Rethinking Sketching as Sampling: Linear Transforms of Graph Signals

Fernando Gama[†], Antonio G. Marques[‡], Gonzalo Mateos[*] and Alejandro Ribeiro[†]

***Abstract*—Sampling of bandlimited graph signals has well-documented merits for dimensionality reduction, affordable storage, and online processing of streaming network data. Most existing sampling methods are designed to minimize the error incurred when reconstructing the original signal from its samples. Oftentimes these parsimonious signals serve as inputs to computationally-intensive linear transformations (e.g., graph filters). Hence, interest shifts from reconstructing the signal itself towards instead approximating the output of the prescribed linear operator efficiently. In this context, we propose a novel sampling scheme that leverages the bandlimitedness of the input as well as the transformation whose output we wish to approximate. We formulate problems to jointly optimize sample selection and a sketch of the target linear transformation, so when the latter is affordably applied to the sampled input signal the result is close to the desired output. The developed sampling plus reduced-complexity processing pipeline is particularly useful for streaming data, where the linear transform has to be applied fast and repeatedly to successive inputs.**

***Index Terms*—Sketching, sampling, graph signal processing, streaming, linear transforms**

I. Introduction

Propelled by the desire of analyzing and processing network data supported on irregular domains, there has been a growing interest in broadening the scope of traditional signal processing techniques to signals defined on graphs [1], [2]. Noteworthy representatives include sampling of graph signals, linear graph filtering and the graph Fourier transform (GFT) [3], [4], all of them instances of linear problems. This is not surprising since linear models are ubiquitous in science and engineering, due in part to their generality, conceptual simplicity, and mathematical tractability. Along with heterogeneity and lack of regularity, data are increasingly high dimensional and this curse of dimensionality not only raises statistical challenges, but also major computational hurdles even for linear models. In particular, these limiting factors can hinder processing of streaming data, where say a massive linear operator has to be repeatedly and efficiently applied to a sequence of input (graph) signals [5]. These Big Data challenges motivated a recent body of work collectively addressing so-termed *sketching* problems [6], which seek computationally-efficient solutions to a subset of (typically inverse) linear problems. The basic idea is to *draw a sketch* of the linear model such that the resulting linear transform is lower dimensional, while still offering quantifiable approximation error guarantees. To achieve this, a fat random projection matrix is designed to pre-multiply and reduce the dimensionality of the linear operator matrix, in such way that the resulting matrix sketch retains the most important structure of the model. The input vector has to be adapted to the sketched operator as well, and to that end the same random projections are applied to the signal in a way often agnostic to the input statistics.

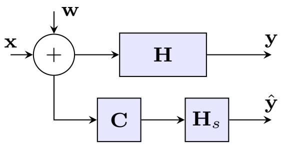


Fig. 1. Knowing the linear transformation $\mathbf{H}$ and having access to a stream of noisy inputs, we want to design the sampling matrix $\mathbf{C}$ and the reduced linear transformation (sketch) $\mathbf{H}_s$ so that $\hat{\mathbf{y}}$ approximates $\mathbf{y}$.

Although random projection methods offer an elegant dimensionality-reduction alternative for several Big Data problems, they face some shortcomings: i) sketching each new input signal entails a nontrivial computational cost, which can be a bottleneck in streaming applications; ii) the design of the random projection matrix does not take into account any a priori information on the input; and iii) the guarantees offered are probabilistic. Alternatively one can think of reducing complexity by simply retaining a few samples of each input. Different from random sketching, under stationarity the sampling can remain fixed and it incurs negligible online complexity. Sampling schemes typically build on a parsimonious model for the signals of interest, which in the case of graph signals is either smoothness or bandlimitedness – i.e., a sparse representation in the graph Fourier domain [7], [8]. Most existing sampling methods are designed with the objective of reconstructing the original graph signal, and do not account for subsequent processing the signal may undergo; see, e.g., [9] for a recent exception.

In this sketching context and with reference to Fig. 1, we propose a novel sampling scheme for signals $\mathbf{x}$ that are bandlimited on a graph (Section II), that also leverages the transformation $\mathbf{H}$ whose output we wish to approximate. In Section III we formulate problems to jointly optimize the sample selection matrix $\mathbf{C}$ and a sketch $\mathbf{H}_s$ of $\mathbf{H}$, so when $\mathbf{H}_s$ is applied to the sampled input signal $\mathbf{Cx}$ the result is close to the desired output $\mathbf{y}$. The pragmatic setting where the input signal is corrupted by noise $\mathbf{w}$ is also investigated in Section III-B. The general premise is to shift all the computational burden of designing $\mathbf{H}_s$ and $\mathbf{C}$ to an off-line phase (see Section III-C for low-complexity heuristics), so that the online stage only entails acquiring the specific samples and processing them via $\mathbf{H}_s$. Numerical tests in Section IV corroborate the

Work in this paper is supported by the NSF award CCF-1217963 and Spanish MINECO TEC2013-41604-R. [†]Dept. of Electrical and Systems Eng., Univ. of Pennsylvania. [‡]Dept. of Signal Theory and Comms., King Juan Carlos Univ. [*]Dept. of Electrical and Comp. Eng., Univ. of Rochester. Emails: {fgama,aribeiro}@seas.upenn.edu, antonio.garcia.marques@urjc.es, gmateosb@ece.rochester.edu

978-1-5386-3954-2/16/$31.00 ©2016 IEEE

effectiveness of the novel methods when adopted to classify handwritten digits from the MNIST database [10], using as few as 20 input pixels. Conclusions are drawn in Section V.

