# Distributed Adaptive Learning of Graph Signals

Paolo Di Lorenzo, *Member, IEEE*, Paolo Banelli, *Member, IEEE*, Sergio Barbarossa, *Fellow, IEEE*, and Stefania Sardellitti, *Member, IEEE*

***Abstract*—The aim of this paper is to propose distributed strategies for adaptive learning of signals defined over graphs. Assuming the graph signal to be bandlimited, the method enables distributed reconstruction, with guaranteed performance in terms of mean-square error, and tracking from a limited number of sampled observations taken from a subset of vertices. A detailed mean-square analysis is carried out and illustrates the role played by the sampling strategy on the performance of the proposed method. Finally, some useful strategies for distributed selection of the sampling set are provided. Several numerical results validate our theoretical findings, and illustrate the performance of the proposed method for distributed adaptive learning of signals defined over graphs.**

***Index Terms*—Graph signal processing, sampling on graphs, adaptation and learning over networks, distributed estimation.**

## I. Introduction

OVER the last few years, there was a surge of interest in the development of processing tools for the analysis of signals defined over a graph, or graph signals for short, in view of the many potential applications spanning from sensor networks, social media, vehicular networks, big data or biological networks [1]–[3]. Graph signal processing (GSP) considers signals defined over a discrete domain having a very general structure, represented by a graph, and subsumes classical discrete-time signal processing as a very simple case. Several processing methods for signals defined over a graph were proposed in [2], [4]–[6], and one of the most interesting aspects is that these analysis tools come to depend on the graph topology. A fundamental role in GSP is of course played by spectral analysis, which passes through the definition of the Graph Fourier Transform (GFT). Two main approaches for GFT have been proposed in the literature, based on the projection of the signal onto the eigenvectors of either the graph Laplacian, see, e.g., [1], [7], [8], or of the adjacency matrix, see, e.g. [2], [9]. The first approach is more suited to handle *undirected* graphs and builds on the clustering properties of the graph Laplacian eigenvectors and the minimization of the $\ell_2$ norm graph total variation; the second approach applies also to *directed* graphs and builds on the interpretation of the adjacency operator as a graph shift operator, which paves the way for all linear shift-invariant filtering methods for graph signals [10], [11].

One of the basic and interesting problems in GSP is the development of a *sampling theory* for signals defined over graphs, whose aim is to recover a bandlimited (or approximately bandlimited) graph signal from a subset of its samples. A seminal contribution was given in [7], later extended in [12] and, very recently, in [9], [13], [14], [15], [16]. Several reconstruction methods have been proposed, either iterative as in [14], [17], or single shot, as in [9], [13], [18]. Frame-based approaches for the reconstruction of graph signals from subsets of samples have also been proposed in [7], [13], [14]. Furthermore, as shown in [9], [13], dealing with graph signals, the recovery problem may easily become ill-conditioned, depending on the location of the samples. Thus, for any given number of samples, the sampling set plays a fundamental role in the conditioning of the recovery problem. This makes crucial to search for strategies that optimize the selection of the sampling set over the graph. The theory developed in the last years for GSP was then applied to solve specific learning tasks, such as semi-supervised classification on graphs [19], graph dictionary learning [20], smooth graph signal recovery from random samples [21]–[24], inpainting [25], denoising [26], and adaptive estimation [27].

Almost all previous art considers centralized processing methods for graph signals. In many practical systems, data are collected in a distributed network, and sharing local information with a central processor is either unfeasible or not efficient, owing to the large size of the network and volume of data, time-varying network topology, bandwidth/energy constraints, and/or privacy issues. Centralized processing also calls for sufficient resources to transmit the data back and forth between the nodes and the fusion center, which limits the autonomy of the network, and may raise robustness concerns as well, since the central processor represents a bottleneck and an isolate point of failure. In addition, a centralized solution may limit the ability of the nodes to adapt in real-time to time-varying scenarios. Motivated by these observations, in this paper we focus on distributed techniques for graph signal processing. Some distributed methods were recently proposed in the literature, see, e.g. [28]–[30]. In [28], a distributed algorithm for graph signal inpainting is proposed; the work in [29] considers distributed processing of graph signals exploiting graph spectral dictionaries; finally, reference [30] proposes a distributed tracking method for

Manuscript received September 20, 2016; revised January 25, 2017 and March 29, 2017; accepted May 9, 2017. Date of publication May 25, 2017; date of current version June 16, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yuichi Tanaka. The work of P. Di Lorenzo was supported by the "Fondazione Cassa di Risparmio di Perugia." *(Corresponding author: Paolo Di Lorenzo.)*

P. Di Lorenzo and P. Banelli are with the Department of Engineering, University of Perugia, Perugia 06125, Italy (e-mail: paolo.dilorenzo@unipg.it; paolo.banelli@unipg.it).

S. Barbarossa and S. Sardellitti are with the Department of Information Engineering, Electronics, and Telecommunications, Sapienza University of Rome, Rome 00184, Italy (e-mail: sergio.barbarossa@uniroma1.it; stefania.sardellitti@uniroma1.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2017.2708035

time-varying bandlimited graph signals, assuming perfect observations (i.e., there is no measurement noise) and a fixed sampling strategy.

*Contributions of the paper:* In this work, we propose distributed strategies for adaptive learning of graph signals. The main contributions are listed in the sequel.

1) We formulate the problem of *distributed learning of graph signals* exploiting a *probabilistic sampling* scheme over the graph;
2) We provide necessary and sufficient *conditions for adaptive reconstruction* of the signal from the graph samples;
3) We apply diffusion adaptation methods to solve the problem of learning graph signals in a distributed manner. The resulting algorithm is a generalization of diffusion adaptation strategies where nodes sample data from the graph with some given probability.
4) We provide a detailed mean square analysis that illustrates the role of the probabilistic sampling strategy on the performance of the proposed algorithm.
5) We design useful strategies for the *distributed selection* of the (expected) sampling set. To the best of our knowledge, this is the first strategy available in the literature for distributed selection of graph signal's samples.

The work merges, for the first time in the literature, the well established field of adaptation and learning over networks, see, e.g., [31]–[39], with the emerging area of signal processing on graphs, see, e.g., [1]–[3]. The proposed method exploits the graph structure that describes the observed signal and, under a bandlimited assumption, enables adaptive reconstruction and tracking from a limited number of observations taken over a subset of vertices in a totally distributed fashion. Interestingly, the graph topology plays an important role both in the processing and communication aspects of the algorithm. A detailed mean-square analysis illustrates the role of the sampling strategy on the reconstruction capability, stability, and performance of the proposed algorithm. Thus, based on these results, we also propose a distributed method to select the set of sampling nodes in an efficient manner. An interesting feature of our proposed strategy is that this subset is allowed to vary over time, provided that the *expected* sampling set satisfies specific conditions enabling signal reconstruction. We expect that the proposed tools will represent a key technology for the distributed proactive sensing of cyber physical systems, where an effective control mechanism requires the availability of data-driven sampling strategies able to monitor the overall system by only checking a limited number of nodes.

The paper is organized as follows. In Section II, we introduce some basic GSP tools. Section III introduces the proposed distributed algorithm for adaptive learning of graph signals, illustrating also the conditions enabling signal reconstruction from a subset of samples. In Section IV we carry out a detailed mean-square analysis, whereas Section V is devoted to the development of useful strategies enabling the selection of the sampling set in a totally distributed fashion. Then, in Section VI we report several numerical simulations, aimed at assessing the validity of the theoretical analysis and the performance of the proposed algorithm. Finally, Section VII draws some conclusions.

## II. Graph Signal Processing Tools

In this section, we introduce some useful concepts from GSP that will be exploited along the paper. Let us consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ composed of $N$ nodes $\mathcal{V} = \{1, 2, ..., N\}$, along with a set of weighted edges $\mathcal{E} = \{a_{ij}\}_{i,j\in\mathcal{V}}$, such that $a_{ij} > 0$, if there is a link from node $j$ to node $i$, or $a_{ij} = 0$, otherwise. The adjacency matrix $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ is the collection of all the weights $a_{ij}, i, j = 1, \ldots, N$. The degree of node $i$ is $k_i := \sum_{j=1}^{N} a_{ij}$, and the degree matrix $\mathbf{K}$ is a diagonal matrix having the node degrees on its diagonal. The Laplacian matrix is defined as: $\mathbf{L} = \mathbf{K} - \mathbf{A}$. If the graph is *undirected*, the Laplacian matrix is symmetric and positive semi-definite, and admits the eigendecomposition $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, where $\mathbf{U}$ collects all the eigenvectors of $\mathbf{L}$ in its columns, whereas $\mathbf{\Lambda}$ contains the eigenvalues of $\mathbf{L}$. It is well known from spectral graph theory [40] that the eigenvectors of $\mathbf{L}$ are well suited for representing clusters, since they are signal vectors that minimize the $\ell_2$-norm graph total variation.

A signal $\boldsymbol{x}$ over a graph $\mathcal{G}$ is defined as a mapping from the vertex set to the set of complex numbers, i.e. $\boldsymbol{x} : \mathcal{V} \to \mathbb{C}$. In many applications, the signal $\boldsymbol{x}$ admits a compact representation, i.e., it can be expressed as:

$$\boldsymbol{x} = \mathbf{U}\boldsymbol{s} \tag{1}$$

where $\boldsymbol{s}$ is exactly (or approximately) sparse. As an example, in all cases where the graph signal exhibits clustering features, i.e. it is a smooth function within each cluster, but it is allowed to vary arbitrarily from one cluster to the other, the representation in (1) is compact, i.e., $\boldsymbol{s}$ is sparse. A key example is cluster analysis in semi-supervised learning, where a constant signal (label) is associated to each cluster [41]. The GFT $\boldsymbol{s}$ of a signal $\boldsymbol{x}$ is defined as the projection onto the orthogonal set of eigenvectors of the Laplacian [1], i.e.,

$$\boldsymbol{s} = \mathbf{U}^H\boldsymbol{x}. \tag{2}$$

The GFT has been defined in alternative ways, see, e.g., [1], [2], [8], [9]. In this paper, we basically follow the approach based on the Laplacian matrix, assuming an undirected graph structure, but the theory could be extended to handle directed graphs considering, e.g., a graph Fourier basis as proposed in [2]. Also, we denote the support of $\boldsymbol{s}$ in (1) as $\mathcal{F} = \{i \in \{1, \ldots, N\} : s_i \neq 0\}$, and the *bandwidth* of the graph signal $\boldsymbol{x}$ is defined as the cardinality of $\mathcal{F}$, i.e. $|\mathcal{F}|$. Clearly, combining (1) with (2), if the signal $\boldsymbol{x}$ exhibits a clustering behavior, in the sense specified above, the GFT is the way to recover the sparse vector $\boldsymbol{s}$. Finally, given a subset of vertices $\mathcal{S} \subseteq \mathcal{V}$, we define a vertex-limiting operator as the matrix

$$\mathbf{D}_{\mathcal{S}} = \mathrm{diag}\{\mathbf{1}_{\mathcal{S}}\}, \tag{3}$$

where $\mathbf{1}_{\mathcal{S}}$ is the set indicator vector, whose $i$-th entry is equal to one, if $i \in \mathcal{S}$, or zero otherwise.

## III. Distributed Learning of Graph Signals

We consider the problem of learning a (possibly time-varying) graph signal from observations taken from a subset of vertices of the graph. The problem fits well, e.g., to a wireless sensor network (WSN) scenario, where the nodes are observing a

spatial field related to some physical parameter of interest. Let us assume that the field is either fixed or slowly varying over both the time domain and the vertex (space) domain. Suppose now that the WSN is equipped with nodes that, at every time instant, can decide wether to take (noisy) observations of the underlying signal or not, depending on, e.g., energy constraints, failures, limited memory and/or processing capabilities, etc. Our purpose is to build adaptive techniques that allow the recovery of the field values at each node, pursued using recursive and distributed techniques as new data arrive. In this way, the information is processed on the fly by all nodes and the data diffuse across the network by means of a real-time sharing mechanism.

Let us consider a signal $\boldsymbol{x}^o = \{x_i^o\}_{i=1}^N \in \mathbb{C}^N$ defined over the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. To enable sampling of $\boldsymbol{x}^o$ without loss of information, the following is assumed:

*Assumption 1* (Bandlimited): The signal $\boldsymbol{x}^o$ is $\mathcal{F}$-bandlimited on the (time-invariant) graph $\mathcal{G}$, i.e., its spectral content is different from zero only on the set of indices $\mathcal{F}$. ■

Under Assumption 1, if the support $\mathcal{F}$ is known beforehand, from (1), the graph signal $\boldsymbol{x}^o$ can be cast in compact form as:

$$\boldsymbol{x}^o = \mathbf{U}_{\mathcal{F}} \boldsymbol{s}^o, \tag{4}$$

where $\mathbf{U}_{\mathcal{F}} \in \mathbb{C}^{N \times |\mathcal{F}|}$ collects the subset of columns of matrix $\mathbf{U}$ in (1) associated to the frequency indices $\mathcal{F}$, and $\boldsymbol{s}^o \in \mathbb{C}^{|\mathcal{F}| \times 1}$ is the vector of GFT coefficients of the frequency support of the graph signal $\boldsymbol{x}^o$. Let us assume that streaming and noisy observations of the graph signal are sampled over a (possibly time-varying) subset of vertices. In such a case, the observation taken by node $i$ at time $n$ can be expressed as:

$$y_i[n] \;=\; d_i[n]\,(x_i^o + v_i[n]) = d_i[n]\left(\boldsymbol{c}_i^H\,\boldsymbol{s}^o + v_i[n]\right), \tag{5}$$

$i = 1, \ldots, N$, where $^H$ denotes complex conjugate-transposition; $d_i[n] = \{0, 1\}$ is a random sampling binary coefficient, which is equal to 1 if node $i$ is taking the observation at time $n$, and 0 otherwise; $v_i[n]$ is zero-mean, spatially and temporally independent observation noise, with variance $\sigma_i^2$; also, in (5) we have used (4), where $\boldsymbol{c}_i^H \in \mathbb{C}^{1 \times |\mathcal{F}|}$ denotes the $i$-th row of matrix $\mathbf{U}_{\mathcal{F}}$. In the sequel, we assume that each node $i$ has local knowledge of its corresponding regression vector $\boldsymbol{c}_i$ in (5). This is a reasonable assumption even in the distributed scenario considered in this paper. For example, if neighbors in the processing graph can communicate with each other, either directly or via multi-hop routes, there exist many techniques that enable the distributed computation of eigenparameters of matrices describing sparse topologies such as the Laplacian or the adjacency, see, e.g., [42]–[44]. The methods are mainly based on the iterative application of distributed power iteration and consensus methods in order to iteratively compute the desired eigenparameters of the Laplacian (or adjacency) matrix, see, e.g., [44] for details. Since we consider graph signals with time-invariant topology, such procedures can be implemented offline during an initialization phase of the network to compute the required regression vectors in a totally distributed fashion. In the case of time-varying graphs, the distributed procedure should be adapted over time, but might result unpractical for large dynamic networks.

