

Learning over Multitask Graphs-Part I: Stability Analysis

Roula Nassif, Stefan Vlaski, Cédric Richard, Ali H Sayed

▶ **To cite this version:**

Roula Nassif, Stefan Vlaski, Cédric Richard, Ali H Sayed. Learning over Multitask Graphs-Part I: Stability Analysis. IEEE Open Journal of Signal Processing, 2020, 1, pp.28-45. 10.1109/OJSP.2020.2989038 . hal-03347210

HAL Id: hal-03347210
https://hal.science/hal-03347210

Submitted on 17 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IEEE Open Journal of Signal Processing

Received 15 November 2019; revised 12 April 2020; accepted 14 April 2020. Date of publication 21 April 2020; date of current version 19 May 2020. The review of this article was arranged by Associate Editor Selin Aviyente.

Digital Object Identifier 10.1109/OJSP.2020.2989038

Learning Over Multitask Graphs—Part I: Stability Analysis

ROULA NASSIF [1,2] (Member, IEEE), STEFAN VLASKI [1] (Member, IEEE), CÉDRIC RICHARD [3] (Senior Member, IEEE), AND ALI H. SAYED [1] (Fellow, IEEE)

[1] Institute of Electrical Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
[2] American University of Beirut, Beirut 1107 2020, Lebanon
[3] Université de Nice Sophia-Antipolis, 06100 Nice, France

CORRESPONDING AUTHOR: ROULA NASSIF (e-mail: roula.nassif@aub.edu.lb)

The work of A. H. Sayed was supported in part by NSF under Grants CCF-1524250 and ECCS-1407712 and in part by the Swiss National Science Foundation under Grant 205121-184999.

ABSTRACT This paper formulates a multitask optimization problem where agents in the network have individual objectives to meet, or individual parameter vectors to estimate, subject to a smoothness condition over the graph. The smoothness condition softens the transition in the tasks among adjacent nodes and allows incorporating information about the graph structure into the solution of the inference problem. A diffusion strategy is devised that responds to streaming data and employs stochastic approximations in place of actual gradient vectors, which are generally unavailable. The approach relies on minimizing a global cost consisting of the aggregate sum of individual costs regularized by a term that promotes smoothness. We show in this Part I of the work, under conditions on the step-size parameter, that the adaptive strategy induces a contraction mapping and leads to small estimation errors on the order of the small step-size. The results in the accompanying Part II will reveal explicitly the influence of the network topology and the regularization strength on the network performance and will provide insights into the design of effective multitask strategies for distributed inference over networks.

INDEX TERMS Multitask distributed inference, diffusion strategy, smoothness prior, graph Laplacian regularization, gradient noise, stability analysis.

I. INTRODUCTION

Distributed inference allows a collection of interconnected agents to perform parameter estimation tasks from streaming data by relying solely on local computations and interactions with immediate neighbors. Most prior literature focuses on single-task problems, where agents with separable objective functions need to agree on a common parameter vector corresponding to the minimizer of an aggregate sum of individual costs [2]–[11]. Many network applications require more complex models and flexible algorithms than single-task implementations since their agents may need to estimate and track multiple objectives simultaneously [12]–[23]. Networks of this kind are referred to as multitask networks. Although agents may generally have distinct though related tasks to perform, they may still be able to capitalize on inductive transfer between them to improve their performance.

Based on the type of prior information that may be available about how the tasks are related to each other, multitask learning algorithms can be derived by translating the prior information into constraints on the parameter vectors to be inferred [12]–[23]. For example, in [19]–[21], distributed strategies are developed under the assumption that the parameter vectors across the agents overlap partially. A more general scenario is considered in [22] where it is assumed that the tasks across the agents are locally coupled through linear equality constraints. In [23], the parameter space is decomposed into two orthogonal subspaces, with one of the subspaces being common to all agents. There is yet another useful way to model relationships among tasks, namely, to formulate optimization problems with appropriate regularization terms encoding these relationships [13]–[18]. For example, the strategy developed in [13] adds squared ℓ_2-norm co-regularizers to

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

the mean-square-error criterion to promote task similarities, while the strategy in [14] adds ℓ_1-norm co-regularizers to promote piece-wise constant transitions.