II. Preliminaries

Let $N = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ be a network (graph) described by a set of n nodes $\mathcal{V}$, a set $\mathcal{E}$ of edges (i, j) and a weight function $\mathcal{W} : \mathcal{E} \rightarrow \mathbb{R}$ that gives weights to the directed edges. A graph signal $\mathbf{x} \in \mathbb{R}^n$ is defined on the nodes of such a network where each element of the vector $\mathbf{x} = [x_1, \ldots, x_n]^T$ represents a real value present at the node [1], [2]. A graph-shift operator $\mathbf{S} \in \mathbb{R}^{n \times n}$ is introduced in order to describe the impact of the structure of the network on the signal [11]. Matrix $\mathbf{S}$ is such that $[\mathbf{S}]_{i,j} = 0$ whenever $i \neq j$ and $(j, i) \notin \mathcal{E}$. We will focus on normal shifts, so that $\mathbf{S}$ can be decomposed as

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H := \mathbf{V}\text{diag}\left([\lambda_1, \ldots, \lambda_n]^T\right)\mathbf{V}^H. \qquad (1)$$

where the unitary matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{C}^{n \times n}$ contains the eigenvectors of $\mathbf{S}$ and the diagonal matrix $\mathbf{\Lambda} = \text{diag}\left([\lambda_1, \ldots, \lambda_n]^T\right) \in \mathbb{C}^{n \times n}$ the corresponding eigenvalues. Examples of commonly used shift operators that are normal include the Laplacian or the adjacency matrix of symmetric graphs [4], [12], the adjacency of a directed cycle for time signals, and the correlation or precision (inverse covariance) matrix of processes following a graphical model [13].

Bandlimited, stationary graph signals. The eigendecomposition in (1) can be used to define the Graph Fourier Transform (GFT) and the inverse GFT (iGFT) as $\tilde{\mathbf{x}} := \mathbf{V}^H\mathbf{x}$ and $\mathbf{x} = \mathbf{V}\tilde{\mathbf{x}}$, respectively [4]. Vector $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_n]^T$ contains the frequency coefficients of $\mathbf{x}$. A key assumption made throughout this paper is that $\mathbf{x}$ is k-bandlimited on $\mathbf{S}$. This implies that there exists a constant $k \ll n$ such that $\tilde{x}_l = 0$ for all $l > k$. Then, the GFT vector $\tilde{\mathbf{x}}$ can be rewritten as $\tilde{\mathbf{x}} = [\tilde{\mathbf{x}}_k; \mathbf{0}_{n-k}]$, where $\tilde{\mathbf{x}}_k := [\tilde{x}_1, \ldots, \tilde{x}_k]^T$ contains the first k elements of $\tilde{\mathbf{x}}$. Collecting the k eigenvectors associated with the active frequencies in the matrix $\mathbf{V}_k := [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in \mathbb{C}^{n \times k}$, the GFT-iGFT pairs can be rewritten as

$$\tilde{\mathbf{x}}_k = \mathbf{V}_k^H\mathbf{x}, \qquad \mathbf{x} = \mathbf{V}_k\tilde{\mathbf{x}}_k. \qquad (2)$$

Assuming also $\mathbf{x}$ is a realization of a zero-mean random process, we specify next its covariance matrix $\mathbf{R}_x := \mathbb{E}[\mathbf{x}\mathbf{x}^T] \in \mathbb{R}^{n \times n}$. Since the realizations are bandlimited, we consider first the so-called frequency template $\mathbf{T} \in \mathbb{C}^{k \times k}$ defined as

$$\mathbf{T} \ := \ \mathbb{E}\left[\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k^H\right]. \qquad (3)$$

This results in a singular positive semidefinite covariance matrix $\mathbf{R}_x = \mathbb{E}[\mathbf{x}\mathbf{x}^H] = \mathbf{V}_k\mathbb{E}[\tilde{\mathbf{x}}_k\tilde{\mathbf{x}}_k^H]\mathbf{V}_k^H = \mathbf{V}_k\mathbf{T}\mathbf{V}_k^H$ with rank k. If no information is available other than $\mathbf{x}$ being bandlimited, a reasonable choice is to set $\mathbf{T} = \mathbf{I}$ so that $\mathbf{R}_x = \mathbf{V}_k\mathbf{V}_k^H$. This type of spectral templates appear when $\mathbf{x}$ is *graph stationary with respect to* the shift operator $\mathbf{S}$ [13].

Sampling and reconstructing bandlimited signals. If the signal $\mathbf{x}$ is k-bandlimited on $\mathbf{S}$ and the active frequencies are known, then the k values in $\tilde{\mathbf{x}}_k$ suffice to describe the n values in $\mathbf{x}$ [cf. (2)]. Hence, it is conceivable that $\mathbf{x}$ could be recovered from $p \geq k$ properly selected samples. In this context, different sampling schemes for graph signals have been proposed [7], [8]. The most relevant for the problem at hand is the so-called selection sampling [7], where samples correspond to values of the signal $\mathbf{x}$ at a particular set of $p \geq k$ nodes. Specifically, the goal is to design a selection matrix $\mathbf{C} \in \mathcal{C}_{pn}$, where

$$\mathcal{C}_{pn} := \{\mathbf{C} \in \mathbb{R}^{p \times n} : \ \mathbf{C}\mathbf{1}_n = \mathbf{1}_n, \ \mathbf{C}^T\mathbf{1}_n \preceq \mathbf{1}_n, \ C_{ij} \in \{0, 1\}\},$$

so that samples $\mathbf{x}_s = \mathbf{C}\mathbf{x}$ contain enough information to accurately recover $\mathbf{x}$ via suitable interpolation. If samples of $\mathbf{x}$ can be acquired perfectly in the absence of noise, then any $\mathbf{C}$ for which $\mathbf{C}\mathbf{V}_k$ is nonsingular yields perfect recovery since

$$\mathbf{x} = \mathbf{V}_k(\mathbf{C}\mathbf{V}_k)^{-1}\mathbf{x}_s. \qquad (4)$$

