdot)$ to define the homotopy

$$\begin{aligned} h(t) &= \log\det\left[\sigma_x^2\left(\boldsymbol{Y}(t) + \gamma\boldsymbol{W}(\{u\})\right)^{-1}\right] \\ &\quad - \log\det\left[\sigma_x^2\boldsymbol{Y}(t)^{-1}\right] \\ h(t) &= \log\det\left[\left(\boldsymbol{Y}(t) + \gamma\boldsymbol{W}(\{u\})\right)^{-1}\right] \\ &\quad - \log\det\left[\boldsymbol{Y}(t)^{-1}\right] \end{aligned} \tag{21}$$

for $0 \leq t \leq 1$ and $\boldsymbol{Y}(t) = \boldsymbol{I} + \boldsymbol{W}(\mathcal{A}) + \gamma t\left[\boldsymbol{W}(\mathcal{B}) - \boldsymbol{W}(\mathcal{A})\right]$. Note that $h(0) = f_u(\mathcal{A})$ and $h(1) = f_u(\mathcal{B})$. Thus, we can write

$$f_u(\mathcal{B}) = f_u(\mathcal{A}) + \int_0^1 h'(t)dt,$$

where $h'(t)$ is the derivative of $h(t)$ with respect to t. Clearly, if $h'(t) \geq 0$ for $t \in [0, 1]$, then $f_u(\mathcal{S})$ is an increasing set functions and $\log\det[\bar{\boldsymbol{K}}(\mathcal{S})]$ is supermodular.

Surely, since $\frac{d}{dt}\log\det(\boldsymbol{X}(t)^{-1}) = -\mathrm{Tr}\left[\boldsymbol{X}(t)^{-1}\frac{d}{dt}\boldsymbol{X}(t)\right]$, the derivative of (21) yields

$$\begin{aligned} h'(t) &= \gamma\,\mathrm{Tr}\left\{\boldsymbol{Y}(t)^{-1}\left[\boldsymbol{W}(\mathcal{B}) - \boldsymbol{W}(\mathcal{A})\right]\right\} \\ &\quad - \gamma\,\mathrm{Tr}\left\{\left(\boldsymbol{Y}(t) + \gamma\boldsymbol{W}(\{u\})\right)^{-1}\left[\boldsymbol{W}(\mathcal{B}) - \boldsymbol{W}(\mathcal{A})\right]\right\} \\ &= \gamma\,\mathrm{Tr}\left\{\left[\boldsymbol{Y}(t)^{-1} - \left(\boldsymbol{Y}(t) + \gamma\boldsymbol{W}(\{u\})\right)^{-1}\right]\right. \\ &\qquad\qquad \left.\times\left[\boldsymbol{W}(\mathcal{B}) - \boldsymbol{W}(\mathcal{A})\right]\right\} \geq 0, \end{aligned}$$

which is non-negative because it is the trace of a product of PSD matrices [24]. Indeed, from Lemma 1, $\mathcal{A} \subseteq \mathcal{B} \Rightarrow \boldsymbol{W}(\mathcal{B}) \succeq \boldsymbol{W}(\mathcal{A})$. Also, it is straightforward from $\boldsymbol{Y}(t) \succeq 0$, $\boldsymbol{W}(\{v\}) \succeq 0$, and $\gamma \geq 0$ that $\boldsymbol{Y}(t) \preceq \boldsymbol{Y}(t) + \gamma\boldsymbol{W}(\{u\})$. Since matrix inversion is an antitone function [30], it follows that $\boldsymbol{Y}(t)^{-1} \succeq (\boldsymbol{Y}(t) + \gamma\boldsymbol{W}(\{v\}))^{-1}$. ■

Proof of Theorem 3. Once again, part (i) is established from Lemma 1. To obtain part (ii), we need to lower bound α in (13). To do so, start by using the matrix inversion lemma to derive a closed form expression for the increments in (13):

$$\begin{aligned} f(\mathcal{A} \cup \{u\}) - f(\mathcal{A}) &= \sigma_x^2\,\mathrm{Tr}\left[\left(\boldsymbol{Z}(\mathcal{A}) + \gamma\boldsymbol{v}_u\boldsymbol{v}_u^T\right)^{-1} - \boldsymbol{Z}(\mathcal{A})^{-1}\right] \\ &= -\sigma_x^2\,\mathrm{Tr}\left[\frac{\boldsymbol{Z}(\mathcal{A})^{-1}\boldsymbol{v}_u\boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{A})^{-1}}{\gamma^{-1} + \boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{A})^{-1}\boldsymbol{v}_u}\right] \\ &= -\sigma_x^2 \cdot \frac{\boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{A})^{-2}\boldsymbol{v}_u}{\gamma^{-1} + \boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{A})^{-1}\boldsymbol{v}_u}, \end{aligned}$$

with $\boldsymbol{Z}(\mathcal{A}) = \boldsymbol{I} + \gamma\sum_{i\in\mathcal{A}}\boldsymbol{v}_i\boldsymbol{v}_i^T$. Therefore, (13) becomes

$$\alpha = \min_{\substack{\mathcal{A}\subseteq\mathcal{B}\subseteq\mathcal{V} \\ v\notin\mathcal{B}}} \frac{\gamma^{-1} + \boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{B})^{-1}\boldsymbol{v}_u}{\gamma^{-1} + \boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{A})^{-1}\boldsymbol{v}_u} \times \frac{\boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{A})^{-2}\boldsymbol{v}_u}{\boldsymbol{v}_u^T\boldsymbol{Z}(\mathcal{B})^{-2}\boldsymbol{v}_u}. \tag{22}$$