In this paper, and the accompanying Part II [24], we consider multitask inference problems where each agent in the network seeks to minimize an individual cost expressed as the expectation of some loss function. The minimizers of the individual costs are assumed to vary smoothly on the topology captured by the graph Laplacian matrix. The smoothness property softens the transition in the tasks among adjacent nodes and allows incorporating information about the graph structure into the solution of the inference problem. This smoothness property (under which the tasks are similar at neighboring vertices with the strength of similarity specified by the weight between them) is often observed in real world applications. For instance, in monitoring applications, agents in a network need to track multiple targets located on a smooth area [13]. In weather forecasting applications, the decision rules at neighboring weather stations are expected to be similar since they are collecting features arising from similar statistical distributions (see Section IV). Other examples include house price prediction and crime prediction applications (see [25] and the applications therein). In order to exploit the smoothness prior, we formulate in this work the inference problem as the minimization of the aggregate sum of individual costs regularized by a term promoting smoothness, known as the graph-Laplacian regularizer [26], [27]. A diffusion strategy is devised that responds to streaming data and employs stochastic approximations in place of actual gradient vectors, which are generally unavailable. We show in this Part I of the work, under conditions on the step-size learning parameter μ, that the adaptive strategy induces a contraction mapping and that despite gradient noise, it is able to converge in the mean-square-error sense within $O(\mu)$ from the solution of the regularized problem, for sufficiently small μ. The analysis in the current part also reveals how the regularization strength η can steer the convergence point of the network toward many modes starting from the non-cooperative mode where each agent converges to the minimizer of its individual cost and ending with the single-task mode where all agents converge to a common parameter vector corresponding to the minimizer of the aggregate sum of individual costs. We shall also derive in Part II [24] a closed-form expression for the steady-state network mean-square-error relative to the minimizer of the regularized cost. This closed form expression will reveal explicitly the influence of the regularization strength, network topology, gradient noise, and data characteristics, on the network performance. Additionally, a closed-form expression for the steady-state network mean-square-error relative to the minimizers of the individual costs will be also derived in Part II [24]. This expression will provide insights into the design of effective multitask strategies for distributed inference over networks.

There have been many works in the literature studying distributed multitask adaptive strategies and their convergence behavior. Nevertheless, with few exceptions [21], most of these works focus on mean-square-error costs. This paper, and the accompanying Part II [24], generalize distributed multitask inference over networks and applies it to a wide class of individual costs. Furthermore, previous works in this domain tend to show the benefit of multitask learning empirically by simulations. Following some careful and demanding analysis, we establish in Part II [24], which builds on the results of this Part I, a useful expression for the network steady-state performance. This expression provides insights into the learning behavior of multitask networks and clarifies how multitask distributed learning may improve the network performance. Unlike most of the previous works in the field where the performance analyses are conducted in the graph (or vertex) domain, this work focuses instead on analyzing the proposed strategy in the graph spectral domain. As a consequence, the resulting performance expressions reveal explicitly the influence of the eigenvalues and eigenvectors of the graph Laplacian on the network performance. Furthermore, as we will see in this paper, and the accompanying Part II [24], this analysis shows that the multitask algorithm behaves in some situations as a lowpass graph filter.