When samples are noisy, then there are several methods and algorithms for obtaining the optimal selection matrix $\mathbf{C}$; see [7], [8]. Working with $\mathbf{x}_s$ in lieu of $\mathbf{x}$ offers several advantages. Of particular interest here are the computational savings of processing $\mathbf{x}_s$ rather than $\mathbf{x}$, especially when $p \ll n$.

III. Sketching as Graph Signal Sampling

We consider here linear sketching problems for signals that are bandlimited on a graph. With reference to Fig. 1, consider a graph signal $\mathbf{x} \in \mathbb{R}^n$ corrupted by additive noise $\mathbf{w} \in \mathbb{R}^n$, and suppose that a stream of inputs $\mathbf{z} := \mathbf{x} + \mathbf{w}$ is available. Ideally one would like to process each noiseless input $\mathbf{x}$ by a given linear transformation $\mathbf{H} \in \mathbb{R}^{m \times n}$ to generate the output $\mathbf{y} \in \mathbb{R}^m$, where $n \geq m$ and both dimensions are large. Our goal is to find a fixed sampling scheme $\mathbf{C} \in \mathcal{C}_{pn}$ and a fixed sketch $\mathbf{H}_s \in \mathbb{R}^{m \times p}$ of the linear transformation $\mathbf{H}$, such that with $p \leq n$ the signal $\hat{\mathbf{y}} := \mathbf{H}_s\mathbf{z}_s = \mathbf{H}_s\mathbf{C}(\mathbf{x} + \mathbf{w})$ resembles the desired output $\mathbf{y} = \mathbf{H}\mathbf{x}$. The design is performed off-line, assuming that: i) the linear transformation $\mathbf{H}$ is known; ii) the inputs correspond to realizations of a k-bandlimited graph signal whose frequency template $\mathbf{T}$ (hence its covariance $\mathbf{R}_x$) is known; and iii) the noise $\mathbf{w}$ is zero mean, uncorrelated with respect to the input $\mathbf{x}$ and with known positive-definite covariance $\mathbf{\Sigma}_w = \mathbb{E}[\mathbf{w}\mathbf{w}^T] \succ \mathbf{0}$. The design of $\mathbf{C}$ and $\mathbf{H}_s$ is performed jointly as the solution of the following minimization

$$\{\mathbf{C}^*, \mathbf{H}_s^*\} := \underset{\mathbf{C} \in \mathcal{C}_{pn}, \mathbf{H}_s}{\text{argmin}} \ \mathbb{E}\left[\left\|\mathbf{H}_s\mathbf{C}(\mathbf{x} + \mathbf{w}) - \mathbf{H}\mathbf{x}\right\|_2^2\right]. \qquad (5)$$

We consider a streaming setup where matrix $\mathbf{H}$ has to be applied to a succession of inputs signals. Since (5) minimizes an ensemble criterion, finding the optimal solution only requires knowledge of second-order statistics of $\mathbf{x}$. Hence, (5) can be solved off-line, yielding an optimal sketch $\mathbf{H}_s^*$ and selection matrix $\mathbf{C}^*$ that will be the same for all the inputs under the stationarity assumption. As a result, during the online phase one must calculate $\mathbf{H}_s^*\mathbf{C}^*(\mathbf{x} + \mathbf{w})$ instead of $\mathbf{H}(\mathbf{x} + \mathbf{w})$, reducing the long-run computational cost by a factor of p/n.

We first look at the optimal joint design when noise is not present (Section III-A) and then address its noisy counterpart (Section III-B). A number of alternatives to solve approximately the resultant optimizations are outlined in Section III-C.

A. Noiseless case

First, consider the case where $\mathbf{w} = \mathbf{0}$ in Fig. 1. In this noiseless scenario, the desired output is $\mathbf{y} = \mathbf{H}\mathbf{x}$ and the reduced-complexity approximation is given by $\hat{\mathbf{y}} = \mathbf{H}_s\mathbf{C}\mathbf{x}$.