The distributed learning task consists in recovering the graph signal $\boldsymbol{x}^o$ from the noisy, streaming, and partial observations $y_i[n]$ in (5) by means of in-network adaptive processing. Following a least mean squares (LMS) estimation approach [31], [34]–[36], [45], the task can be cast as the cooperative solution of the following optimization problem:

$$\min_{\boldsymbol{s}} \;\sum_{i=1}^{N} \mathbb{E}_{\,\mathrm{d,v}} \left| d_i[n] \left( y_i[n] - \boldsymbol{c}_i^H\,\boldsymbol{s} \right) \right|^2, \tag{6}$$

where $\mathbb{E}_{\,\mathrm{d,v}}(\cdot)$ denotes the expectation operator evaluated over the random variables $\{d_i[n]\}_{i=1}^N$ and $\{v_i[n]\}_{i=1}^N$, and we have exploited $d_i[n]^2 = d_i[n]$ for all $i, n$. In the rest of the paper, to avoid overcrowded symbols, we will drop the subscripts in the expectation symbol referring to the random variables. In the sequel, we first analyze the conditions that enable signal recovery from a subset of samples. Then, we introduce adaptive strategies specifically tailored for the distributed reconstruction of graph signals from a limited number of samples.

### A. Conditions for Adaptive Reconstruction of Graph Signals

In this section, we give a necessary and sufficient condition guaranteeing signal reconstruction from its samples. In particular, assuming the random sampling and observations processes $\boldsymbol{d}[n] = \{d_i[n]\}_{i=1}^N$ and $\boldsymbol{y}[n] = \{y_i[n]\}_{i=1}^N$ to be stationary, the solution of problem (6) is given by the vector $\boldsymbol{s}^o$ that satisfies the normal equations:

$$\left( \sum_{i=1}^{N} \mathbb{E}\{d_i[n]\} \boldsymbol{c}_i \boldsymbol{c}_i^H \right) \boldsymbol{s}^o = \sum_{i=1}^{N} \mathbb{E}\{d_i[n] y_i[n]\}\, \boldsymbol{c}_i. \tag{7}$$

Letting $p_i = \mathbb{E}\{d_i[n]\}$, $i = 1, \ldots, N$, be the probability that node $i$ takes an observation at time $n$, from (7), it is clear that reconstruction of $\boldsymbol{s}^o$ is possible only if the matrix

$$\sum_{i=1}^{N} p_i \boldsymbol{c}_i \boldsymbol{c}_i^H \;= \mathbf{U}_{\mathcal{F}}^H \mathbf{P} \mathbf{U}_{\mathcal{F}} \tag{8}$$

is invertible, with $\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_N)$ denoting a vertex sampling operator as (3), but weighted by the sampling probabilities $\{p_i\}_{i=1}^N$. Let us denote the *expected sampling set* by

$$\overline{\mathcal{S}} = \{i = 1, \ldots, N \,|\, p_i > 0\}.$$

$\overline{\mathcal{S}}$ represents the set of nodes of the graph that collect data with a probability different from zero. From (7) and (8), a necessary condition enabling reconstruction is

$$|\overline{\mathcal{S}}| \geq |\mathcal{F}|, \tag{9}$$

i.e., the number of nodes in the expected sampling set must be greater than equal to the signal bandwidth. However, this condition is not sufficient, because matrix $\mathbf{U}_{\mathcal{F}}^H \mathbf{P} \mathbf{U}_{\mathcal{F}}$ in (8) may loose rank, or easily become ill-conditioned, depending on the graph topology and sampling strategy (defined by $\overline{\mathcal{S}}$ and $\mathbf{P}$). To provide a condition for signal reconstruction, we proceed similarly to [13], [16], [27]. Since $p_i > 0$ for all $i \in \overline{\mathcal{S}}$,

$$\mathrm{rank} \left( \sum_{i=1}^{N} p_i \boldsymbol{c}_i \boldsymbol{c}_i^H \right) = \mathrm{rank} \left( \sum_{i \in \overline{\mathcal{S}}} \boldsymbol{c}_i \boldsymbol{c}_i^H \right), \tag{10}$$

i.e., matrix (8) is invertible if matrix $\sum_{i \in \overline{\mathcal{S}}} \boldsymbol{c}_i \boldsymbol{c}_i^H = \mathbf{U}_{\mathcal{F}}^H \mathbf{D}_{\overline{\mathcal{S}}} \mathbf{U}_{\mathcal{F}}$ has full rank, where $\mathbf{D}_{\overline{\mathcal{S}}}$ is the vertex-limiting operator that

projects onto the expected sampling set $\overline{\mathcal{S}}$. Let us now introduce the operator

$$\mathbf{D}_{\overline{\mathcal{S}}_c} = \mathbf{I} - \mathbf{D}_{\overline{\mathcal{S}}}, \tag{11}$$

which projects onto the complement of the expected sampling set, i.e., $\overline{\mathcal{S}}_c = \{i = 1, \ldots, N \,|\, p_i = 0\}$. Then, exploiting (11), signal reconstruction is possible if

$$\mathbf{U}_{\mathcal{F}}^H \mathbf{D}_{\overline{\mathcal{S}}} \mathbf{U}_{\mathcal{F}} = \mathbf{I} - \mathbf{U}_{\mathcal{F}}^H \mathbf{D}_{\overline{\mathcal{S}}_c} \mathbf{U}_{\mathcal{F}}$$

is invertible, i.e., if condition

$$\boxed{\left\| \mathbf{D}_{\overline{\mathcal{S}}_c} \mathbf{U}_{\mathcal{F}} \right\|_2 < 1} \tag{12}$$

holds true. As shown in [13], [16], condition (12) is related to the localization properties of graph signals: It implies that there are no $\mathcal{F}$-bandlimited signals that are perfectly localized over the set $\overline{\mathcal{S}}_c$. Proceeding as in [13], [27], it is easy to show that condition (12) is necessary and sufficient for signal reconstruction. We remark that, differently from previous works on sampling of graph signals, see, e.g., [7], [9], [12]–[16], condition (12) now depends on the *expected* sampling set. This relaxed condition is due to the iterative nature of the adaptive learning mechanism considered in this paper. As a consequence, the algorithm does not need to collect all the data necessary to reconstruct one-shot the graph signal at each iteration, but can learn acquiring the needed information over time. The only important thing required by condition (12) is that a sufficiently large number of nodes collect data in *expectation* (i.e., belong to the expected sampling set $\overline{\mathcal{S}}$). In the sequel, we introduce the proposed distributed algorithm.

### B. Adaptive Distributed Strategies

In principle, the solution $\boldsymbol{s}^o$ of problem (6) can be computed as the vector that satisfies the linear system in (7). Nevertheless, in many linear regression applications involving online processing of data, the moments in (7) may be either unavailable or time-varying, and thus impossible to update continuously. For this reason, adaptive solutions relying on instantaneous information are usually adopted in order to avoid the need to know the signal statistics beforehand. Furthermore, the solution of (7) would require to collect all the data $\{y_i[n]\}_{i:d_i[n]=1}$, for all $n$, in a single processing unit that performs the computation. In this paper, our emphasis is on distributed, adaptive solutions, where the nodes perform the reconstruction task via online in-network processing only exchanging data between neighbors. To this aim, diffusion techniques were proposed and studied in literature [31]–[33], [46], [47]. In view of their robustness and adaptation properties, diffusion networks have been applied to solve a variety of learning tasks, such as, e.g., resource allocation problems [48], distributed optimization and learning [34], sparse distributed estimation [35], [45], [49], robust system modeling [50], and multi-task networks [37]–[39].

In the sequel, we provide an alternative approach to derive diffusion adaptation strategies with respect to the seminal papers [31], [32]. The derivations will be instrumental to introduce the main assumptions that will be exploited in the mean-square analysis, which will be carried out in the next section. In particular, to ensure the diffusion of information over the entire network, the following is assumed:

*Assumption 2 (Topology):* The communication graph is symmetric and connected, i.e., there exists an undirected path connecting any two vertices of the network. ■

To derive distributed solution methods for problem (6), let us introduce local copies $\{\boldsymbol{s}_i\}_{i=1}^N$ of the global variable $\boldsymbol{s}$, and recast problem (6) in the following equivalent form:

$$\min_{\{\boldsymbol{s}_i\}_{i=1}^N} \; \sum_{i=1}^N \mathbb{E}\left|d_i[n]\left(y_i[n] - \boldsymbol{c}_i^H \boldsymbol{s}_i\right)\right|^2 \tag{13}$$
$$\text{subject to} \quad \boldsymbol{s}_i = \boldsymbol{s}_j \quad \text{for all } i, j = 1, \ldots, N.$$

Under Assumption 2, it is possible to write the constraints in (13) in a compact manner, introducing the disagreement constraint that enforces consensus among the local variables $\{\boldsymbol{s}_i\}_{i=1}^N$ [51]. To this aim, let us denote with $\widetilde{\mathbf{A}} = \{\widetilde{a}_{ij}\}$ the adjacency matrix of the communication graph among the nodes, such that $\widetilde{a}_{ij} > 0$, if there is a communication link from node $j$ to node $i$, or $a_{ij} = 0$, otherwise. Then, under Assumption 2, problem (13) [and (6)] can be rewritten in the following equivalent form:

$$\min_{\{\boldsymbol{s}_i\}_{i=1}^N} \; \sum_{i=1}^N \mathbb{E}\left|d_i[n]\left(y_i[n] - \boldsymbol{c}_i^H \boldsymbol{s}_i\right)\right|^2 \tag{14}$$
$$\text{subject to} \quad \frac{1}{2}\sum_{i=1}^N \sum_{j=1}^N \widetilde{a}_{ij} \|\boldsymbol{s}_j - \boldsymbol{s}_i\|^2 \leq 0.$$

The Lagrangian for problem (14) writes as:

$$\mathcal{L}\left(\{\boldsymbol{s}_i\}_{i=1}^N, \lambda^o\right) = \sum_{i=1}^N \mathbb{E}\left|d_i[n]\left(y_i[n] - \boldsymbol{c}_i^H \boldsymbol{s}_i\right)\right|^2 + \frac{\lambda^o}{2}\sum_{i=1}^N \sum_{j=1}^N \widetilde{a}_{ij} \|\boldsymbol{s}_j - \boldsymbol{s}_i\|^2 \tag{15}$$

with $\lambda^o \geq 0$ denoting the (optimal) Lagrange multiplier associated with the disagreement constraint. At this stage, we do not need to worry about the selection of the Lagrange multiplier $\lambda^o$, because it will be embedded into a set of coefficients that the designer can choose. Then, we proceed by minimizing the Lagrangian function in (15) by means of a steepest descent procedure. Thus, letting $\boldsymbol{s}_i[n]$ be the instantaneous estimate of $\boldsymbol{s}^o$ at node $i$, we obtain:[1]

$$\begin{aligned}\boldsymbol{s}_i[n+1] &= \boldsymbol{s}_i[n] - \mu_i \left[\nabla_{\boldsymbol{s}_i} \mathcal{L}\left(\{\boldsymbol{s}_i[n]\}_{i=1}^N, \lambda^o\right)\right]^* \\ &= \boldsymbol{s}_i[n] + \mu_i \, \mathbb{E}\left\{d_i[n]\boldsymbol{c}_i(y_i[n] - \boldsymbol{c}_i^H \boldsymbol{s}_i[n])\right\} \\ &\quad + \mu_i \lambda^o \sum_{j=1}^N \widetilde{a}_{ij}(\boldsymbol{s}_j[n] - \boldsymbol{s}_i[n])\end{aligned} \tag{16}$$

for all $i = 1, \ldots, N$, where $[\nabla(\cdot)]^*$ denotes the complex gradient operator, and $\mu_i > 0$ are (sufficiently small) step-size coefficients. Now, using similar arguments as in [31], [34]–[36], we

[1] A factor of 2 multiplies (16) when the data are real. This factor was absorbed into the step-sizes $\mu_i$ in (16).

can accomplish the update (16) in two steps by generating an intermediate estimate $\boldsymbol{\psi}_i[n]$ as follows:

$$\boldsymbol{\psi}_i[n] = \boldsymbol{s}_i[n] + \mu_i \, \mathbb{E}\left\{ d_i[n] \boldsymbol{c}_i (y_i[n] - \boldsymbol{c}_i^H \boldsymbol{s}_i[n]) \right\} \quad (17)$$

$$\boldsymbol{s}_i[n+1] = \boldsymbol{\psi}_i[n] + \mu_i \lambda^o \sum_{j=1}^{N} \widetilde{a}_{ij} (\boldsymbol{\psi}_j[n] - \boldsymbol{\psi}_i[n]) \quad (18)$$

where in (18) we have replaced the variables $\{\boldsymbol{s}_i[n]\}_{i,n}$ with the intermediate estimates that are available at the nodes at time $n$, namely, $\{\boldsymbol{\psi}_i[n]\}_{i,n}$. Such kind of substitutions are typically used to derive adaptive diffusion implementations, see, e.g., [31]. Now, from (18), introducing the coefficients

$$w_{ii} = 1 - \mu_i \lambda^o \sum_{j=1}^{N} \widetilde{a}_{ij}, \text{ and } w_{ij} = \mu_i \lambda^o \widetilde{a}_{ij} \text{ for } i \neq j, \quad (19)$$

we obtain

$$\boldsymbol{s}_i[n+1] = \sum_{j=1}^{N} w_{ij} \boldsymbol{\psi}_j[n] \quad (20)$$

where the coefficients $\{w_{ij}\}$ are real, non-negative, weights matching the communication graph and satisfying:

$$w_{ij} = 0 \text{ for } j \notin \mathcal{N}_i, \text{ and } \mathbf{W}\mathbf{1} = \mathbf{1}, \quad (21)$$

where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the matrix with individual entries $\{w_{ij}\}$, and $\mathcal{N}_i = \{j = 1, \ldots, N \,|\, \widetilde{a}_{ij} > 0\} \bigcup \{i\}$ is the extended neighborhood of node $i$, which comprises node $i$ itself. Recursion (17) requires knowledge of the moments $\mathbb{E}\{d_i[n] y_i[n]\}$, which may be either unavailable or time-varying. An adaptive implementation can be obtained by replacing these moments by local instantaneous approximations, say, of the LMS type, i.e. $\mathbb{E}\{d_i[n] y_i[n]\} \approx d_i[n] y_i[n]$, for all $i, n$. The resulting algorithm is reported in Table 1, and will be termed as the Adapt-Then-Combine (ATC) diffusion strategy. The first step in (22) is an adaptation step, where the intermediate estimate $\boldsymbol{\psi}_i[n]$ is updated adopting the current observation taken by node $i$, i.e. $y_i[n]$. The second step is a diffusion step where the intermediate estimates $\boldsymbol{\psi}_j[n]$, from the spatial neighbors $j \in \mathcal{N}_i$, are combined through the weighting coefficients $\{w_{ij}\}$. Finally, given the estimate $\boldsymbol{s}_i[n]$ of the GFT at node $i$ and time $n$, the last step produces the estimate $x_i[n+1]$ of the graph signal value at node $i$ [cf. (5)]. We remark that by reversing the steps (17) and (18) to implement (16), we would arrive at a similar but alternative strategy, known as the Combine-then-Adapt (CTA) diffusion strategy. We continue our discussion by focusing on the ATC algorithm in (22); similar analysis applies to CTA [31].