Notation: All vectors are column vectors. Random quantities are denoted in boldface. Matrices are denoted in capital letters while vectors and scalars are denoted in lower-case letters. The operator $\preceq$ denotes an element-wise inequality; i.e., $a \preceq b$ implies that each entry of the vector a is less than or equal to the corresponding entry of b. The symbol $\mathrm{diag}\{\cdot\}$ forms a matrix from block arguments by placing each block immediately below and to the right of its predecessor. The operator $\mathrm{col}\{\cdot\}$ stacks the column vector entries on top of each other. The symbol $\otimes$ denotes the Kronecker product. We define $x = O(\mu)$ to mean $\lim_{\mu\to 0} \frac{x}{\mu} = c$ for some constant c independent of μ, which is stronger than merely requiring $\lim_{\mu\to 0} \frac{x}{\mu} \leq c$. Table I lists a couple of variables and symbols that will be used throughout the paper.

II. DISTRIBUTED INFERENCE UNDER SMOOTHNESS PRIORS

A. PROBLEM FORMULATION AND ADAPTIVE STRATEGY

We refer to Fig. 1 and consider a connected network (or graph) $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, A\}$, where $\mathcal{N}$ is a set of N agents (nodes), $\mathcal{E}$ is a set of edges connecting agents with particular relations, and A is a symmetric, weighted adjacency matrix. If there is an edge connecting agents k and ℓ, then $[A]_{k\ell} = a_{k\ell} > 0$ reflects the strength of the relation between k and ℓ; otherwise, $[A]_{k\ell} = 0$. We introduce the graph Laplacian L, which is a differential operator defined as $L = D - A$, where the degree matrix D is a diagonal matrix with k-th entry $[D]_{kk} = \sum_{\ell=1}^{N} a_{k\ell}$. Since L is symmetric positive semi-definite, it possesses a complete set of orthonormal eigenvectors. We denote them by $\{v_1, \ldots, v_N\}$. For convenience, we order the set of real, non-negative eigenvalues of L as $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_N = \lambda_{\max}(L)$, where, since the network is connected, there is only one zero eigenvalue with corresponding eigenvector

TABLE 1. Definition of Some Variables Used Throughout the Paper

	Variable
The set of neighbors of node k	$\mathcal{N}_k$
Minimizer at node k	w_k^o
Network minimizer vector	$\mathcal{W}^o = \text{col}\{w_1^o, \ldots, w_N^o\}$
Network regularized solution	$\mathcal{W}_\eta^o$ in (4)
k-th subvector of $\mathcal{W}_\eta^o$	$w_{k,\eta}^o$
Graph Laplacian matrix	L
Extended Laplacian matrix	$\mathcal{L} = L \otimes I_M$
Eigendecomposition of L	V and Λ given by (1)
Extended versions of V and Λ	$\mathcal{V}$ and $\mathcal{J}$ in (22)
Transformed regularized solution	$\overline{\mathcal{W}}_\eta^o$ in (13)
m-th subvector of $\overline{\mathcal{W}}_\eta^o$	$\overline{w}_{m,\eta}^o$ in (14)
Transformed version of $\mathcal{W}^o$	$\overline{\mathcal{W}}^o$ in (21)
m-th subvector of $\overline{\mathcal{W}}^o$	$\overline{w}_m^o$
Gradient noise vector at node k	$\boldsymbol{s}_{k,i}(\cdot)$ in (6)

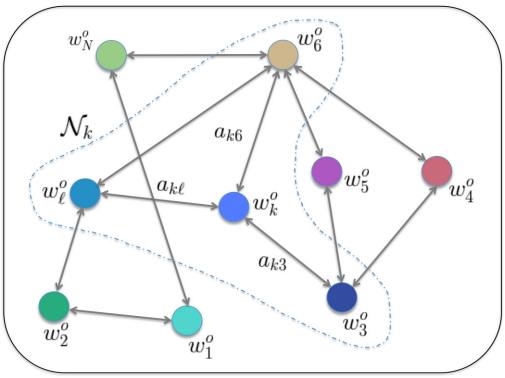


FIGURE 1. Agents linked by an edge can share information. The weight $a_{k\ell}$ over an edge reflects the strength of the relation between w_k^o at node k and w_ℓ^o at node ℓ.