Proposition 1 *Let* $\mathbf{x} \in \mathbb{R}^n$ *be a* k*-bandlimited signal with known spectral template* $\mathbf{T} \in \mathbb{C}^{k\times k}$ *and let* $\mathbf{H} \in \mathbb{R}^{m\times n}$ *be a linear transformation. Let* $\mathbf{H}_s \in \mathbb{R}^{m\times p}$ *be a reduced-input dimensionality sketch of* $\mathbf{H}$*,* $p \leq n$ *and* $\mathbf{C} \in \mathcal{C}_{pn}$ *be a selection matrix. If* $p = k$ *and* $\mathbf{C}^*$ *is designed such that* $\text{rank}\{\mathbf{C}^*\mathbf{V}_k\} = p = k$*, then* $\hat{\mathbf{y}} = \mathbf{H}_s^*\mathbf{C}^*\mathbf{x} = \hat{\mathbf{y}}$ *is equal to the desired output* $\mathbf{y} = \mathbf{H}\mathbf{x}$ *provided that the sketch* $\mathbf{H}_s^*$ *is found as*

$$\mathbf{H}_s^* = \mathbf{H}\mathbf{V}_k(\mathbf{C}^*\mathbf{V}_k)^{-1}. \tag{6}$$

Proof: The mean-squared error (MSE) criterion [cf. (5)] is

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2\right] = \mathbb{E}\left[\|\mathbf{H}\mathbf{x}-\mathbf{H}_s\mathbf{C}\mathbf{x}\|_2^2\right] \\ &\quad= \text{tr}\big[\mathbf{H}\mathbf{R}_x\mathbf{H}^T - 2\mathbf{H}_s\mathbf{C}\mathbf{R}_x\mathbf{H}^T + \mathbf{H}_s\mathbf{C}\mathbf{R}_x\mathbf{C}^T\mathbf{H}_s^T\big]. \end{aligned} \tag{7}$$

Recall that $\text{rank}\{\mathbf{C}\} = p$ for any $\mathbf{C} \in \mathcal{C}_{pn}$. Optimizing the MSE cost over $\mathbf{H}_s$ first, results in (recall $\mathbf{R}_x = \mathbf{V}_k\mathbf{T}\mathbf{V}_k^H$)

$$\mathbf{H}\mathbf{V}_k\mathbf{T}\mathbf{V}_k^H\mathbf{C}^T = \mathbf{H}_s^*\mathbf{C}\mathbf{V}_k\mathbf{T}\mathbf{V}_k^H\mathbf{C}^T. \tag{8}$$

Now, if we set $p = k$ and choose $\mathbf{C}$ such that $\text{rank}\{\mathbf{C}\mathbf{V}_k\} = p = k$, then $(\mathbf{V}_k^H\mathbf{C}^T\mathbf{C}\mathbf{V}_k)^{-1}$ exists. Thus, by post multiplying both sides of (8) by the nonsingular $p \times p$ matrix $\mathbf{C}\mathbf{V}_k(\mathbf{V}_k^H\mathbf{C}^T\mathbf{C}\mathbf{V}_k)^{-1}\mathbf{T}^{-1}$, one arrives at

$$\mathbf{H}\mathbf{V}_k = \mathbf{H}_s^*\mathbf{C}\mathbf{V}_k.$$

Finally, because $\text{rank}\{\mathbf{C}\mathbf{V}_k\} = p = k$, then $(\mathbf{C}\mathbf{V}_k)^{-1}$ exists, so we obtain the closed-form solution for $\mathbf{H}_s^*$ given by (6). ■

All in all, in the absence of noise it suffices to first set $p = k$ to find a selection matrix $\mathbf{C} \in \mathcal{C}_{pn}$ such that $\text{rank}\{\mathbf{C}\mathbf{V}_k\} = p = k$, and then obtain $\mathbf{H}_s$ as per Proposition 1. This ensures that $\mathbf{y}$ can be formed error-free using p samples of $\mathbf{x}$ via $\mathbf{y} = \hat{\mathbf{y}} = \mathbf{H}_s\mathbf{C}\mathbf{x}$. Clearly, selecting $p \geq k$ will also do the job, provided that $\text{rank}\{\mathbf{C}\mathbf{V}_k\} \geq k$. We close by noting that $\text{rank}\{\mathbf{C}\mathbf{V}_k\} = k$ is the same condition for exact recovery with selection sampling [7]. This is expected, since in the noiseless case here the design of $\mathbf{C}$ decouples from that of $\mathbf{H}_s$. As a result, existing methods to determine the most informative nodes in sampling scenarios are also applicable here [9].

B. Noisy case

When noise is present, the noise model must be accounted for in the minimization in (5). To that end, observe first that if $\mathbf{C} \in \mathcal{C}_{pn}$ is a $p \times n$ selection matrix, then $\mathbf{C}\mathbf{C}^T = \mathbf{I}_p$ is the identity matrix of size $p\times p$. Moreover, $\mathbf{C}^T\mathbf{C} = \text{diag}(\mathbf{c})$ where $\mathbf{c} \in \{0,1\}^n$ is a sampling vector containing p ones, located in the places corresponding to the nodes to be sampled. Now, define the covariance matrix of the output signal $\mathbf{y}$ as $\mathbf{R}_y := \mathbb{E}[\mathbf{y}\mathbf{y}^T] = \mathbf{H}\mathbf{R}_x\mathbf{H}^T \in \mathbb{R}^{m\times m}$. Also, define $\bar{\mathbf{C}}_\alpha := \text{diag}(\mathbf{c})/\alpha$ as the rescaled sampling vector in matrix form, where $\alpha > 0$ is the rescaling parameter. Finally, define the auxiliary matrix $\bar{\mathbf{\Sigma}}_\alpha := \mathbf{R}_x + \mathbf{\Sigma}_w - \alpha\mathbf{I}_n$.

With these definitions, the following proposition asserts that (5) is equivalent to a mixed-binary optimization problem, with linear objective and linear matrix inequality (LMI) constraints.