*Remark 1:* The strategy (22) allows each node in the network to estimate and track the (possibly time-varying) GFT of the graph signal $\boldsymbol{x}^o$. From (22), it is interesting to notice that, while sampling nodes (i.e., those such that $d_i[n] = 1$) take and process the observations $y_i[n]$ of the graph signal, the role of the other nodes (i.e., those such that $d_i[n] = 0$) is only to allow the propagation of information coming from neighbors through a diffusion mechanism that percolates over all the network. From a complexity point of view, at every iteration $n$, the strategy in (22) requires that a node $i$ performs $O(3|\mathcal{F}|)$ computations, if $d_i[n] = 1$, and $O(2|\mathcal{F}|)$ computations, if $d_i[n] = 0$. Instead, from a communication point of view, each node in the network must transmit to its neighbors a vector composed of $|\mathcal{F}|$ (complex) values at every iteration $n$.

**Table 1:** ATC diffusion for graph signal learning.

**Data:** $\boldsymbol{s}_i[0]$ chosen at random for all $i$; $\{w_{ij}\}_{i,j}$ satisfying (21); (sufficiently small) step-sizes $\mu_i > 0$. Then, for each time $n \geq 0$ and for each node $i$, repeat:

$$\boldsymbol{\psi}_i[n] = \boldsymbol{s}_i[n] + \mu_i d_i[n] \boldsymbol{c}_i (y_i[n] - \boldsymbol{c}_i^H \boldsymbol{s}_i[n]) \quad \text{(adaptation step)}$$

$$\boldsymbol{s}_i[n+1] = \sum_{j \in \mathcal{N}_i} w_{ij} \boldsymbol{\psi}_j[n] \quad \text{(diffusion step)} \quad (22)$$

$$x_i[n+1] = \boldsymbol{c}_i^H \boldsymbol{s}_i[n+1] \quad \text{(reconstruction step)}$$

In this work, we assume that processing and communication graphs have in general distinct topologies. We remark that both graphs play an important role in the proposed distributed processing strategy (22). First, the processing graph determines the structure of the regression data $\boldsymbol{c}_i$ used in the adaptation step of (22). In fact, $\{\boldsymbol{c}_i^H\}_i$ are the rows of the matrix $\mathbf{U}_{\mathcal{F}}$, whose columns are the eigenvectors of the Laplacian matrix associated with the set of support frequencies $\mathcal{F}$. Then, the topology of the communication graph determines how the processed information is spread all over the network through the diffusion step in (22). This illustrates how, when reconstructing graph signals in a distributed manner, we have to take into account both the processing and communication aspects of the problem. ■

In the next section, we analyze the mean-square behavior of the proposed method, enlightening the role played by the sampling strategy on the final performance.

## IV. Mean-Square Performance Analysis

In this section, we analyze the performance of the ATC strategy in (22) in terms of its mean-square behavior. We remark that the analysis carried out in this section differs from classical derivations used for diffusion adaptation algorithms, see, e.g., [36]. First of all, the observation model in (5) is different from classical models generally adopted in the adaptive filtering literature, see, e.g. [52]. Also, due to the sampling operation and the presence of deterministic regressors [cf. (5)], each node cannot reconstruct the signal using only its own data (i.e., using stand-alone LMS adaptation), and must necessarily cooperate with other nodes in order to exploit information coming from other parts of the network. These issues require the development of a dedicated (non-trivial) analysis (see, e.g., Theorem 1 and the Appendix) to prove the mean-square stability of the proposed algorithm.

From now on, we view the estimates $\boldsymbol{s}_i[n]$ as realizations of a random process and analyze the performance of the ATC diffusion algorithm in terms of its mean-square behavior. To do

so, we introduce the error quantities

$$\boldsymbol{e}_i[n] = \boldsymbol{s}_i[n] - \boldsymbol{s}^o, \quad i = 1, \ldots, N,$$

and the network vector

$$\boldsymbol{e}[n] = \mathrm{col}\{\boldsymbol{e}_1[n], \ldots, \boldsymbol{e}_N[n]\}. \tag{23}$$

We also introduce the block diagonal matrix

$$\mathbf{M} = \mathrm{diag}\{\mu_1 \mathbf{I}_{|\mathcal{F}|}, \ldots, \mu_N \mathbf{I}_{|\mathcal{F}|}\}, \tag{24}$$

the extended block weighting matrix

$$\widehat{\mathbf{W}} = \mathbf{W} \otimes \mathbf{I}_{|\mathcal{F}|}, \tag{25}$$

where $\otimes$ denotes the Kronecker product operation, and the extended sampling operator

$$\widehat{\mathbf{D}}[n] = \mathrm{diag}\left\{d_1[n]\mathbf{I}_{|\mathcal{F}|}, \ldots, d_N[n]\mathbf{I}_{|\mathcal{F}|}\right\}. \tag{26}$$

We further introduce the block quantities:

$$\mathbf{Q} = \mathrm{diag}\left\{\boldsymbol{c}_1 \boldsymbol{c}_1^H, \ldots, \boldsymbol{c}_N \boldsymbol{c}_N^H\right\}, \tag{27}$$

$$\boldsymbol{g}[n] = \mathrm{col}\{\boldsymbol{c}_1 v_1[n], \ldots, \boldsymbol{c}_N v_N[n]\}. \tag{28}$$

Then, exploiting (23)–(28), we conclude from (22) that the following relation holds for the error vector:

$$\boldsymbol{e}[n+1] = \widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{D}}[n]\mathbf{Q})\boldsymbol{e}[n] + \widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{D}}[n]\boldsymbol{g}[n]. \tag{29}$$

This relation tells us how the network error vector evolves over time. As we can notice from (29), the sampling strategy affects the recursion in two ways: (a) it modifies the iteration matrix $\widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{D}}[n]\mathbf{Q})$ of the error; (b) it selects the noise contribution $\widehat{\mathbf{D}}[n]\boldsymbol{g}[n]$ only from a subset of nodes at time $n$. Relation (29) will be the launching point for the mean-square analysis carried out in the sequel. Before moving forward, we introduce an independence assumption on the sampling strategy, and a small step-sizes assumption.

*Assumption 3 (Independent sampling):* The sampling process $\{d_i[t]\}$ is temporally and spatially independent, for all $i = 1, \ldots, N$ and $t \leq n$. ■

*Assumption 4 (Small step-sizes):* The step-sizes $\{\mu_i\}$ are sufficiently small so that terms that depend on higher-order powers of $\{\mu_i\}$ can be neglected. ■

We now proceed by illustrating the mean-square stability and steady-state performance of the algorithm in (22).

### A. Mean-Square Stability

We now examine the behavior of the mean-square deviation $\mathbb{E}\|\boldsymbol{e}_i[n]\|^2$ for any of the nodes in the graph. Following energy conservation arguments [31], [36], we can establish the following variance relation:

$$\mathbb{E}\|\boldsymbol{e}[n+1]\|^2_{\boldsymbol{\Sigma}} = \mathbb{E}\|\boldsymbol{e}[n]\|^2_{\boldsymbol{\Sigma}'} + \mathbb{E}\{\boldsymbol{g}[n]^H \widehat{\mathbf{D}}[n]\mathbf{M}\widehat{\mathbf{W}}^T \boldsymbol{\Sigma}\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{D}}[n]\boldsymbol{g}[n]\} \tag{30}$$

where $\boldsymbol{\Sigma}$ is any Hermitian nonnegative-definite matrix that we are free to choose, and

$$\boldsymbol{\Sigma}' = \mathbb{E}(\mathbf{I} - \mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M})\widehat{\mathbf{W}}^T \boldsymbol{\Sigma}\widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{D}}[n]\mathbf{Q}). \tag{31}$$

Moreover, setting

$$\mathbf{G} = \mathbb{E}\left[\boldsymbol{g}[n]\boldsymbol{g}[n]^H\right] = \mathrm{diag}\left\{\sigma_1^2 \boldsymbol{c}_1 \boldsymbol{c}_1^H, \ldots, \sigma_N^2 \boldsymbol{c}_N \boldsymbol{c}_N^H\right\}, \tag{32}$$

we can rewrite (30) in the form

$$\mathbb{E}\|\boldsymbol{e}[n+1]\|^2_{\boldsymbol{\Sigma}} = \mathbb{E}\|\boldsymbol{e}[n]\|^2_{\boldsymbol{\Sigma}'} + \mathrm{Tr}(\boldsymbol{\Sigma}\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{G}\mathbf{M}\widehat{\mathbf{W}}^T) \tag{33}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace operator,

$$\widehat{\mathbf{P}} = \mathbb{E}\{\widehat{\mathbf{D}}[n]\} = \mathbf{P} \otimes \mathbf{I}_{|\mathcal{F}|},$$

and we have exploited the relation [cf. (26), (32), and Assumption 3]

$$\mathbb{E}\{\widehat{\mathbf{D}}[n]\boldsymbol{g}[n]\boldsymbol{g}[n]^H\widehat{\mathbf{D}}[n]\} = \mathbb{E}\{\widehat{\mathbf{D}}[n]\boldsymbol{g}[n]\boldsymbol{g}[n]^H\} = \widehat{\mathbf{P}}\mathbf{G}.$$

Let $\boldsymbol{\sigma} = \mathrm{vec}(\boldsymbol{\Sigma})$ and $\boldsymbol{\sigma}' = \mathrm{vec}(\boldsymbol{\Sigma}')$, where the $\mathrm{vec}(\cdot)$ notation stacks the columns of $\boldsymbol{\Sigma}$ on top of each other and $\mathrm{vec}^{-1}(\cdot)$ is the inverse operation. We will use interchangeably the notation $\|\boldsymbol{e}\|^2_{\boldsymbol{\sigma}}$ and $\|\boldsymbol{e}\|^2_{\boldsymbol{\Sigma}}$ to denote the same quantity $\boldsymbol{e}^H\boldsymbol{\Sigma}\boldsymbol{e}$. Using the Kronecker product property $\mathrm{vec}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{C}) = (\mathbf{C}^T \otimes \mathbf{A})\mathrm{vec}(\boldsymbol{\Sigma})$, we can vectorize both sides of (31) and conclude that (31) can be replaced by the simpler linear vector relation: $\boldsymbol{\sigma}' = \mathrm{vec}(\boldsymbol{\Sigma}') = \mathbf{H}\boldsymbol{\sigma}$, where $\mathbf{H}$ is the $N^2|\mathcal{F}|^2 \times N^2|\mathcal{F}|^2$ matrix:

$$\begin{aligned}\mathbf{H} &= \mathbb{E}\{(\mathbf{I} - \mathbf{Q}^T\widehat{\mathbf{D}}[n]\mathbf{M})\widehat{\mathbf{W}}^T \otimes (\mathbf{I} - \mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M})\widehat{\mathbf{W}}^T\} \\ &= (\mathbf{I}\otimes\mathbf{I})(\mathbf{I} - \mathbf{I}\otimes\mathbf{Q}\widehat{\mathbf{P}}\mathbf{M} - \mathbf{Q}^T\widehat{\mathbf{P}}\mathbf{M}\otimes\mathbf{I} \\ &\quad + \mathbb{E}\{\mathbf{Q}^T\widehat{\mathbf{D}}[n]\mathbf{M}\otimes\mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M}\})(\mathbf{W}^T\otimes\mathbf{W}^T).\end{aligned} \tag{34}$$

The last term in (34) can be computed in closed form. In particular, from (24), (26), and (27), it is easy to see how the term $\mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M}$ (and $\mathbf{Q}^T\widehat{\mathbf{D}}[n]\mathbf{M}$) in (34) has a block diagonal structure, which can be recast as:

$$\mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M} = \sum_{i=1}^{N} \mu_i d_i[n]\mathbf{C}_i, \tag{35}$$

where $\mathbf{C}_i = \mathbf{R}_i \otimes \boldsymbol{c}_i\boldsymbol{c}_i^H$, and $\mathbf{R}_i = \mathrm{diag}(\boldsymbol{r}_i)$, with $\boldsymbol{r}_i$ denoting the $i$-th canonical vector. Thus, exploiting (35), we obtain

$$\begin{aligned}&\mathbb{E}\{\mathbf{Q}^T\widehat{\mathbf{D}}[n]\mathbf{M}\otimes\mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M}\} \\ &\qquad = \sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i\mu_j m_{i,j}^{(2)} \mathbf{C}_i^T \otimes \mathbf{C}_j\end{aligned} \tag{36}$$

where, exploiting Assumption 3, we have

$$m_{i,j}^{(2)} = \mathbb{E}\{d_i[n]d_j[n]\} = \begin{cases} p_i, & \text{if } i = j; \\ p_i p_j, & \text{if } i \neq j. \end{cases} \tag{37}$$

Substituting (36) in (34), we obtain the final expression:

$$\begin{aligned}\mathbf{H} = (\mathbf{I}\otimes\mathbf{I})\Big(&\mathbf{I} - \mathbf{I}\otimes\mathbf{Q}\widehat{\mathbf{P}}\mathbf{M} - \mathbf{Q}^T\widehat{\mathbf{P}}\mathbf{M}\otimes\mathbf{I} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i\mu_j m_{i,j}^{(2)} \mathbf{C}_i^T \otimes \mathbf{C}_j\Big)(\mathbf{W}^T\otimes\mathbf{W}^T).\end{aligned} \tag{38}$$

Now, using the property $\mathrm{Tr}(\mathbf{\Sigma X}) = \mathrm{vec}(\mathbf{X}^T)^T \boldsymbol{\sigma}$, we can rewrite (33) as follows:

$$\mathbb{E}\|\boldsymbol{e}[n+1]\|^2_{\boldsymbol{\sigma}} = \mathbb{E}\|\boldsymbol{e}[n]\|^2_{\mathbf{H}\boldsymbol{\sigma}} + \mathrm{vec}(\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{G}\mathbf{M}\widehat{\mathbf{W}}^T)^T\boldsymbol{\sigma}. \tag{39}$$

The following theorem guarantees the asymptotic mean-square stability (i.e., stability in the mean and mean-square sense) of the diffusion strategy (22).