$v_1 = \frac{1}{\sqrt{N}}\mathbb{1}_N$ [28]. Thus, the Laplacian can be decomposed as:

$$L = V\Lambda V^\top, \tag{1}$$

where $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_N\}$ and $V = [v_1, \ldots, v_N]$.

Let $w_k \in \mathbb{R}^M$ denote some parameter vector at agent k and let $\mathcal{W} = \text{col}\{w_1, \ldots, w_N\}$ denote the collection of parameter vectors from across the network. We associate with each agent k a risk function $J_k(w_k) : \mathbb{R}^M \to \mathbb{R}$ assumed to be strongly convex. In most learning and adaptation problems, the risk function is expressed as the expectation of a loss function $Q_k(\cdot)$ and is written as $J_k(w_k) = \mathbb{E}\, Q_k(w_k; \boldsymbol{x}_k)$, where $\boldsymbol{x}_k$ denotes the random data. The expectation is computed over the distribution of this data. We denote the unique minimizer of $J_k(w_k)$ by w_k^o. We introduce a common assumption on the risks $\{J_k(w_k)\}$. This condition is satisfied by many objective functions of interest in learning and adaptation such as quadratic and logistic risks. Besides, regularization is a common technique to ensure strong convexity (see, e.g., [7], [10]). It should be noted that regularized logistic regression risks are considered in the weather forecasting application in Section IV and that mean-square-error (quadratic) risks are considered in the simulation section of Part II [24].

Assumption 1 (Strong convexity): It is assumed that the individual costs $J_k(w_k)$ are each twice differentiable and strongly convex such that the Hessian matrix function $H_k(w_k) = \nabla_{w_k}^2 J_k(w_k)$ is uniformly bounded from below and above, say, as:

$$0 < \lambda_{k,\min} I_M \le H_k(w_k) \le \lambda_{k,\max} I_M, \tag{2}$$

where $\lambda_{k,\min} > 0$ for $k = 1, \ldots, N$. □

In many situations, there is prior information available about $\mathcal{W}^o = \text{col}\{w_1^o, \ldots, w_N^o\}$. In the current work, the prior belief we want to enforce is that the target signal $\mathcal{W}^o$ is smooth with respect to the underlying weighted graph. References [13]–[15], [18] provide variations for such problems for the special case of mean-square-error costs and linear data models. Here we treat general convex costs. Let $\mathcal{L} = L \otimes I_M$. The smoothness of $\mathcal{W}$ can be measured in terms of a quadratic form of the graph Laplacian [26], [27], [29]:

$$S(\mathcal{W}) = \mathcal{W}^\top \mathcal{L}\mathcal{W} = \frac{1}{2}\sum_{k=1}^{N}\sum_{\ell\in\mathcal{N}_k} a_{k\ell}\|w_k - w_\ell\|^2, \tag{3}$$

where $\mathcal{N}_k$ is the set of neighbors of k, i.e., the set of nodes connected to agent k by an edge. Fig. 1 provides an illustration. The smaller $S(\mathcal{W})$ is, the smoother the signal $\mathcal{W}$ on the graph is. Intuitively, given that the weights are non-negative, $S(\mathcal{W})$ shows that $\mathcal{W}$ is considered to be smooth if nodes with a large $a_{k\ell}$ on the edge connecting them have similar weight values $\{w_k, w_\ell\}$. Our objective is to devise and study a strategy that solves the following regularized problem:

$$\mathcal{W}_\eta^o = \arg\min_{\mathcal{W}} J^{\text{glob}}(\mathcal{W}) = \sum_{k=1}^{N} J_k(w_k) + \frac{\eta}{2}\mathcal{W}^\top\mathcal{L}\mathcal{W}, \tag{4}$$