Proposition 2 *The solution of* (5) *is given by* $\mathbf{C}^*$ *and* $\mathbf{H}_s^* = \mathbf{H}_s^*(\mathbf{C}^*)$*, where*

$$\mathbf{H}_s^*(\mathbf{C}) = \mathbf{H}\mathbf{R}_x\mathbf{C}^T\left(\mathbf{C}(\mathbf{R}_x+\mathbf{\Sigma}_w)\mathbf{C}^T\right)^{-1} \tag{9}$$

and $\mathbf{C}^*$ *can be obtained as the solution to the problem*

$$\min_{\mathbf{C}\in\mathcal{C}_{pn}} \ \text{tr}\left[\mathbf{R}_y - \mathbf{H}\mathbf{R}_x\mathbf{C}^T\left(\mathbf{C}(\mathbf{R}_x+\mathbf{\Sigma}_w)\mathbf{C}^T\right)^{-1}\mathbf{C}\mathbf{R}_x\mathbf{H}^T\right]. \tag{10}$$

Moreover, (10) *is equivalent to*

$$\begin{aligned} &\min_{\mathbf{c}\in\{0,1,\}^n,\mathbf{Y}} \ \text{tr}\left[\mathbf{Y}\right] \\ &\quad\text{s.\,t. } \bar{\mathbf{C}}_\alpha = \text{diag}(\mathbf{c})/\alpha\ ,\ \mathbf{c}^T\mathbf{1}_n = p \\ &\qquad\begin{bmatrix} \mathbf{Y}-\mathbf{R}_y+\mathbf{H}\mathbf{R}_x\bar{\mathbf{C}}_\alpha\mathbf{R}_x\mathbf{H}^T & \mathbf{H}\mathbf{R}_x\bar{\mathbf{C}}_\alpha \\ \bar{\mathbf{C}}_\alpha\mathbf{R}_x\mathbf{H}^T & \bar{\mathbf{\Sigma}}_\alpha^{-1}+\bar{\mathbf{C}}_\alpha \end{bmatrix} \succeq \mathbf{0} \end{aligned} \tag{11}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times m}$ *is an auxiliary variable and* $\alpha > 0$ *is any scalar satisfying* $\bar{\mathbf{\Sigma}}_\alpha = (\mathbf{R}_x + \mathbf{\Sigma}_w - \alpha\mathbf{I}_n) \succ \mathbf{0}$.

Proof: The objective function in (5) can be rewritten as

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{y}-\hat{\mathbf{y}}\|_2^2\right] = \mathbb{E}\left[\|\mathbf{H}\mathbf{x}-\mathbf{H}_s\mathbf{C}(\mathbf{x}+\mathbf{w})\|_2^2\right] \\ &\quad= \text{tr}\left[\mathbf{H}\mathbf{R}_x\mathbf{H}^T - 2\mathbf{C}\mathbf{R}_x\mathbf{H}^T\mathbf{H}_s + \mathbf{H}_s\mathbf{C}(\mathbf{R}_x+\mathbf{\Sigma}_w)\mathbf{C}^T\mathbf{H}_s^T\right] \end{aligned} \tag{12}$$

since $\mathbf{x}$ and $\mathbf{w}$ are assumed independent. Solving for $\mathbf{H}_s$ yields

$$\mathbf{H}_s^*(\mathbf{C}) = \mathbf{H}\mathbf{R}_x\mathbf{C}^T\left(\mathbf{C}(\mathbf{R}_x+\mathbf{\Sigma}_w)\mathbf{C}^T\right)^{-1}$$

establishing (9). Matrix $\mathbf{C} \in \mathcal{C}_{pn}$ is full rank since it selects p *distinct* nodes, then $\mathbf{C}(\mathbf{R}_x + \mathbf{\Sigma}_w)\mathbf{C}^T$ has rank p and thus it is invertible [14]. Substituting the expression for $\mathbf{H}_s^*(\mathbf{C})$ into (12), yields (10). The inverse in the objective of (10) can be written as [9]

$$\begin{aligned} &(\mathbf{C}(\mathbf{R}_x+\mathbf{\Sigma}_w)\mathbf{C}^T)^{-1} = \left(\alpha\mathbf{I}_p - \alpha\mathbf{I}_p + \mathbf{C}(\mathbf{R}_x+\mathbf{\Sigma}_w)\mathbf{C}^T\right)^{-1} \\ &\quad= \alpha^{-1}\mathbf{I}_p - \alpha^{-2}\mathbf{C}(\bar{\mathbf{\Sigma}}_\alpha^{-1} + \alpha^{-1}\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T, \end{aligned} \tag{13}$$