*Theorem 1* (Mean-square stability) Assume model (5), condition (12), Assumptions 2, 3, and 4 hold. Then, for any initial condition and choice of the matrices $\mathbf{W}$ satisfying (21) and $\mathbf{1}^T\mathbf{W} = \mathbf{1}^T$, the algorithm (22) is mean-square stable.

*Proof:* Let $\boldsymbol{r} = \mathrm{vec}(\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{G}\mathbf{M}\widehat{\mathbf{W}}^T)$. From (39), we get

$$\mathbb{E}\|\boldsymbol{e}[n]\|^2_{\boldsymbol{\sigma}} = \mathbb{E}\|\boldsymbol{e}[0]\|^2_{\mathbf{H}^n\boldsymbol{\sigma}} + \boldsymbol{r}^T\sum_{l=0}^{n-1}\mathbf{H}^l\boldsymbol{\sigma} \tag{40}$$

where $\mathbb{E}\|\boldsymbol{e}[0]\|^2$ is the initial condition. We first note that if $\mathbf{H}$ is stable, $\mathbf{H}^n \to \mathbf{0}$ as $n \to \infty$. In this way, the first term on the RHS of (40) vanishes asymptotically. At the same time, the convergence of the second term on the RHS of (40) depends only on the geometric series of matrices $\sum_{l=0}^{\infty}\mathbf{H}^l$, which is known to be convergent to a finite value if the matrix $\mathbf{H}$ is a stable matrix [53]. Thus, the stability of matrix $\mathbf{H}$ is a sufficient condition for the convergence of the mean-square recursion $\mathbb{E}\|\boldsymbol{e}[n]\|^2_{\boldsymbol{\sigma}}$ in (40) to a steady-state value.

To verify the stability of $\mathbf{H}$, we use the following approximation, which is accurate under Assumption 4, see, e.g., [31], [34], [35]. Then, we approximate (34) as:[2]

$$\begin{aligned}\mathbf{H} \;\approx\; & (\mathbf{I}\otimes\mathbf{I})(\mathbf{I}-\mathbf{I}\otimes\mathbf{Q}\widehat{\mathbf{P}}\mathbf{M}-\mathbf{Q}^T\widehat{\mathbf{P}}\mathbf{M}\otimes\mathbf{I}\\ & +\mathbf{Q}^T\widehat{\mathbf{P}}\mathbf{M}\otimes\mathbf{Q}\widehat{\mathbf{P}}\mathbf{M})(\mathbf{W}^T\otimes\mathbf{W}^T) \;=\; \mathbf{B}^T\otimes\mathbf{B}^H\end{aligned} \tag{41}$$

with $\mathbf{B}$ given by

$$\mathbf{B} = \widehat{\mathbf{W}}(\mathbf{I}-\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}). \tag{42}$$

Thus, from (41), exploiting the properties of the Kronecker product, we deduce that matrix $\mathbf{H}$ in (34) is stable if matrix $\mathbf{B}$ in (42) is also stable. Under the assumptions of Theorem 1, in the Appendix, we provide the proof of the stability of matrix $\mathbf{B}$ in (42). This concludes the proof of Theorem 1. ■

*Remark 2:* The assumptions used in Theorem 1 are *sufficient* conditions for graph signal reconstruction using the ATC diffusion algorithm in (22). In particular, (12) requires that the network collects samples from a sufficiently large number of nodes on average, and guarantees the existence of a *unique* solution of the normal equations in (7). Furthermore, (12) and Assumption 4 are also instrumental to prove the stability of matrix $\mathbf{B}$ in (42) [and of $\mathbf{H}$ in (34)] and, consequently, the stability in the mean and mean-square sense of the diffusion algorithm in (22) (see the Appendix). ■

[2]It is immediate to see that (41) can be obtained from (38) by replacing the term $\mathbb{E}\{\mathbf{Q}^T\widehat{\mathbf{D}}[n]\mathbf{M}\otimes\mathbf{Q}\widehat{\mathbf{D}}[n]\mathbf{M}\}$ with $\mathbf{Q}^T\widehat{\mathbf{P}}\mathbf{M}\otimes\mathbf{Q}\widehat{\mathbf{P}}\mathbf{M}$. This step coincides with substituting the terms $p_i$ in (36)–(37) with $p_i^2$, for all $i = 1,\ldots,N$. Such approximation appears in (41) only in the term $\mathbf{Q}^T\widehat{\mathbf{P}}\mathbf{M}\otimes\mathbf{Q}\widehat{\mathbf{P}}\mathbf{M} = O(\mu^2_{\max})$ and, consequently, under Assumption 4 it is assumed to produce a negligible deviation from (38).

*Remark 3:* In Theorem 1, we require the matrix $\mathbf{W}$ to be doubly stochastic. Note that, from the definition of weights $\{w_{ij}\}$ in (19), under Assumption 2, this further condition would imply that the step-sizes $\mu_i$ must be chosen constant for all $i$. However, as a consequence of Theorem 1, our strategy works for *any* choice of doubly stochastic matrices $\mathbf{W}$, without imposing the constraint that the step-sizes must be chosen constant for all $i$. Several possible combination rules have been proposed in the literature, such as the Laplacian or the Metropolis-Hastings weights, see, e.g. [31], [51], [54]. ■

### B. Steady-State Performance

Taking the limit as $n \to \infty$ (assuming convergence conditions are satisfied), we deduce from (39) that:

$$\lim_{n\to\infty}\mathbb{E}\|\boldsymbol{e}[n]\|^2_{(\mathbf{I}-\mathbf{H})\boldsymbol{\sigma}} = \mathrm{vec}(\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{G}\mathbf{M}\widehat{\mathbf{W}}^T)^T\boldsymbol{\sigma}. \tag{43}$$

Expression (43) is a useful result: it allows us to derive several performance metrics through the proper selection of the free weighting parameter $\boldsymbol{\sigma}$ (or $\mathbf{\Sigma}$), as was done in [31]. For example, the Mean-Square Deviation (MSD) for any node $i$ is defined as the steady-state value $\mathbb{E}|\widetilde{x}_i[n]|^2$, as $n \to \infty$, where $\widetilde{x}_i[n] = x_i[n] - x_i^o[n]$, for all $i = 1,\ldots,N$, with $x_i[n]$ defined in (22). From (22), this value can be obtained by computing $\lim_{n\to\infty}\mathbb{E}\|\boldsymbol{e}[n]\|^2_{\mathbf{T}_i}$, with a block weighting matrix $\mathbf{T}_i = \mathbf{R}_i\otimes\boldsymbol{c}_i\boldsymbol{c}_i^H$, where $\mathbf{R}_i = \mathrm{diag}(\boldsymbol{r}_i)$, with $\boldsymbol{r}_i$ denoting the $i$-th canonical vector. Then, from (43), the MSD at node $i$ can be obtained as:

$$\begin{aligned}\mathrm{MSD}_i &= \lim_{n\to\infty}\mathbb{E}\,|\widetilde{x}_i[n]|^2 = \lim_{n\to\infty}\mathbb{E}\|\boldsymbol{e}[n]\|^2_{\mathbf{R}_i\otimes\boldsymbol{c}_i\boldsymbol{c}_i}\\ &= \mathrm{vec}(\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{G}\mathbf{M}\widehat{\mathbf{W}}^T)^T(\mathbf{I}-\mathbf{H})^{-1}\mathrm{vec}\left(\mathbf{R}_i\otimes\boldsymbol{c}_i\boldsymbol{c}_i^H\right).\end{aligned} \tag{44}$$

Finally, letting $\widetilde{\boldsymbol{x}}[n] = \{\widetilde{x}_i[n]\}_{i=1}^N$, from (44), the network MSD is given by:

$$\begin{aligned}\mathrm{MSD} &= \lim_{n\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{x}}[n]\|^2\\ &= \mathrm{vec}(\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{G}\mathbf{M}\widehat{\mathbf{W}}^T)^T(\mathbf{I}-\mathbf{H})^{-1}\boldsymbol{q},\end{aligned} \tag{45}$$

where $\boldsymbol{q} = \mathrm{vec}(\sum_{i=1}^N\mathbf{R}_i\otimes\boldsymbol{c}_i\boldsymbol{c}_i^H) = \mathrm{vec}(\mathbf{Q})$ [cf. (27)]. In the sequel, we will confirm the validity of these theoretical expressions by comparing them with numerical results.

## V. Distributed Graph Sampling Strategies

As illustrated in the previous sections, the properties of the proposed distributed algorithm in (22) for graph signal reconstruction strongly depend on the expected sampling set $\overline{\mathcal{S}}$. Thus, building on the results obtained in Section IV, it is fundamental to devise (possibly distributed) strategies that design the set $\overline{\mathcal{S}}$, with the aim of reducing the computational/memory burden while still guaranteing provable theoretical performance. To this purpose, in this section we propose a distributed method that iteratively selects vertices from the graph in order to build an expected sampling set $\overline{\mathcal{S}}$ that enables reconstruction with a limited number of nodes, while ensuring guaranteed performance of the learning task.

To select the best sampling strategy, one should optimize some (global) performance criterion, e.g. the MSD in (45), with respect to the expected sampling set $\overline{\mathcal{S}}$, or, equivalently, the weighted vertex limiting operator $\widehat{\mathbf{P}}$. However, the solution of such problem would require global information about the entire graph to be collected into a single processing unit. To favor distributed implementations, we propose to consider a different (but related) performance metric for the selection of the sampling set, which comes directly from the solution of the normal equations in (7). In particular, to allow reconstruction of the graph signal, a good sampling strategy should select a sufficiently large number of vertices $i \in \mathcal{V}$ to favor the invertibility of the matrix in (8). In the sequel, we assume that the probabilities $\{p_i\}_{i=1}^N$ are given, either because they are known apriori or can be estimated locally at each node. In this context, the design of the sampling probabilities $\{p_i\}_{i=1}^N$ is an important task, which will be tackled in a future work.

Let us then consider the general selection problem:

$$\mathcal{S}^* = \arg\max_{\overline{\mathcal{S}}} \; h(\overline{\mathcal{S}}) = f\left( \sum_{i \in \overline{\mathcal{S}}} \frac{p_i}{1+\sigma_i^2} \boldsymbol{c}_i \boldsymbol{c}_i^H \right) \qquad (46)$$

$$\text{subject to} \quad |\overline{\mathcal{S}}| = M$$

where $\overline{\mathcal{S}}$ is the expected sampling set; $M$ is the given number of vertices to be selected; the weighting terms $p_i/(1+\sigma_i^2)$ take into account (possibly) heterogeneous sampling and noise conditions at each node; and $f(\cdot) : \mathbb{C}^{|\mathcal{F}| \times |\mathcal{F}|} \rightarrow \mathbb{R}$ is a function that measures the degree of invertibility of the matrix in its argument, e.g., the (logarithm of) pseudo-determinant, as proposed in [13], [27], [55], or the minimum eigenvalue, as proposed in [9]. As an example, taking $f(\cdot)$ as the (logarithm of) pseudo-determinant function, the solution of problem (46) aims at selecting $M$ rows $\boldsymbol{c}_i^H$ of matrix $\mathbf{U}_{\mathcal{F}}$, properly weighted by the terms $\sqrt{p_i/(1+\sigma_i^2)}$, such that the volume of the parallelepiped built by these vectors is maximized. Thus, intuitively, the method will tend to select vertices with: (a) large sampling probabilities $p_i$'s; (b) low noise variances $\sigma_i^2$'s; and (c) such that their corresponding regression vectors $\boldsymbol{c}_i$'s are large in magnitude and as orthogonal as possible. However, since the formulation in (46) translates inevitably into a selection problem, whose solution in general requires an exhaustive search over all the possible combinations, the complexity of such procedure becomes intractable also for graph signals of moderate dimensions. To cope with these issues, in the sequel we will provide an efficient, albeit sub-optimal, greedy strategy that tackles the problem of selecting the (expected) sampling set in a distributed fashion.

The greedy approach is described in Table 2. The simple idea underlying the proposed approach is to iteratively add to the sampling set those vertices of the graph that lead to the largest increment of the performance metric $h\left(\overline{\mathcal{S}}\right)$ in (46). In particular, the implementation of the distributed algorithm in Table 2 proceeds as follows. Given the current instance of the (expected) sampling set $\overline{\mathcal{S}}$, at Step 1, each node $j \notin \overline{\mathcal{S}}$ evaluates locally the value of the objective function $h\left(\overline{\mathcal{S}} \cup j\right)$ that the network would achieve if node $j$ was added to $\overline{\mathcal{S}}$. Then, in step 2, the network finds the maximum among the local values computed at the previous step. This task can be easily obtained with a distributed iterative procedure as, e.g., a maximum consensus algorithm [56], which is guaranteed to converge in a number of iterations less than equal to $\mathcal{D}$, with $\mathcal{D}$ denoting the diameter of the network. A node $j \notin \overline{\mathcal{S}}$ can then understand if it is the one that has achieved the maximum by simply comparing the value $h\left(\overline{\mathcal{S}} \cup j\right)$ computed at step 1, with the result of the distributed procedure in Step 2. The node $s^*$, which has achieved the maximum value at step 2, is then added to the expected sampling set. Finally, the weighted regression vector associated to the selected node, i.e. $\sqrt{p_{s^*}/(1+\sigma_{s^*}^2)}\boldsymbol{c}_{s^*}$, is diffused over the network through a flooding process, which terminates in a number of iterations less than or equal to $\mathcal{D}$. This allows each node not belonging to the sampling set to evaluate step 1 of the algorithm at the next round. This procedure continues until the network has added $M$ nodes to the expected sampling set.