in a distributed manner where each agent is interested in estimating the k-th sub-vector of $\mathcal{W}_\eta^o = \text{col}\{w_{1,\eta}^o, \ldots, w_{N,\eta}^o\}$. The tuning parameter $\eta \ge 0$ controls the trade-off between the two components of the objective function. Reference [1] provides a theoretical motivation for the optimization framework where it is shown that, under a Gaussian Markov random field assumption, solving problem (4) is equivalent to finding a maximum a posteriori (MAP) estimate for $\boldsymbol{\mathcal{W}}$. We are particularly interested in solving the problem in the stochastic setting when the distribution of the data $\boldsymbol{x}_k$ in $J_k(w_k) = \mathbb{E}\, Q_k(w_k; \boldsymbol{x}_k)$ is generally unknown. This means that the risks $J_k(w_k)$ and their gradients $\nabla_{w_k} J_k(w_k)$ are unknown. As such, approximate gradient vectors need to be employed. A common construction in stochastic approximation theory is to employ the following approximation at iteration i:

$$\widehat{\nabla_{w_k} J_k}(w_k) = \nabla_{w_k} Q_k(w_k; \boldsymbol{x}_{k,i}), \tag{5}$$

where $\boldsymbol{x}_{k,i}$ represents the data observed at iteration i. The difference between the true gradient and its approximation is called the gradient noise $\boldsymbol{s}_{k,i}(\cdot)$:

$$\boldsymbol{s}_{k,i}(w) \triangleq \nabla_{w_k} J_k(w) - \widehat{\nabla_{w_k} J_k}(w). \tag{6}$$

Each agent can employ a stochastic gradient descent update to estimate $w_{k,\eta}^o$:

$$\begin{aligned}\boldsymbol{w}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla_{w_k} J_k}(\boldsymbol{w}_{k,i-1}) \\ - \mu\eta \sum_{\ell\in\mathcal{N}_k} a_{k\ell}(\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{\ell,i-1}),\end{aligned} \tag{7}$$

where $\mu > 0$ is a small step-size parameter. In this implementation, each agent k collects from its neighbors the estimates $\boldsymbol{w}_{\ell,i-1}$, and performs a stochastic-gradient descent update on:

$$\bar{J}_{k,i-1}(w_k) \triangleq J_k(w_k) + \frac{\eta}{2}\sum_{\ell\in\mathcal{N}_k} a_{k\ell}\|w_k - \boldsymbol{w}_{\ell,i-1}\|^2. \tag{8}$$

By introducing an auxiliary variable $\boldsymbol{\psi}_{k,i}$, strategy (7) can be implemented in an incremental manner:

$$\begin{cases}\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla_{w_k} J_k}(\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} = \boldsymbol{\psi}_{k,i} - \mu\eta\sum_{\ell\in\mathcal{N}_k} a_{k\ell}(\boldsymbol{\psi}_{k,i} - \boldsymbol{\psi}_{\ell,i}),\end{cases} \tag{9}$$

where we replaced $(\boldsymbol{w}_{k,i-1} - \boldsymbol{w}_{\ell,i-1})$ in the second step by the difference $(\boldsymbol{\psi}_{k,i} - \boldsymbol{\psi}_{\ell,i})$ since we expect $\boldsymbol{\psi}_{k,i}$ to be an improved estimate compared to $\boldsymbol{w}_{k,i-1}$. Note that if we introduce the coefficients:

$$c_{k\ell} = \begin{cases}1 - \mu\eta\sum_{\ell\in\mathcal{N}_k} a_{k\ell}, & k = \ell \\ \mu\eta a_{k\ell}, & \ell \in \mathcal{N}_k \setminus \{k\} \\ s0, & \ell \notin \mathcal{N}_k\end{cases} \tag{10}$$

then recursion (9) can be written in the diffusion form [6]–[10]:

$$\begin{cases}\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\widehat{\nabla_{w_k} J_k}(\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} = \sum_{\ell\in\mathcal{N}_k} c_{k\ell}\boldsymbol{\psi}_{\ell,i},\end{cases} \tag{11}$$

where the second step is a combination step. If we collect the scalars $\{c_{k\ell}\}$ into the matrix $C = [c_{k\ell}]$, then the entries of C are non-negative for small enough μ and its columns and rows add up to one, i.e., C is a doubly-stochastic matrix. We shall continue with form (9) because the second step in (9) makes the dependence on η explicit. We will show later that by varying the value of η we can make the algorithm behave in different ways from fully non-cooperative to fully single-task with many other modes in between.