To bound (22), first notice that $\lambda_{\max}\left[\boldsymbol{Z}(\mathcal{C})\right] \leq 1 + \gamma$ and $\lambda_{\min}\left[\boldsymbol{Z}(\mathcal{C})\right] \geq 1$ for any set $\mathcal{C}$. These bounds are achieved for $\mathcal{V}$

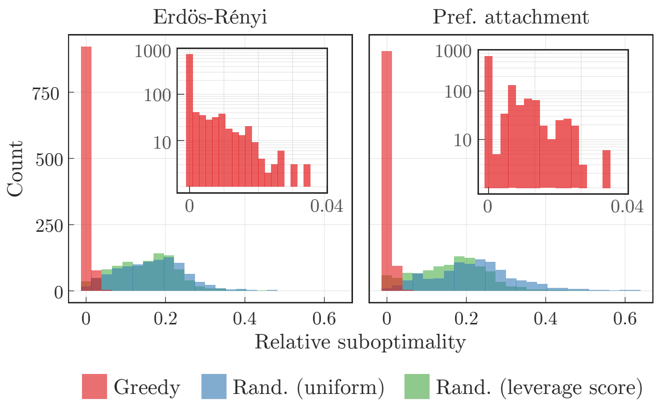

Fig. 2. Relative suboptimality of sampling schemes (log det)

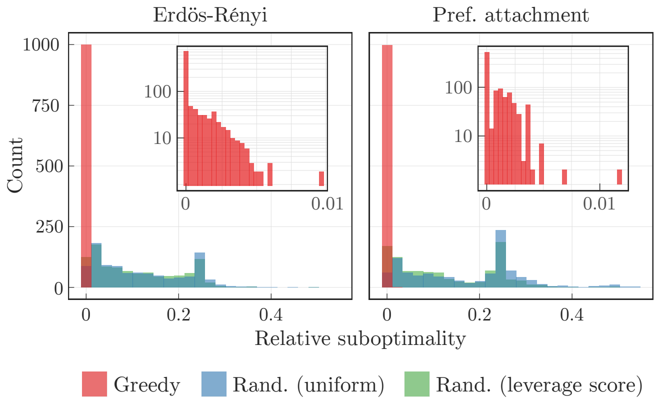

Fig. 3. Relative suboptimality of sampling schemes (MSE)

and the empty set, respectively. Then, using the Rayleigh quotient inequalities [24], (22) can be bounded by

$$\begin{aligned}\alpha &\geq \frac{\gamma^{-1} + \|\boldsymbol{v}_u\|_2^2 (1+\gamma)^{-1}}{\gamma^{-1} + \|\boldsymbol{v}_u\|_2^2} \times \frac{\|\boldsymbol{v}_u\|_2^2 (1+\gamma)^{-2}}{\|\boldsymbol{v}_u\|_2^2} \\ &= \frac{\gamma^{-1} + 1 + \|\boldsymbol{v}_u\|_2^2}{\gamma^{-1} + \|\boldsymbol{v}_u\|_2^2} \cdot (1+\gamma)^{-3} \triangleq \alpha'. \qquad (23)\end{aligned}$$

Finally, to obtain the expression in Theorem 3, notice that (23) is decreasing with respect to $\|\boldsymbol{v}_u\|_2^2$. Indeed,

$$\frac{\partial \alpha'}{\partial \|\boldsymbol{v}_u\|_2^2} = \frac{1-(1+\gamma)^3}{(\gamma^{-1}+\|\boldsymbol{v}_u\|_2^2)^2(1+\gamma)^6} \leq 0.$$

Since $\boldsymbol{v}_u$ is a row of $\boldsymbol{V}_{\mathcal{K}}$, i.e., it is composed of a subset of elements of a unit vector, $\|\boldsymbol{v}_u\|_2^2 \leq 1$ and we obtain the desired result. ■