**Table 2:** Distributed Graph Sampling Strategy.

$Input\ Data$: $M$, the number of sampling nodes. $\overline{\mathcal{S}} \equiv \emptyset$.
$Output\ Data$: $\overline{\mathcal{S}}$, the expected sampling set.
$Function$:
while $|\overline{\mathcal{S}}| < M$
1) Each node $j$ computes locally $h\left(\overline{\mathcal{S}} \cup j\right)$, for all $j \notin \overline{\mathcal{S}}$;
2) Distributed selection of the maximum: find
$$s^* = \arg\max_{j \notin \overline{\mathcal{S}}} \; h\left(\overline{\mathcal{S}} \cup j\right)$$
3) $\overline{\mathcal{S}} \leftarrow \overline{\mathcal{S}} \cup \{s^*\}$;
4) Diffusion of $\sqrt{\frac{p_{s^*}}{1+\sigma_{s^*}^2}}\,\boldsymbol{c}_{s^*}$ over the network;
end

In principle, there is no insurance that the selection path followed by the algorithm in Table 2 is the best one. In general, the performance of the proposed distributed strategy will be sub-optimal with respect to an exhaustive search procedure over all the possible combinations. Nevertheless, selecting the function $h\left(\overline{\mathcal{S}}\right)$ in (46) as the logarithm of the pseudo determinant, it is possible to prove that $h\left(\overline{\mathcal{S}}\right)$ is a monotone sub-modular function, and that greedy selection strategies (e.g., Table 2) achieve performance within $1 - 1/e$ of the optimal combinatorial solution [57], [58]. From a communication point of view, in the worst case, the procedure in Table 2 requires that each node exchanges $M\mathcal{D}(1+2|\mathcal{F}|)$ scalar values to accomplish the distributed task of sampling set selection. This procedure can be run offline once for all during the initialization phase of the network, when the set of sampling nodes must be decided. In the case of time-varying scenarios, e.g. switching graph topologies, link failures, time-varying spectral properties of the graph signal, the procedure should be repeated periodically in order to cope with such dynamicity. Of course, the procedure might result unpractical in the case of large, rapidly time-varying graphs. In such a case, future investigations are needed for practical and efficient implementations of distributed adaptive graph sampling strategies.

In the sequel, we will illustrate numerical results assessing the performance achieved by the proposed sampling strategies.

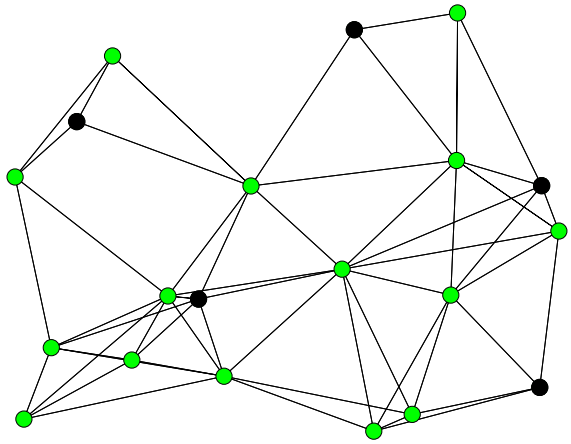

Fig. 1. Network graph, and sampling set (black nodes).

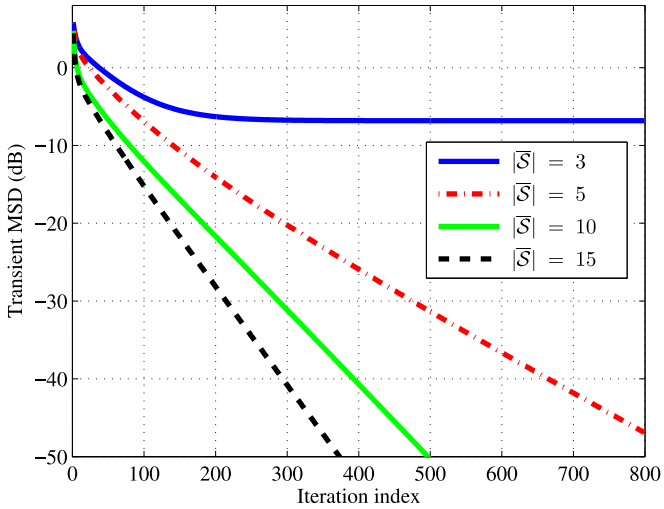


Fig. 2. Convergence behavior: Transient MSD in the noise-free case, considering static sampling. $|\mathcal{F}| = 5$.

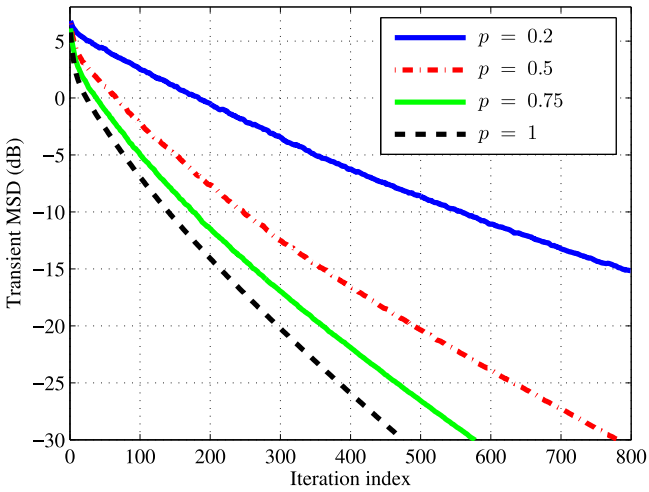


Fig. 3. Convergence behavior: Transient MSD in the noise-free case, considering random sampling. $|\mathcal{F}| = 5$, $|\overline{\mathcal{S}}| = 5$.

## VI. Numerical Results

In this section, we illustrate some numerical simulations aimed at assessing the performance of the proposed strategy for distributed learning of signals defined over graphs. First, we will illustrate the convergence properties of the proposed algorithm in absence of observation noise. Second, we will confirm the theoretical results in (39) and (44)–(45), which quantify the transient behavior and steady-state performance of the algorithm. Third, we will illustrate how the choice of the sampling strategy (see, e.g., Table 2) affects the performance of the proposed algorithm. Fourth, we will evaluate the tracking capabilities of the proposed technique, considering the presence of stochastic processes evolving over the graph. Finally, we apply the proposed strategy to estimate and track the spatial distribution of electromagnetic power in the context of cognitive radio networks.

*1) Convergence in the Noise-Free Case*: Let us consider a network composed of $N = 20$ nodes, whose topology (for both processing and communication tasks) is depicted in Fig. 1. We generate a graph signal from (1) having a spectral content limited to the first five eigenvectors of the Laplacian matrix of the graph in Fig. 1. Thus, the signal bandwidth is equal to $|\mathcal{F}| = 5$. For simplicity, we use the graph illustrated in Fig. 1 for both communication and processing tasks. To illustrate the perfect reconstruction capabilities of the proposed method in absence of noise, in this simulation we set $v_i[n] = 0$ for all $i, n$ in (5). Then, in Fig. 2 we report the transient behavior of the squared error $\|\widetilde{\boldsymbol{x}}[n]\|^2$ obtained by the ATC algorithm in (22), where $\widetilde{\boldsymbol{x}}[n] = \{x_i[n] - x_i^o\}_{i=1}^N$, with $x_i[n]$ defined in (22) for all $i$. In particular, we report four behaviors, each one associated to a different static sampling strategy (i.e., $p_i = 1$ for all $i \in \overline{\mathcal{S}}$), with $|\overline{\mathcal{S}}|$ equal to 3, 5, 10, and 15, respectively. The samples are chosen according to the distributed strategy proposed in Table 2, where the function $f(\cdot)$ is chosen to be the logarithm of the pseudo-determinant. From now on, we will denote this choice as the Max-Det sampling strategy. Also, we set $p_i = 1$, and $\sigma_i^2 = 0$, for all $i$ (because nor noise nor sampling probability play any role in the selection of the samples). An example of graph sampling in the case $|\overline{\mathcal{S}}| = 5$ is given in Fig. 1, where the black vertices correspond to the sampling nodes. The step-sizes $\mu_i$ in (22) are chosen equal to 0.5 for all $i$; the combination weights $\{w_{ij}\}$ are selected using the Metropolis rule [54], where $\widetilde{a}_{ij} = 1$ if nodes $i$ and $j$ are connected, and $\widetilde{a}_{ij} = 0$ otherwise. As we can notice from Fig. 2, as long as condition (12) is satisfied (see Section III-A), the algorithm drives to zero the error asymptotically, thus perfectly reconstructing the entire signal from a limited number of samples in a totally distributed manner. In particular, as expected, increasing the number of sampling nodes, the learning rate of the algorithm improves. On the contrary, when $|\overline{\mathcal{S}}| < |\mathcal{F}|$, e.g., in the case $|\overline{\mathcal{S}}| = 3$, condition (12) cannot be satisfied in any way (i.e., the signal is downsampled), and the algorithm cannot reconstruct the graph signal, as shown in Fig. 2.

To illustrate the convergence properties of the proposed strategy in the presence of probabilistic sampling (i.e., $0 < p_i < 1$ for $i \in \overline{\mathcal{S}}$), in Fig. 3 we report the average transient behavior of the squared error $\|\widetilde{\boldsymbol{x}}[n]\|^2$ obtained by the ATC algorithm in (22), considering different values of sampling probability

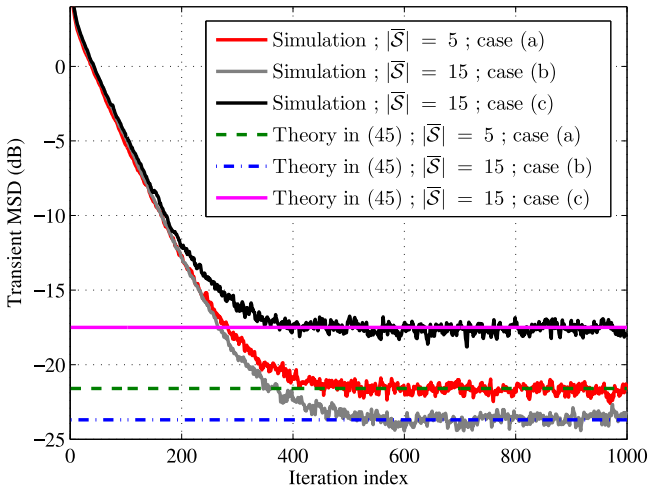


Fig. 4. Mean-Square performance: Transient MSD, and theoretical steady-state MSD, for different values of $|\overline{\mathcal{S}}|$. $|\mathcal{F}| = 5$.

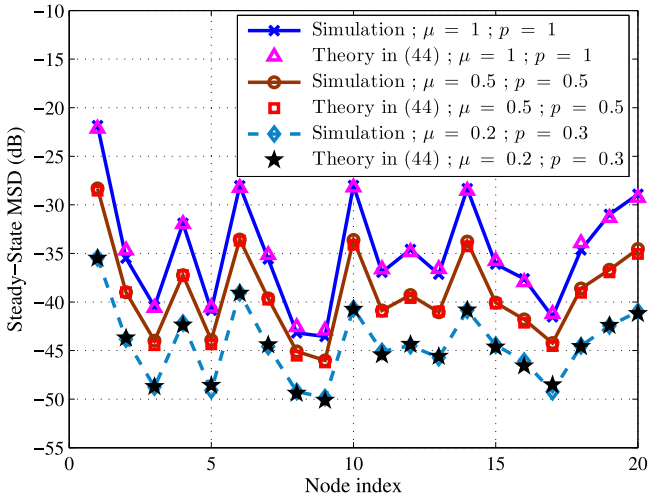


Fig. 5. Mean-Square performance: Theoretical and numerical steady-state MSD versus node index. $|\mathcal{F}| = 5$, $|\overline{\mathcal{S}}| = 10$.

$p_i = p$ for all $i \in \overline{\mathcal{S}}$. The signal bandwidth is equal to $|\mathcal{F}| = 5$, and the expected sampling set is composed of 5 nodes selected according to the Max-Det sampling strategy. The results are averaged over 100 independent simulations. The step-sizes and the combination weights are chosen as before. As we can notice from Fig. 3, since $\overline{\mathcal{S}}$ satisfies condition (12), i.e., the network collects samples from a sufficient number of nodes on average, the algorithm drives to zero the error for any value of $p$. As expected, increasing the sampling probability at each node, the learning rate of the proposed algorithm improves.

*2) Mean-Square Performance:* Now, we illustrate the mean-square behavior of the proposed strategy in the presence of observation noise in (5). As a first example, we report the transient behavior of the network MSD obtained by the ATC algorithm in (22), versus the iteration index, for different number of nodes collecting samples from the network: (a) $|\overline{\mathcal{S}}| = 5$; (b) $|\overline{\mathcal{S}}| = 15$; (c) $|\overline{\mathcal{S}}| = 15$. The difference between the three cases (a), (b) and (c) is also in the observation noise. In particular, in (a) and (b), the noise at the sampling nodes is chosen to be zero-mean, Gaussian, with variance chosen at random between 0 and 0.1. In case (c), the noise variance is chosen equal to case (a) for the first $|\overline{\mathcal{S}}| = 5$ nodes belonging also to case (a), whereas it is chosen equal to 0.4 for the remaining 10 sampling nodes. The expected sampling set is chosen according to the Max-Det strategy, and the sampling probabilities are set equal to $p_i = 0.8$ for all $i \in \overline{\mathcal{S}}$. The signal bandwidth is equal to $|\mathcal{F}| = 5$. The combination weights are chosen as before, and the step-sizes are selected in order to match the learning rates of the algorithm. The curves are averaged over 200 independent simulations, and the corresponding theoretical steady-state values in (45) are reported for the sake of comparison. As we can notice from Fig. 4, the theoretical predictions match well the simulation results. Furthermore, we notice how, when varying the number of nodes collecting samples, the algorithm might lead to lower or larger steady-state errors. This illustrates that, when reconstructing a graph signal in the presence of noise, it is not always better to increase the number of samples, as this implies an increment of the overall noise injected in the algorithm. In particular, the steady-state performance can improve or degrade by enlarging the sampling set, depending on the distribution of noise over the network. Intuitively, if the noise variance is almost uniform and low at each node of the network, it is convenient to add samples to the algorithm [as from case (a) to case (b)], which improves its learning rate/steady-state performance tradeoff. On the contrary, if some nodes have very noisy observations, it might be not convenient to take their samples (as from case (a) to case (c)), as this might lead to a performance degradation.