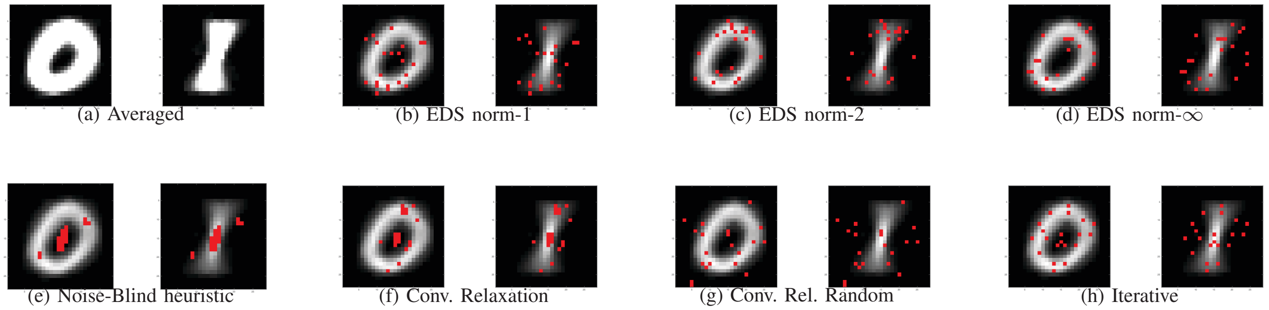
where $\alpha \neq 0$ is a rescaling parameter, and we used the Woodbury Matrix Identity. Note that α has to be such that $\bar{\mathbf{\Sigma}}_\alpha = (\mathbf{R}_x + \mathbf{\Sigma}_w - \alpha\mathbf{I}_n)$ is still invertible. Substituting (13) into (10) and recalling that $\mathbf{C}^T\mathbf{C} = \text{diag}(\mathbf{c})$, we have that

$$\begin{aligned} \min_{\mathbf{c}\in\{0,1\}^n,\bar{\mathbf{C}}_\alpha} \ &\text{tr}\Big[\mathbf{R}_y - \mathbf{H}\mathbf{R}_x\bar{\mathbf{C}}_\alpha\mathbf{R}_x\mathbf{H}^T \\ &\qquad + \mathbf{H}\mathbf{R}_x\bar{\mathbf{C}}_\alpha\left(\bar{\mathbf{\Sigma}}_\alpha^{-1}+\bar{\mathbf{C}}_\alpha\right)^{-1}\bar{\mathbf{C}}_\alpha\mathbf{R}_x\mathbf{H}^T\Big] \\ &\text{s.\,t. } \bar{\mathbf{C}}_\alpha = \text{diag}(\mathbf{c})/\alpha\ ,\ \mathbf{c}^T\mathbf{1}_n = p. \end{aligned} \tag{14}$$

Note that in (14) we optimize over a binary vector $\mathbf{c} \in \mathbb{R}^n$ with exactly p nonzero entries, instead of a binary matrix $\mathbf{C} \in \mathcal{C}_{pn}$. The p nonzero elements in $\mathbf{c}$ indicate the nodes to be sampled. Problem (14) can be reformulated as

$$\begin{aligned} &\min_{\mathbf{c}\in\{0,1\}^n,\mathbf{Y}} \ \text{tr}\left[\mathbf{Y}\right] \\ &\quad\text{s.\,t. } \mathbf{R}_y - \mathbf{H}\mathbf{R}_x\bar{\mathbf{C}}_\alpha\mathbf{R}_x\mathbf{H}^T \\ &\qquad\quad + \mathbf{H}\mathbf{R}_x\bar{\mathbf{C}}_\alpha\left(\bar{\mathbf{\Sigma}}_\alpha^{-1}+\bar{\mathbf{C}}_\alpha\right)^{-1}\bar{\mathbf{C}}_\alpha\mathbf{R}_x\mathbf{H}^T \preceq \mathbf{Y} \\ &\qquad\quad \bar{\mathbf{C}}_\alpha = \text{diag}(\mathbf{c})/\alpha\ ,\ \mathbf{c}^T\mathbf{1}_n = p \end{aligned} \tag{15}$$

where $\mathbf{Y} \in \mathbb{R}^{m\times m}$, $\mathbf{Y} \succeq 0$ is an auxiliary optimization variable. Using the Schur-complement lemma for *positive definite* matrices, problem (15) can be written as (11). Hence, to complete the proof we need to show that $\bar{\mathbf{\Sigma}}_\alpha^{-1} + \bar{\mathbf{C}}_\alpha \succeq \mathbf{0}$ so that the aforementioned lemma can be invoked. To that end, suppose first that $\alpha < 0$. Then we have that $\bar{\mathbf{\Sigma}}_\alpha \succ \mathbf{0}$ and



(a) Averaged (b) EDS norm-1 (c) EDS norm-2 (d) EDS norm-∞

(e) Noise-Blind heuristic (f) Conv. Relaxation (g) Conv. Rel. Random (h) Iterative

Fig. 2. (a) Averaged image of all digits in the test set. (b)–(h) Selected pixels to use for classification of the digits according to each strategy. It is observed how the methods for reconstruction (b)–(d) tend to select pixels around the annulus that determine the digit 0, especially on top and bottom, which also help in reconstructing the digit 1. On the other hand, methods for classification (e)–(f) tend to distribute the pixel selection both in the center and in the annulus on the sides, which are the pixels that best help distinguish a zero (no pixels in the center) and a one (no pixels on the sides of the annulus).