5. SIMULATIONS

This section presents numerical results for the reconstruction performance of three sampling set selection schemes: *greedy* from (15) and the *uniform* and *leverage score* randomized methods from [7]. For illustration purposes, undirected graphs were simulated using the *Erdös-Rényi* model, in which an edge is placed between two nodes with probability $p = 0.2$, and the *preferential attachment* model [33], in which nodes are added one at a time and connected to a node already in the graph with probability proportional to its degree. To evaluate the *relative suboptimality* measure in (16), we use graphs with $n = 10$ nodes since the optimal sampling set can only be found by exhaustive search. The bandlimited graph signals were then generated taking $\boldsymbol{V}_{\mathcal{K}}$ in (2) to be the eigenvectors of the graph adjacency matrix relative to the four eigenvalues with largest magnitude ($|\mathcal{K}| = 4$). The random vectors $\bar{\boldsymbol{x}}$ in (2) and $\boldsymbol{w}$ in (3) were taken to be zero-mean Gaussian random variables with covariance matrices $\boldsymbol{\Lambda} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}_w = \sigma_w^2 \boldsymbol{I}$, with $\sigma_w^2 = 10^{-2}$ (SNR = 20 dB). The size of the sampling set is taken to be $|\mathcal{S}| = |\mathcal{K}| = 4$.

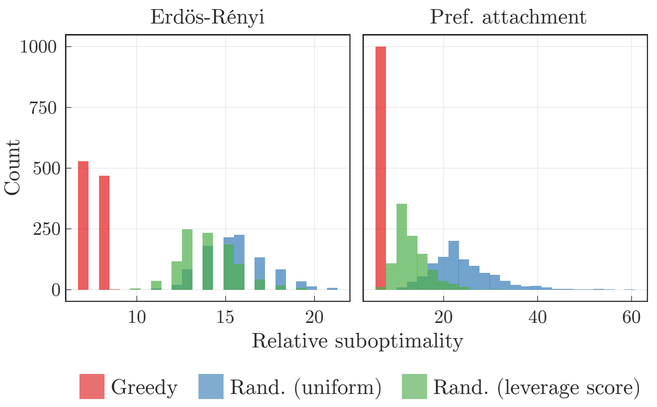

Fig. 4. Sampling set size for 90% reduction of MSE over empty set

Fig. 2 and 3 display histograms of the relative suboptimality for 1000 realizations of graphs and graph signals using the log det and MSE cost functions, respectively. Note that greedy sampling always obtains a sampling set with a relative suboptimality bounded by Theorems 1, 2, and 3. In fact, it typically behaves much better (see details in Fig. 2 and 3). In these experiments, it obtained the optimal set at least 40% of the time. Randomized sampling schemes, however, do not perform as well for single problem instances, since these methods are more appropriate when several sampling sets of the graph signal are considered. Indeed, the performance measures in [7] hold in expectation over sampling realizations. Note that leverage score sampling is shown in Fig. 2 for completeness, since it is an approximation of the optimal sampling distribution for the MSE [7].

Since evaluating the relative suboptimality for larger graphs is untractable, we turn to measuring the sampling set size required to yield a certain MSE reduction. Fig. 4 displays the distribution of the sampling set size required to achieve a 90% reduction in the MSE with respect to the empty set. The plots were obtained from 1000 graphs and signals realizations with $n = 100$ nodes and $\boldsymbol{V}_{\mathcal{K}}$ in (2) containing the eigenvectors relative to the seven eigenvalues with largest magnitude ($|\mathcal{K}| = 7$). Note that, although Theorem 3 estimates that the MSE would require sets considerably larger to recover the same near-optimal guarantees as supermodular functions, greedy sampling obtained a sampling set of size $|\mathcal{K}|$ in more than 50% of the realizations. Moreover, as noted in [7], leverage score sampling has similar performance to uniform sampling for Erdös-Rényi graphs, but gives better results for the preferential attachment model.

6. CONCLUSION

This work addressed the issue of sampling set selection in the context of GSP. It started by deriving the optimal graph signal reconstruction (interpolation) operator for any given sampling set. Then, it showed that, although the sampling set selection problem is combinatorial, its solution can be approximated using greedy algorithms. By introducing the concept of approximate supermodularity, near-optimal guarantees were given for two important reconstruction figures of merit, namely the $\log\det$ of the error covariance matrix and the MSE, thus justifying the success of greedy sampling schemes.

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," 2016, arXiv:1512.05405v2.

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

[10] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," 2016, arXiv:1507.08822v3.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] A. Krause and D. Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014.

[17] 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.

[18] 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.

[19] 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.

[20] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *Int. Conf. on Mach. Learning*, 2011.

[21] F. Bach, "Learning with submodular functions: A convex optimization perspective," *Foundations and Trends in Machine Learning*, vol. 6[2-3], pp. 145–373, 2013.

[22] L.F.O. Chamon and A. Ribeiro, "Greedy sampling of graph signals," 2016, Available at http://www.seas.upenn.edu/∼luizf.

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

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

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

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

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

[28] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[29] 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.

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

[31] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57[2], pp. 451–462, 2009.

[32] A. Krause and V. Cevher, "Submodular dictionary selection for sparse representation," in *Int. Conf. on Mach. Learning*, 2010.

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