As a further example aimed at validating the theoretical results in (44), in Fig. 5 we report the behavior of the theoretical steady-state MSD values achieved at each vertex of the graph, comparing them with simulation results, for different values of the sampling probability $p$, and of the step-sizes $\mu_i = \mu$ for all $i$. The numerical results are obtained averaging over 200 independent simulations and 500 samples of squared error after convergence of the algorithm. The signal bandwidth is equal to $|\mathcal{F}| = 5$, and the expected sampling set is composed of $|\overline{\mathcal{S}}| = 10$ nodes. We can notice from Fig. 5 how the theoretical values in (44) predict well the simulation results. As expected, reducing the step-size and the sampling probability, the steady-state MSD of the algorithm improves.

Finally, in Fig. 6, we validate the theoretical expression for the transient MSD in (39), comparing it with numerical results, for different values of the step-sizes $\mu_i = \mu$ for all $i$. The numerical results are obtained averaging over 200 independent simulations, the signal bandwidth is equal to $|\mathcal{F}| = 2$, $p_i = 0.5$ for all $i \in \overline{\mathcal{S}}$, and $|\overline{\mathcal{S}}| = 10$ nodes. We can notice from Fig. 6 how the theoretical behaviors in (39) predict well the numerical results. As expected, reducing the step-size, the algorithm becomes slower, but the steady-state MSD improves.

*3) Effect of Sampling Strategy:* As previously remarked, it is fundamental to assess the performance of the algorithm in (22) with respect to the strategy adopted to select the expected sampling set $\overline{\mathcal{S}}$. Indeed, when sampling a graph signal, what matters is not only the number of samples, but also (and most important) where the samples are taken. From (45), we can in fact deduce that the sampling set plays a fundamental role, since

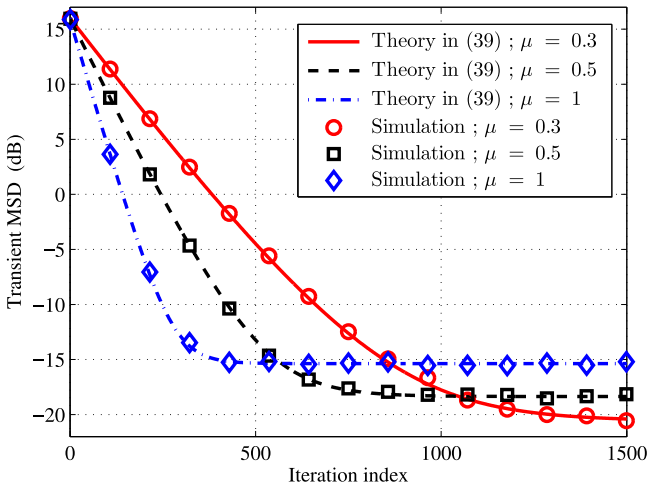


Fig. 6. Mean-Square performance: Numerical and theoretical transient MSD, for different values of $\mu$. $|\mathcal{F}| = 2$, $|\overline{\mathcal{S}}| = 10$.

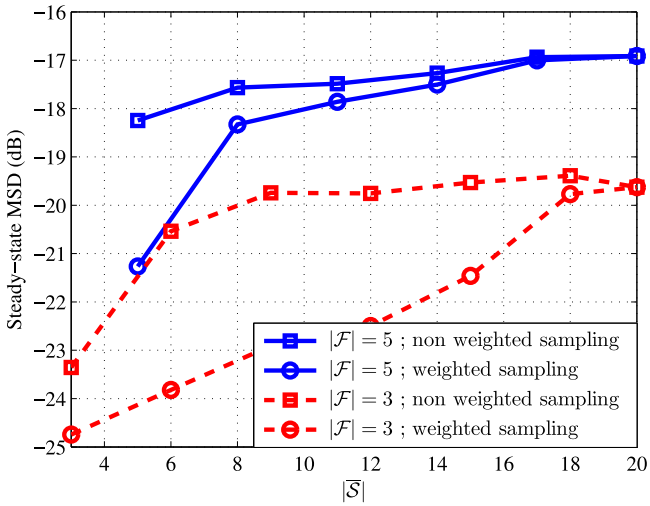


Fig. 7. Effect of sampling: Steady-state MSD versus $|\overline{\mathcal{S}}|$, for different graph signal bandwidths and sampling strategies.

it affects the performance of the proposed strategy in two ways: (a) it determines the stability of the iteration matrix $\mathbf{B}$ in (42), i.e., $\mathbf{H}$ in (45); (b) it allows us to select the nodes that inject noise into the system. As a first example, we aim at illustrating the performance obtained by the algorithm in (22) under different noise conditions at each node in the network, thus illustrating how selecting samples in a right manner can help reduce the effect of noisy nodes. In particular, we adopt the Max-Det sampling strategy, where the sampling probabilities are set equal to $p_i = 0.8$ for all $i \in \overline{\mathcal{S}}$. The noise at each node is chosen to be zero-mean, Gaussian, with a variance chosen uniformly random between 0 and 0.1. The step-sizes are $\mu_i = 0.5$ for all $i$, and the combination weights are chosen as before; we also consider the graph in Fig. 1. Then, in Fig. 7, we report the steady-state MSD obtained by the algorithm in (22), versus $|\overline{\mathcal{S}}|$, for different values of bandwidth $|\mathcal{F}|$ of the graph signal. The curves are averaged over 500 independent simulations. In particular, we

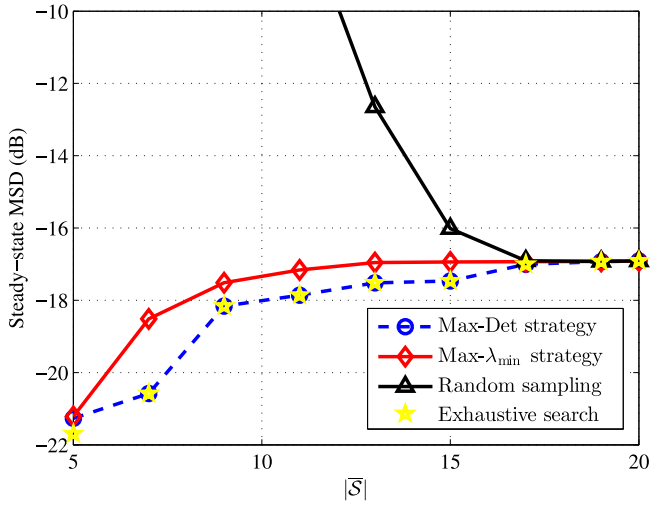


Fig. 8. Effect of sampling: Steady-state MSD versus $|\overline{\mathcal{S}}|$, for different sampling strategies.

consider two variants of the sampling strategy: (a) a weighted strategy as in Table 2, where each local element is weighted by the variance $\sigma_i^2$ of the noise for all $i$ (see, e.g., (46)); and (b) a non-weighted strategy, corresponding to setting $\sigma_i^2 = 0$ for all $i$, in Table 2. As we can notice from Fig. 7, the weighted strategy always outperforms the non-weighted method; this happens because the weighted strategy tends to select sampling nodes with smaller noise variance, thus leading to better performance. Interestingly, the gain is larger at lower bandwidths, thanks to the larger freedom that the method has in the selection of the (noisy) samples.

As a further example, in Fig. 8, we illustrate the steady-state MSD of the algorithm in (22) comparing the performance obtained by four different sampling strategies, namely: (a) the Max-Det strategy (obtained setting $f(\cdot)$ as the logarithm of the pseudo-determinant in Table 2); (b) the Max-$\lambda_{\min}$ strategy (obtained setting $f(\cdot) = \lambda_{\min}(\cdot)$ in Table 2); (c) the random sampling strategy, which simply picks at random $|\overline{\mathcal{S}}|$ nodes; and (d) the exhaustive search procedure aimed at minimizing the MSD in (45) over all the possible sampling combinations. In general, the latter strategy cannot be performed for large graphs and/or in a distributed fashion, and is reported only as a benchmark. We consider a signal bandwidth equal to $|\mathcal{F}| = 5$, the sampling probabilities are set equal to $p_i = 0.8$ for all $i \in \overline{\mathcal{S}}$, and the results are averaged over 500 independent simulations. The step-sizes and the combination weights are chosen as before. As we can notice from Fig. 8, the algorithm in (22) with random sampling can perform quite poorly, especially at low number of sampling nodes. Comparing the other sampling strategies, we notice from Fig. 8 that the Max-Det strategy outperforms all the others, giving good performance also at low number of sampling nodes ($|\overline{\mathcal{S}}| = 5$ is the minimum number of nodes that allows signal reconstruction). Interestingly, even if the proposed Max-Det strategy is a greedy approach, it shows performance that are comparable to the exhaustive search procedure, which represents the best possible performance achievable by a sampling strategy in terms of MSD. As previously mentioned, this

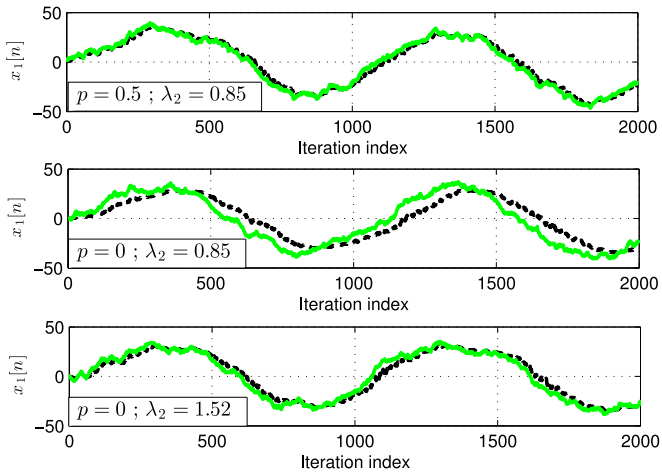


Fig. 9. Tracking behavior: Graph signal estimate (dashed) and true signal (solid) versus iteration index, for different values of sampling probability $p$ and graph algebraic connectivity $\lambda_2$.

good behavior is due to the monotonicity and sub-modularity properties of the objective function used in the Max-Det strategy, which ensures that the greedy selection strategy in Table 2 achieves performance that are very close to the optimal combinatorial solution [57], [58]. Finally, comparing the Max-$\lambda_{\min}$ strategy with the Max-Det strategy, we notice how the latter leads to better performance, because it considers all the modes of the matrix in (46), as opposed to the single mode associated to the minimum eigenvalue considered by the Max-$\lambda_{\min}$ strategy. This analysis suggests that an optimal design of the sampling strategy for graph signals should take into account processing complexity (in terms of number of sampling nodes), prior knowledge (e.g., graph structure, noise distribution), and achievable mean-square performance.

*4) Tracking of Time-Varying Graph Signals:* In this example, we illustrate the tracking capabilities of the proposed distributed methods in the presence of (slowly) time-varying signals evolving over the graph. To this aim, we generate a time-varying signal such that its graph Fourier transform (with respect to the graph in Fig. 1, having algebraic connectivity $\lambda_2 = 0.85$) evolves over time as: $\boldsymbol{s}^o[n+1] = \vartheta\, \boldsymbol{s}^o[n] + \boldsymbol{u}[n]$, where $\boldsymbol{s}^o[n] \in \mathbb{R}^{|\mathcal{F}|}$, $|\mathcal{F}| = 5$, $\vartheta = 0.99$, $\boldsymbol{u}[n] = \sin(2\pi f_o n)\mathbf{1} + \boldsymbol{w}[n]$, $f_o = 10^{-3}$, and $\boldsymbol{w}[n]$ is a zero-mean, Gaussian noise vector with identity covariance matrix. The corresponding graph signal at time $n$ is then obtained as $\boldsymbol{x}^o[n] = \mathbf{U}_{\mathcal{F}} \boldsymbol{s}^o[n]$. Thus, in Fig. 9 (top), we report the behavior of the estimate of the graph signal $x_i[n]$ in (22), for $i = 1$, using a dashed line. We also report the behavior of the true signal $x_i^o[n]$, using a solid line. The expected sampling set is composed of 10 nodes, and is selected according to the Max-Det sampling strategy; the sampling probabilities are set equal to $p_i = 0.5$ for all $i \in \overline{\mathcal{S}}$. In Fig. 9 (middle) we repeat the same experiment but setting the sampling probability of node 1 equal to $p_i = 0$, i.e., the node never observes the signal. Finally, in Fig. 9 (bottom), we consider the case in which the sampling probability of node 1 is equal to zero, but the connectivity of the communication graph linking the nodes is larger than before, having now an algebraic connectivity $\lambda_2 = 1.52$. The step-sizes are chosen equal to $\mu_i = 1$ for all $i$; the combination weights are selected as before. As we can notice from Fig. 9, the algorithm shows good tracking performance in all cases. As expected, the tracking capability is good in the case of Fig. 9 (top), when node 1 belongs to the expected sampling set and observes the signal for half of the time. Remarkably, also in the cases of Fig. 9 (middle) and (bottom), even if node 1 does not directly observe the signal at its location (i.e., $p = 0$), the algorithm can still guarantee good tracking performance thanks to the real-time diffusion of information among nodes in the graph. Finally, comparing Fig. 9 (middle) and (bottom), we can notice how a larger connectivity of the communication graph boosts the tracking capabilities of the network thanks to the faster information sharing among the nodes.

*5) Application Example - Power Spatial Density Estimation in Cognitive Networks:* In this example, we apply the proposed distributed framework to power density cartography in cognitive radio (CR) networks. We consider a 5G scenario, where a dense deployment of radio access points (RAPs) is envisioned to provide a service environment characterized by very low latency and high rate access. Each RAP collects data related to the transmissions of primary users (PUs) at its geographical position, and communicates with other RAPs with the aim of implementing advanced cooperative sensing techniques. The aim of the CR network is then to build a map of power spatial density (PSD) transmitted by PUs, while processing the received data on the fly and envisaging proper sampling techniques that enable a proactive sensing of the environment from only a limited number of RAP's measurements.

Let us then consider an operating region of 200 $\text{m}^2$ where 150 RAPs are randomly deployed to produce a map of the spatial distribution of power generated by the transmissions of four active primary users. The PU's emit electromagnetic radiation with power equal to 10 mW. The propagation medium is supposed to introduce a free-space path loss attenuation on the PU's transmissions. The graph among RAPs is built from a distance based model, i.e., stations that are sufficiently close to each other are connected through a link. In Fig. 10, we illustrate a pictorial description of the scenario, and of the resulting graph signal. For simplicity, we use the graph illustrated in Fig. 10 for both communication and processing tasks. We assume that each RAP is equipped with an energy detector, which estimates the received signal using 100 samples, considering an additive white Gaussian noise with variance $\sigma_v^2 = 10^{-4}$. The resulting signal is not perfectly bandlimited, but it turns out to be smooth over the graph, i.e., neighbor nodes observe similar values. This implies that sampling such signals inevitably introduces aliasing during the reconstruction process. However, even if we cannot find a limited (lower than $N$) set of frequencies where the signal is completely localized, the greatest part of the signal energy is concentrated at low frequencies. This means that if we process the data using a sufficient number of observations and (low) frequencies, we should still be able to reconstruct the signal with a satisfactory performance.