$\bar{\mathbf{C}}_\alpha = \alpha^{-1}\text{diag}(\mathbf{c}) \preceq \mathbf{0}$, so that $\bar{\mathbf{\Sigma}}_\alpha^{-1} + \bar{\mathbf{C}}_\alpha \succeq \mathbf{0}$ may not be positive definite. Suppose now that $\alpha > 0$. Then $\bar{\mathbf{C}}_\alpha \succeq \mathbf{0}$ and there always exists a sufficiently small positive α such that $\bar{\mathbf{\Sigma}}_\alpha \succ \mathbf{0}$ since $\mathbf{\Sigma}_w \succ \mathbf{0}$. This implies that if α is chosen such that $\alpha > 0$ and $\bar{\mathbf{\Sigma}}_\alpha = \mathbf{R}_x + \mathbf{\Sigma}_w - \alpha\mathbf{I}_n \succ \mathbf{0}$ (which are the conditions stated in the proposition), then $\bar{\mathbf{\Sigma}}_\alpha^{-1} + \bar{\mathbf{C}}_\alpha$ is positive definite and problems (14) and (11) are equivalent. ■

In words, the problem in (5) can be solved in two steps. First the optimal sketch $\mathbf{H}_s^*$ is found as a function of $\mathbf{C}$ via (9), and then this $\mathbf{H}_s^*(\mathbf{C})$ is substituted into (5) to formulate the optimization in (11), which depends only on $\mathbf{C}$. From an algorithmic perspective, the order is reversed. First, we find $\mathbf{C}$ by "solving" the binary optimization in either (10) or (11), and then the resulting $\mathbf{C}^*$ is substituted into (9) to find $\mathbf{H}_s^*$ in $O(nmp)+O(p^3)$ complexity. Different heuristics to tackle the non-convex binary optimization over $\mathbf{C}$ are discussed next.

C. Heuristic Approaches

A natural first alternative is to relax the binary constraint

$$\mathbf{c} \in \{0,1\}^n \underset{\text{Relaxation}}{\Rightarrow} \mathbf{0}_n \preceq \mathbf{c} \preceq \mathbf{1}_n \tag{16}$$

so that problem (11) becomes convex, and is thus efficiently solved. Once a solution to the relaxation is obtained, two ways of recovering a binary vector $\mathbf{c}$ are considered. The first one is a deterministic method referred to as thresholding, which simply consists of setting the largest p elements to 1 and the rest to 0. The second one consists on normalizing the solution so that it sums to 1, which can be viewed as a probability distribution over the nodes. The sampled nodes are then drawn at random from this distribution, see [15]. Although not pursued here, also pertinent are formulations that penalize the objective with $\|\mathbf{c}\|_1$ and leverage ℓ_1-norm minimization advances to promote sparsity on $\mathbf{c}$.

A second method is to ignore the noise altogether, and select node so that $\mathbf{H}\mathbf{R}_x^{1/2}\text{diag}(\mathbf{c})\mathbf{R}_x^{1/2}\mathbf{H}^T$ is as close as possible to $\mathbf{H}\mathbf{R}_x\mathbf{H}^T = \mathbf{R}_y$ [cf. (10)]. This implies that the samples chosen correspond to the rows of $\mathbf{R}_x^{1/2}\mathbf{H}^T$ with highest ℓ_2 norm, a scheme referred to as the noise-blind heuristic.

The last approach relies on a greedy iterative scheme. Specifically, instead of searching over all possible $\binom{n}{p}$ sampling configurations, we run p iterations and in each of them we try nodes individually and retain the one yielding the best solution. This way only $n(n-1)\cdots(n-(p-1)) < n^p$ evaluations of the objective function (10) are required. Greedy algorithms have well-documented merits for sample selection, even for non-submodular objectives like the one in (10).

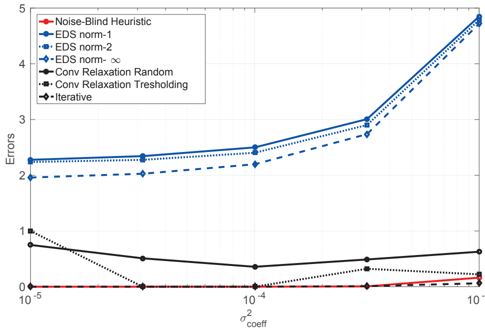


Fig. 3. Average number of classification errors (out of 200 images to classify) as a function of the noise coefficient σ^2_{coeff} that determines the noise power. For each noise coefficient, 500 realizations were carried out. It is observed that the proposed methods in this work yield better performance than traditional sampling methods.

IV. Numerical experiments

The problem of classifying handwritten digits from the MNIST database [10] is considered here to validate the methods proposed. This particular classification task is typically carried out in two steps: first the images are transformed to the PCA domain and then a linear classifier on the principal components is implemented [16]. However, the first step can often be computationally expensive. Since the PCA transformation is linear, our approach is to subsume both the linear classifier and the PCA transform into a single linear operator $\mathbf{H}$, which is then approximated by sampling pixels directly – hence, boosting the online classification speed.

We will constrain ourselves to the classification of the black and white images of digits 0 and 1. The images have size 28×28 and by taking 5000 images of each digit from the training set and vectorizing them, estimates of the mean and the covariance matrix of the images are obtained. These estimates are used to compute the PCA transform. Images are assumed k-bandlimited with $k = 20$, since this is the number of principal components needed to achieve perfect classification in the PCA domain. The total number of pixels is $n = 784$. The linear classifier is an SVM applied directly to the PCA domain, thus $m = 1$. The total number of images to be classified are 100 for each digit. The result of *averaging* all the 0s and 1s considered is shown in Fig. 2(a). The methods proposed in this work are compared to three experimental design sampling methods (EDS) proposed in [17]. Specifically, we implement sampling with replacement with weights according to the: 1) ℓ_∞ norm of the rows of $\mathbf{V}$; 2) ℓ_1 norm of the rows of $\mathbf{V}$; and 3) ℓ_2 norm of the rows of $\mathbf{V}$. This latter case is the method proposed in [15].