An example of PSD cartography based on the proposed diffusion algorithm is shown in Fig. 11, where we simulate a dynamic situation where the four PU's switch between idle and active modes in the order shown in Fig. 10 every $10^4$ time instants.

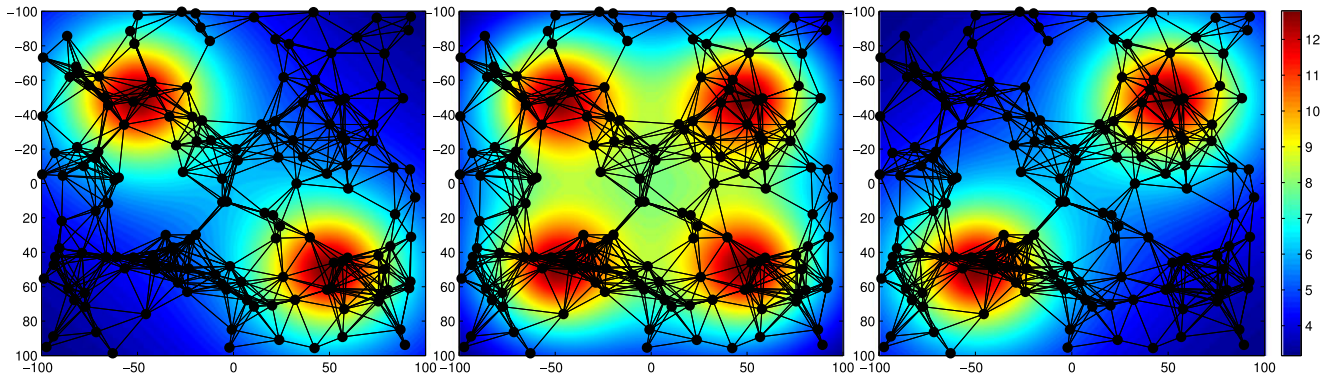

Fig. 10. PSD estimation in Cognitive Networks: PSD at different time instants, and topology of the cognitive network.

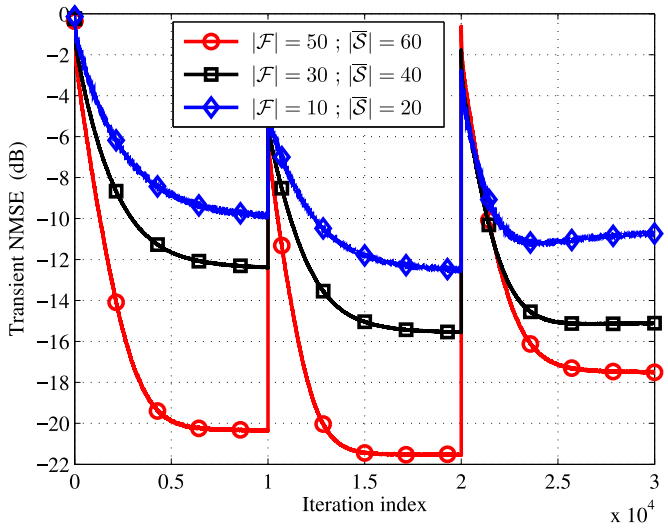


Fig. 11. PSD estimation in Cognitive Networks: Transient normalized MSD for different values of $|\mathcal{F}|$ and $|\overline{\mathcal{S}}|$.

In particular, in Fig. 11, we show the behavior of the transient normalized MSD, for different values of $|\overline{\mathcal{S}}|$ and bandwidths used for processing. The step-size is chosen equal to 1, the sampling probabilities are $p_i = 0.5$ for all $i$, while the adopted sampling strategy is the Max-Det strategy proposed in Table 2. From Fig. 11, we can see how the proposed technique can track time-varying scenarios. Furthermore, as expected, its steady-state performance and learning rate improve with increase in the number of nodes collecting samples and bandwidths used for processing.

## VII. Conclusions

In this paper, we have proposed distributed strategies for adaptive learning of graph signals. The method hinges on the structure of the underlying graph to process data and, under a bandlimited assumption, enables adaptive reconstruction and tracking from a limited number of observations taken over a subset of vertices in a totally distributed fashion. An interesting feature of our proposed method is that the sampling set is allowed to vary over time, and the convergence properties depend only on the expected set of sampling nodes. Furthermore, the graph topology plays an important role both in the processing and communication aspect of the problem. A detailed mean square analysis is also provided, thus illustrating the role of the sampling strategy on the reconstruction capability, stability, and mean-square performance of the proposed algorithm. Based on this analysis, some useful strategies for the distributed selection of the (expected) sampling set are also provided. Finally, several numerical results are reported to validate the theoretical findings, and illustrate the performance of the proposed method for distributed adaptive learning of signals defined over graphs.

This paper represents the first work that merges the well established field of adaptation and learning over networks, and the emerging topic of signal processing over graphs. Several interesting problems are still open, e.g., distributed reconstruction in the presence of directed and/or switching graph topologies, online identification of the graph signal support from streaming data, distributed inference of the (possibly unknown) graph signal topology, adaptation of the sampling strategy to time-varying scenarios, optimization of the sampling probabilities, just to name a few. We plan to investigate on these exciting problems in our future works.

## Appendix
## Stability of matrix $\mathbf{B}$ in (42)

Taking the expectation of both sides of (29), and exploiting Assumption 3, we conclude that the mean-error vector evolves according to the following dynamics:

$$\mathbb{E}\boldsymbol{e}[n+1] = \widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q})\mathbb{E}\boldsymbol{e}[n] = \mathbf{B}\,\mathbb{E}\boldsymbol{e}[n]. \tag{47}$$

To prove stability of matrix $\mathbf{B}$ in (42) (and, consequently, the mean stability of the algorithm in (22)), we proceed by showing that the sequence $\boldsymbol{e}[n]$ in (47) asymptotically vanishes for any initial condition. To this aim, let $\boldsymbol{y}[n] = \mathbb{E}\boldsymbol{e}[n]$, and consider its decomposition as:

$$\boldsymbol{y}[n] = \overline{\boldsymbol{y}}[n] + \widetilde{\boldsymbol{y}}[n], \tag{48}$$

where $\overline{\boldsymbol{y}}[n]$ represents the average vector over all nodes, and $\widetilde{\boldsymbol{y}}[n]$ is a disagreement error, respectively given by:

$$\overline{\boldsymbol{y}}[n] = \mathbf{J}\,\boldsymbol{y}[n] = (\mathbf{1} \otimes \mathbf{I})\,\hat{\boldsymbol{y}}[n], \tag{49}$$

$$\widetilde{\boldsymbol{y}}[n] = \mathbf{J}_{\perp}\,\boldsymbol{y}[n], \tag{50}$$

with

$$\hat{\boldsymbol{y}}[n] = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_i[n], \tag{51}$$

$$\mathbf{J} = \frac{1}{N}\,\mathbf{1}\mathbf{1}^T \otimes \mathbf{I}, \quad \text{and} \quad \mathbf{J}_\perp = \mathbf{I} - \mathbf{J}. \tag{52}$$

In the sequel, we will show that both $\overline{\boldsymbol{y}}[n]$ (or, equivalently, $\hat{\boldsymbol{y}}[n]$) and $\widetilde{\boldsymbol{y}}[n]$ asymptotically converge to zero, thus proving the convergence in the mean of the algorithm and the stability of matrix $\mathbf{B}$ in (42). From (49) and (47), we obtain

$$\begin{aligned}\overline{\boldsymbol{y}}[n+1] \;&=\; \mathbf{J}\widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q})\boldsymbol{y}[n] \\ &\overset{(a)}{=}\; \mathbf{J}\boldsymbol{y}[n] - \mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\boldsymbol{y}[n] \\ &\overset{(b)}{=}\; (\mathbf{I} - \mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q})\overline{\boldsymbol{y}}[n] - \mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\,\widetilde{\boldsymbol{y}}[n]\end{aligned} \tag{53}$$

where in (a) we have used $\mathbf{J}\widehat{\mathbf{W}} = \mathbf{J}$ [cf. (21), (25) and (52)]; and in (b) we have exploited (49) and (48). Similarly, from (50) and (47), we get

$$\begin{aligned}\widetilde{\boldsymbol{y}}[n+1] \;&=\; \mathbf{J}_\perp\widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q})\boldsymbol{y}[n] \\ &\overset{(a)}{=}\; \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{J}_\perp\boldsymbol{y}[n] - \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\,\boldsymbol{y}[n] \\ &\overset{(b)}{=} \mathbf{J}_\perp\widehat{\mathbf{W}}(\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q})\widetilde{\boldsymbol{y}}[n] - \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\,\overline{\boldsymbol{y}}[n]\end{aligned} \tag{54}$$

where in (a) we have exploited the relation $\mathbf{J}_\perp\widehat{\mathbf{W}} = \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{J}_\perp$ [cf. (21), (25) and (52)]; and in (b) we have used (50) and (48). Now, combining the recursions (53) and (54), we obtain

$$\begin{bmatrix}\overline{\boldsymbol{y}}[n+1] \\ \widetilde{\boldsymbol{y}}[n+1]\end{bmatrix} = \begin{pmatrix}\mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22}\end{pmatrix}\begin{bmatrix}\overline{\boldsymbol{y}}[n] \\ \widetilde{\boldsymbol{y}}[n]\end{bmatrix}, \tag{55}$$

where

$$\mathbf{Z}_{11} \;=\; \mathbf{I} - \mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}, \tag{56}$$

$$\mathbf{Z}_{12} \;=\; -\mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}, \tag{57}$$

$$\mathbf{Z}_{21} \;=\; -\mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}, \tag{58}$$

$$\mathbf{Z}_{22} \;=\; \mathbf{J}_\perp\widehat{\mathbf{W}}\left(\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\right). \tag{59}$$

A necessary and sufficient condition that guarantee the convergence to zero of the sequence in (55) is that matrix

$$\mathbf{Z} = \begin{pmatrix}\mathbf{Z}_{11} & \mathbf{Z}_{12} \\ \mathbf{Z}_{21} & \mathbf{Z}_{22}\end{pmatrix} \tag{60}$$

is stable [53]. We proceed by showing that, under Assumption 4, the eigenvalues of matrix $\mathbf{Z}$ in (60) are approximatively determined only by the eigenvalues of $\mathbf{Z}_{11}$ and $\mathbf{Z}_{22}$. From (60), the characteristic polynomial of $\mathbf{Z}$ is given by:

$$\begin{aligned}p(\lambda) &= \det\left(\mathbf{Z} - \lambda\mathbf{I}\right) \\ &\overset{(a)}{=}\; \det\left((\mathbf{Z}_{22} - \lambda\mathbf{I})(\mathbf{Z}_{11} - \lambda\mathbf{I}) - \mathbf{Z}_{21}\mathbf{Z}_{12}\right) \\ &\overset{(b)}{\simeq}\; \det(\mathbf{Z}_{22} - \lambda\mathbf{I})\det(\mathbf{Z}_{11} - \lambda\mathbf{I})\end{aligned} \tag{61}$$

where $(a)$ holds for $2\times 2$ block matrices [59, p. 4], since $\mathbf{Z}_{11} - \lambda\mathbf{I}$ and $\mathbf{Z}_{12}$ commute [cf. (56) and (57)]; and $(b)$ follows from the small-step size Assumption 4, as proved next. Indeed, expanding the argument of the determinant in (61a) we obtain:

$$\begin{aligned}&(\mathbf{Z}_{22} - \lambda\mathbf{I})(\mathbf{Z}_{11} - \lambda\mathbf{I}) - \mathbf{Z}_{21}\mathbf{Z}_{12} \\ &\qquad = \mathbf{Z}_{22}\mathbf{Z}_{11} - \mathbf{Z}_{21}\mathbf{Z}_{12} - (\mathbf{Z}_{22} + \mathbf{Z}_{11})\lambda + \lambda^2\mathbf{I}.\end{aligned} \tag{62}$$

Thus, if under Assumption 4 we have

$$\mathbf{Z}_{22}\mathbf{Z}_{11} - \mathbf{Z}_{21}\mathbf{Z}_{12} \approx \mathbf{Z}_{22}\mathbf{Z}_{11}, \tag{63}$$

from (62) and (61a), we can conclude that (61b) holds, i.e.,

$$(\mathbf{Z}_{22} - \lambda\mathbf{I})(\mathbf{Z}_{11} - \lambda\mathbf{I}) - \mathbf{Z}_{21}\mathbf{Z}_{12} \approx (\mathbf{Z}_{22} - \lambda\mathbf{I})(\mathbf{Z}_{11} - \lambda\mathbf{I}).$$

Now, from (56)–(59), we easily obtain:

$$\begin{aligned}\mathbf{Z}_{22}\mathbf{Z}_{11} &= \mathbf{J}_\perp\widehat{\mathbf{W}} - \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q} - \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q} \\ &\quad + \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q} \\ \mathbf{Z}_{21}\mathbf{Z}_{12} &= \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q} \\ \mathbf{Z}_{22}\mathbf{Z}_{11} - \mathbf{Z}_{21}\mathbf{Z}_{12} &= \mathbf{J}_\perp\widehat{\mathbf{W}} - \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q} - \mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}.\end{aligned} \tag{64}$$

It is then clear that, using Assumption 4 and thus neglecting the term $\mathbf{J}_\perp\widehat{\mathbf{W}}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\mathbf{J}\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q} = O(\mu_{\max}^2)$ in (64) with respect to the constant term and the term $O(\mu_{\max})$ contained in the expression of $\mathbf{Z}_{22}\mathbf{Z}_{11}$, we obtain (63). As previously mentioned, this proves that the approximation made in (61b) holds under the small step-sizes Assumption 4.

From (61), we conclude that, for sufficiently small step-sizes, the eigenvalues of matrix $\mathbf{Z}$ in (60) are approximatively given by the eigenvalues of $\mathbf{Z}_{11}$ and $\mathbf{Z}_{22}$ in (56) and (59), i.e. matrix $\mathbf{Z}$ is stable if matrices $\mathbf{Z}_{11}$ and $\mathbf{Z}_{22}$ are also stable. This means that the iteration matrix in (55) can be considered approximatively diagonal for the purpose of stability analysis. Thus, in the sequel, we analyze the stability of the recursion in (55), considering separately the behavior of the mean vector $\overline{\boldsymbol{y}}[n]$ and of the fluctuation $\widetilde{\boldsymbol{y}}[n]$, under the aforementioned diagonal approximation.