For the first simulation, we consider that $p = k = 20$ pixels are sampled and add a fixed noise with covariance $\mathbf{\Sigma}_w = \sigma^2 \mathbf{I}_n$, where $\sigma^2 = \sigma^2_{\text{coeff}} \cdot \|\hat{\boldsymbol{\mu}}\|^2$ with $\sigma^2_{\text{coeff}} = 10^{-4}$, and $\hat{\boldsymbol{\mu}}$ is the estimated mean. Selected pixels are illustrated in Fig. 2(b)–(h). The estimated error rate is the average error across 500 realizations. It yields an error rate of 0% for the SVM classifier using the full image. For the EDS reconstruction techniques with norm-1, norm-2 and norm-∞ weights the error rates are 0%, 4.81% and 1.2%, respectively. For the noise-blind heuristic there is an error rate of 0.5% and for the convex relaxation thresholding technique, the convex relaxation random technique and the iterative heuristic, the error is 0%, 0.3% and 0%, respectively. All in all, this first example gives rise to the following findings: f1) the classification performance using only $p = 20$ pixels is very close to that using the full image ($n = 784$ pixels) but with only a 2.55% of the online computational cost; and f2) the proposed schemes (especially the most sophisticated ones) tend to work better than existing alternatives. To confirm these findings we run two additional set of experiments.

The second simulation fixes the number of samples $p = k = 20$ and considers different noise values σ^2_{coeff}. Again, 500 realizations are carried out for each value of σ^2_{coeff}. The number of errors (out of 200 images) averaged across the realizations is plotted in Fig. 3. It confirms that the methods proposed in this work yield a reconstruction performance better than that of traditional methods [cf. f2)]. Finally, the third simulation fixes the noise to $\sigma^2_{\text{coeff}} = 10^{-4}$ and varies the number of samples p from 16 to 35. The results, averaged across 500 noise realizations, are shown in Fig. 4. In general, the proposed methods work better than traditional sampling methods [cf. f2)], exhibiting an error rate below 0.5% for all p.

V. Conclusions

We studied a class of linear sketching problems for streaming signals that are bandlimited on a graph. To effect computational savings during online operation, we formulated an optimization problem to jointly design the sampling scheme to reduce the dimensionality of the input, as well as a sketch of the desired transformation to affordably process the resulting samples. Since the resultant problem was non-convex, different suboptimal schemes were proposed and their performance compared in the context of handwritten digit classification.

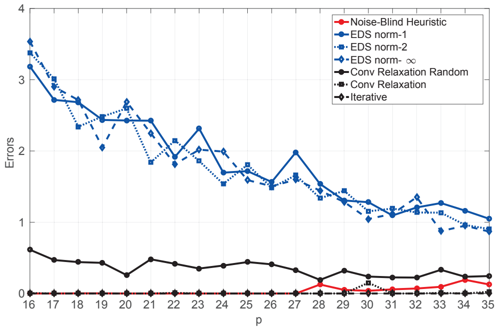


Fig. 4. Number of classification errors (out of 200 images to classify) as a function of the number of samples p. It is observed that the proposed methods in this work yield better performance than traditional sampling methods.

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, 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," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, September 2014.

[3] S. K. Narang and A. Ortega, "Perfect Reconstruction Two-Channel Wavelet FIlter Banks for Graph Structured Data," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2786–2799, June 2012.

[4] A. Sandyhaila and J. M. F. Moura, "Discrete Signal Processing on Graphs: Frequency Analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, June 2014.

[5] D. K. Berberidis, V. Kekatos, and G. B. Giannakis, "Online censoring for large-scale regressions with application to streaming big data," *arXiv preprint arXiv:1507.07536v1*, 2016.

[6] D. P. Woodruff, "Sketching as a Tool for Numerical Linear Algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1-2, pp. 1–157, 2014.

[7] S. Chen, R. Varma, A. Sandryhaila, and J. Kovačević, "Discrete Signal Processing on Graphs: Sampling Theory," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6510–6523, December 2015.

[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, no. 7, pp. 1832–1843, April 2016.

[9] S. Liu, S. P. Chepuri, M. Fardad, E. Masazade, G. Leus, and P. K. Varshney, "Sensor Selection for Estimation with Correlated Measurement Noise," *IEEE Trans. Signal Process.*, vol. PP, no. 99, pp. 1832–1843, April 2016.

[10] Y. Le Cun, C. Cortes, and C. J. C. Burges, "The MNIST Database of handwritten digits," Website, http://yann.lecun.com/exdb/mnist/, 2015-08-18.

[11] M. Püschel and J. M. F. Moura, "Algebraic Signal Processing Theory: Foundation and 1-D Time," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3575–3585, August 2008.

[12] A. Sandryhaila and J. M. F. Moura, "Discrete Signal Processing on Graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, April 2013.

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

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 1st ed. Cambridge, UK: Cambridge University Press, 1985.

[15] G. Puy, N. Tremblay, R. Gribonval, and P. Vandergheynst, "Random Sampling of Bandlimited Graph Signals," *eprint arXiv:1511.05118*, November 2015.

[16] J. E. Jackson, *A User's Guide to Principal Components*. New York, NY: John Wiley and Sons, 1991.

[17] R. Varma, S. Chen, and J. Kovačević, "Spectrum-Blind Signal Recovery on Graphs," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. Cancún, México: IEEE, 13–16 December 2015, pp. 81–84.