*Convergence of $\overline{\boldsymbol{y}}[n]$:* We now study the recursion

$$\overline{\boldsymbol{y}}[n+1] = \mathbf{Z}_{11}\,\overline{\boldsymbol{y}}[n].$$

For convenience, exploiting (49), (56), and (52), we equivalently recast the previous recursion in terms of $\hat{\boldsymbol{y}}[n]$, as:

$$\hat{\boldsymbol{y}}[n+1] = \left(\mathbf{I} - \frac{1}{N}\left(\mathbf{1}^T \otimes \mathbf{I}\right)\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\left(\mathbf{1}\otimes\mathbf{I}\right)\right)\hat{\boldsymbol{y}}[n]. \tag{65}$$

The recursion (65) converges to zero if the two following conditions hold: (a) matrix

$$\mathbf{V} = \frac{1}{N}\left(\mathbf{1}^T \otimes \mathbf{I}\right)\mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\left(\mathbf{1}\otimes\mathbf{I}\right) = \frac{1}{N}\sum_{i\in\overline{\mathcal{S}}}\mu_i p_i \boldsymbol{c}_i\boldsymbol{c}_i^H \tag{66}$$

is invertible (i.e., full rank); (b) and $|1 - \lambda_{\max}(\mathbf{V})| < 1$. Proceeding as in (10)–(12), the invertibility of matrix (66) is guaranteed under condition (12). Then, if matrix (66) is full rank, exploiting the inequality

$$\lambda_{\max}(\mathbf{V}) \;=\; \frac{1}{N}\left\|\sum_{i\in\overline{\mathcal{S}}}\mu_i p_i \boldsymbol{c}_i\boldsymbol{c}_i^H\right\| \;\le\; \frac{\mu_{\max}}{N}\sum_{i\in\overline{\mathcal{S}}} p_i\|\boldsymbol{c}_i\|^2,$$

condition (b) is guaranteed if the step-sizes satisfy:

$$0 < \mu_i \leq \mu_{\max} < \frac{2}{\frac{1}{N}\sum_{i\in\overline{\mathcal{S}}} p_i \|\boldsymbol{c}_i\|^2}, \quad \text{for all } i \in \overline{\mathcal{S}}, \text{ (67)}$$

which hold true under Assumption 4. Thus, under conditions (12) and assumption 4, $\hat{\boldsymbol{y}}[n]$ (and $\overline{\boldsymbol{y}}[n]$) converges to zero for all initial conditions, i.e., matrix $\mathbf{Z}_{11}$ is stable.

*Convergence of* $\widetilde{\boldsymbol{y}}[n]$*:* We now study the recursion

$$\widetilde{\boldsymbol{y}}[n+1] = \mathbf{Z}_{22}\, \widetilde{\boldsymbol{y}}[n],$$

which converges to zero if $\mathbf{Z}_{22}$ is stable. From (59), we have

$$\rho(\mathbf{Z}_{22}) \leq \|\mathbf{J}_\perp \widehat{\mathbf{W}}\|\, \|\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\|, \quad (68)$$

with $\rho(\mathbf{X})$ denoting the spectral radius of a matrix $\mathbf{X}$. Under Assumption 2, we have [cf. (25) and (52)]

$$\left\|\mathbf{J}_\perp \widehat{\mathbf{W}}\right\| = \left\|\left(\mathbf{W} - \frac{1}{N}\mathbf{1}\mathbf{1}^T\right) \otimes \mathbf{I}\right\| < 1. \quad (69)$$

Thus, from (68) and (69), $\rho(\mathbf{Z}_{22}) < 1$, i.e., matrix $\mathbf{Z}_{22}$ in (59) is stable, if $\|\mathbf{I} - \mathbf{M}\widehat{\mathbf{P}}\mathbf{Q}\| \leq 1$, which holds true under Assumption 4. In conclusion, matrix $\mathbf{Z}$ in (60) is stable, and the sequence $\boldsymbol{y}[n]$ in (48) [i.e., $\mathbb{E}\boldsymbol{e}[n]$ in (47)] asymptotically vanishes for all possible initial conditions. This proves the stability of matrix $\mathbf{B}$ in (42).

## ACKNOWLEDGMENT

The authors would like thank the anonymous reviewers for the detailed suggestions that improved the manuscript.

## 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, "Discrete signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[3] 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, Sep. 2014.

[4] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014.

[5] 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, Jun. 2012.

[6] S. K. Narang and A. Ortega, "Compact support biorthogonal wavelet filterbanks for arbitrary undirected graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4673–4685, Oct. 2013.

[7] I. Z. Pesenson, "Sampling in Paley-Wiener spaces on combinatorial graphs," *Trans. Amer. Math. Soc.*, vol. 360, no. 10, pp. 5603–5627, 2008.

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

[9] 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, Dec. 2015.

[10] 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. 3572–3585, Aug. 2008.

[11] M. Püschel and J. M. F. Moura, "Algebraic signal processing theory: 1-D space," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3586–3599, Aug. 2008.

[12] S. Narang, A. Gadde, and A. Ortega, "Signal processing techniques for interpolation in graph structured data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 5445–5449.

[13] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo, "Signals on graphs: Uncertainty principle and sampling," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4845–4860, Sep. 2016.

[14] X. Wang, P. Liu, and Y. Gu, "Local-set-based graph signal reconstruction," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2432–2444, May 2015.

[15] A. G. Marquez, S. Segarra, G. Leus, and A. Ribeiro, "Sampling of graph signals with successive local aggregations," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1832–1843, Apr. 2016.

[16] M. Tsitsvero and S. Barbarossa, "On the degrees of freedom of signals on graphs," in *Proc. Eur. Signal Process. Conf.*, Nice, Sep. 2015, pp. 1521–1525.

[17] S. K. Narang, A. Gadde, E. Sanou, and A. Ortega, "Localized iterative methods for interpolation in graph structured data," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, Dec. 2013, pp. 491–494.

[18] S. Segarra, A. G. Marques, G. Leus, and A. Ribeiro, "Reconstruction of graph signals through percolation from seeding nodes," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4363–4378, Aug. 2016.

[19] A. Sandryhaila and J. M. Moura, "Classification via regularization on graphs." in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, Dec. 2013, pp. 495–498.

[20] D. Thanou, D. I. Shuman, and P. Frossard, "Parametric dictionary learning for graph signals," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, Dec. 2013, pp. 487–490.

[21] D. Zhou and B. Schölkopf, "A regularization framework for learning from graph data," in *Proc. ICML Workshop Statist. Relational Learn. Connections Fields*, Jul. 2004, vol. 15, pp. 67–68.

[22] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[23] S. Chen, A. Sandryhaila, J. M. Moura, and J. Kovacevic, "Signal recovery on graphs: Variation minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4609–4624, Sep. 2015.

[24] S. Chen, R. Varma, A. Singh, and J. Kovačević, "Signal recovery on graphs: Fundamental limits of sampling strategies," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 539–554, Dec. 2016.

[25] S. Chen *et al.*, "Signal inpainting on graphs via total variation minimization," in *Proc. IEEE Conf. Acoust., Speech, Signal Process.*, Florence, May 2014, pp. 8267–8271.

[26] S. Chen, A. Sandryhaila, J. M. Moura, and J. Kovacevic, "Signal denoising on graphs via graph filtering," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Atlanta, GA, USA, Dec. 2014, pp. 872–876.

[27] P. Di Lorenzo, S. Barbarossa, P. Banelli, and S. Sardellitti, "Adaptive least mean squares estimation of graph signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 555–568, Dec. 2016.

[28] S. Chen, A. Sandryhaila, and J. Kovacevic, "Distributed algorithm for graph signal inpainting," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, QLD, Australia, Mar. 2015, pp. 3731–3735.

[29] D. Thanou and P. Frossard, "Distributed signal processing with graph spectral dictionaries," in *Proc. Allerton Conf. Commun., Control, Comput.*, Monticello, Sep. 2015, pp. 1391–1398.

[30] X. Wang, M. Wang, and Y. Gu, "A distributed tracking algorithm for reconstruction of graph signals," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 728–740, Jun. 2015.

[31] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[32] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[33] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4795–4810, Sep. 2010.

[34] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[35] P. Di Lorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, Mar. 2013.

[36] A. Sayed, "Adaptation, learning, and optimization over networks," *Found. Trends Mach. Learn.*, vol. 7, no. 4/5, pp. 311–801, 2014.

[37] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.
[38] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.
[39] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks with common latent representations," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 563–579, Apr. 2017.
[40] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: American Mathematical Society, 1997.
[41] A. Gadde, A. Anis, and A. Ortega, "Active semi-supervised learning using sampling theory for graph signals," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2014, pp. 492–501.
[42] D. Kempe and F. McSherry, "A decentralized algorithm for spectral analysis," in *Proc. ACM Symp. Theory Comput.*, Chicago, Jun. 2004, pp. 561–568.
[43] A. Bertrand and M. Moonen, "Seeing the bigger picture: How nodes can learn their place within a complex ad hoc network topology," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 71–82, May 2013.
[44] P. Di Lorenzo and S. Barbarossa, "Distributed estimation and control of algebraic connectivity over random graphs," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5615–5628, Nov. 2014.
[45] P. Di Lorenzo, "Diffusion adaptation strategies for distributed estimation over Gaussian Markov random fields," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5748–5760, Nov. 2014.
[46] J. Fernández-Bes, J. A. Azpicueta-Ruiz, M. T. Silva, and J. Arenas-García, "A novel scheme for diffusion networks with least-squares adaptive combiners," in *Proc. 2012 IEEE Int. Workshop Mach. Learn. Signal Process.*, Santander, Sep. 2012, pp. 1–6.
[47] C. G. Lopes, L. F. Chamon, and V. H. Nascimento, "Towards spatially universal adaptive diffusion networks," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Atlanta, GA, USA, Dec. 2014, pp. 803–807.
[48] P. Di Lorenzo, S. Barbarossa, and A. H. Sayed, "Bio-inspired decentralized radio access based on swarming mechanisms over adaptive networks," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3183–3197, Jun. 2013.
[49] P. Di Lorenzo, S. Barbarossa, and A. H. Sayed, "Distributed spectrum estimation for small cell networks based on sparse diffusion adaptation," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1261–1265, Dec. 2013.
[50] S. Chouvardas, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4692–4707, Oct. 2011.
[51] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Distributed detection and estimation in wireless sensor networks," *Signal Process.*, vol. 2, pp. 329–408, 2014.
[52] A. H. Sayed, *Adaptive Filters*. Hoboken, NJ, USA: Wiley, 2011.
[53] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[54] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *J. Parallel Distrib. Comput.*, vol. 67, no. 1, pp. 33–46, 2007.
[55] S. P. Chepuri and G. Leus, "Subsampling for graph power spectrum estimation," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, Rio de Janeiro, Brazil, Jul. 2016.
[56] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1520–1533, Sep. 2004.
[57] S. P. Chepuri and G. Leus, "Sparsity-promoting sensor selection for nonlinear measurement models," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 684–698, Feb. 2015.
[58] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *Proc. IEEE Conf. Decis. Control*, Atlanta, GA, USA, Dec. 2010, pp. 2572–2577.
[59] J. R. Silvester, "Determinants of block matrices," *Math. Gazette*, vol. 84, no. 501, pp. 460–467, 2000.

**Paolo Di Lorenzo** (S'10–M'13) received the M.Sc. degree in 2008 and the Ph.D. degree in electrical engineering in 2012, both from University of Rome "Sapienza," Italy. He is an Assistant Professor in the Department of Engineering, University of Perugia, Italy. He has participated in the European research projects FREEDOM, SIMTISYS, and TROPIC. His current research interests are in signal processing theory and methods, distributed optimization, adaptation and learning over networks, and graph signal processing. He is an Associate Editor of the *Eurasip Journal on Advances in Signal Processing*. He received three best student paper awards at IEEE SPAWC'10, EURASIP EUSIPCO'11, and IEEE CAMSAP'11, respectively. He is also recipient of the 2012 GTTI (Italian national group on telecommunications and information theory) award for the Best Ph.D. Thesis in information technologies and communications.

**Paolo Banelli** (S'90–M'99) received the Laurea degree (*cum laude*) in electronics engineering and the Ph.D. degree in telecommunications from the University of Perugia, Italy, in 1993 and 1998, respectively. In 2005, he was appointed as Associate Professor in the Department of Electronic and Information Engineering, University of Perugia, where he has been an Assistant Professor since 1998. His research interests include signal processing for wireless communications, with emphasis on multicarrier transmissions, signal processing for biomedical applications, spectrum sensing for cognitive radio, waveform design for 5G communications, and recently graph signal processing. In 2009, he was a General Cochair of the IEEE International Symposium on Signal Processing Advances for Wireless Communications. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the *EURASIP Journal on Advances in Signal Processing*.

**Sergio Barbarossa** (S'84–M'88–F'12) received the M.Sc. and Ph.D. degrees in electrical engineering from Sapienza University of Rome, Italy, in 1984 and 1988, respectively. He is a Full Professor with the University of Rome "Sapienza." He received the 2010 EURASIP Technical Achievements Award and the 2000 and 2014 IEEE Best Paper Awards from the IEEE Signal Processing Society. He served as IEEE Distinguished Lecturer in 2012–2013. Prof. Barbarossa is an IEEE Fellow and an EURASIP Fellow. He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1998–2000 and 2004–2006) and the IEEE SIGNAL PROCESSING MAGAZINE. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS. He is the coauthor of papers that received the Best Student Paper Award at ICASSP 2006, SPAWC 2010, EUSIPCO 2011, and CAMSAP 2011. His current research interests include topological data analysis, signal processing over graphs, mobile-edge computing, and 5G networks.

**Stefania Sardellitti** (M'12) received the M.Sc. degree in electronic engineering from the University of Rome "Sapienza," Italy, in 1998 and the Ph.D. degree in electrical and information engineering from the University of Cassino, Italy, in 2005. Since 2005 she is an appointed professor of digital communications at the University of Cassino, Italy. She is a research assistant at the Department of Information, Electronics and Telecommunications, University of Rome, Sapienza, Italy. She received the 2014 IEEE Best Paper Award from the IEEE Signal Processing Society. She has participated in the European projects WINSOC, FREEDOM, and TROPIC. Her research interests are in the area of statistical signal processing, mobile cloud computing, femtocell networks and wireless sensor networks, with emphasis on distributed optimization.