B. SUMMARY OF MAIN RESULTS

Before delving into the study of the learning capabilities of (9) and its performance limits, we summarize in this section, for the benefit of the reader, the main conclusions of this Part I, and its accompanying Part II [24]. One key observation is that the smoothing parameter η can be regarded as an effective tuning parameter that controls the nature of the learning process. The value of η can vary from $\eta = 0$ to $\eta \to \infty$. At one end, when $\eta = 0$, the learning algorithm reduces to a non-cooperative mode of operation where each agent acts individually and estimates its own local model, w_k^o. On the other hand, when $\eta \to \infty$, the learning algorithm moves to a single-mode of operation where all agents cooperate to estimate a *single* parameter (namely, the Pareto solution of the aggregate cost function). For any values of η in the range $0 < \eta < \infty$, the network behaves in a multitask mode where agents seek their individual models while at the same time ensuring that these models satisfy certain smoothness and closeness conditions dictated by the value of η. We are not only interested in a qualitative description of the network behavior. Instead, we would like to characterize these models in a quantitative manner by deriving expressions that allow us to predict performance as a function of η and, therefore, fine tune the network to operate in different scenarios.

To begin with, recall that the objective of the multitask strategy (9) is to exploit similarities among neighboring agents in an attempt to improve the overall network performance in approaching the collection of individual minimizer $\mathcal{W}^o$ by means of local communications. In light of the fact that algorithm (9) has been derived as an (incremental) gradient descent recursion for the regularized cost (4), whose minimizer $\mathcal{W}_\eta^o$ is in general different from $\mathcal{W}^o$, the limiting point of algorithm (9) will therefore be generally different from $\mathcal{W}^o$, the actual objective of the multitask learning problem. This mismatch is the "cost" of enforcing smoothness. The analysis in the paper will reveal that the mismatch is a function of the similarity between the individual minimizers $\{w_k^o\}$, of second-order properties of the individual costs, of the network topology captured by L, and of the regularization strength η. In particular, future expression (32) will allow us to understand the interplay between these quantities which is important for the design of effective multitask strategies. The key conclusion will be that, while the bias (difference between $\mathcal{W}_\eta^o$ and $\mathcal{W}^o$) will in general increase as the regularization strength η increases, the *size* of this increase is determined by the smoothness of $\mathcal{W}^o$ which is in turn function of the network topology captured by L. The more similar the tasks at neighboring agents are, the smaller the bias will be. This result, while intuitive, is reassuring, as it implies that as long as $\mathcal{W}^o$ is sufficiently smooth, the bias induced by regularization will remain small, even for moderate regularization strengths η.

The analysis in this work (this Part I, and its accompanying Part II [24]) also quantifies the *benefit* of cooperation, namely, the objective of improving the mean-square deviation around the limiting point of the algorithm. This analysis is challenging due to coupling among agents, and the multi-task nature of the learning process (where agents have individual targets but need to meet certain smoothness and closeness conditions with their neighbors). Section III in this Part I and Sections III and IV in Part II [24], and the supporting appendices, are devoted to carrying out this analysis in depth leading, for example to Theorem 1 in Part II [24]. This theorem gives expressions for the mean-square-deviation (MSD) relative to $\mathcal{W}_\eta^o$. The expressions reveal the effect of the step-size parameter μ, regularization strength η, network topology, and data characteristics (captured by the smoothness profile, second-order properties of the costs, and second-order moments of the gradient noise) on the size of the steady-state mean-square-error performance. The results established in Theorem 1 and

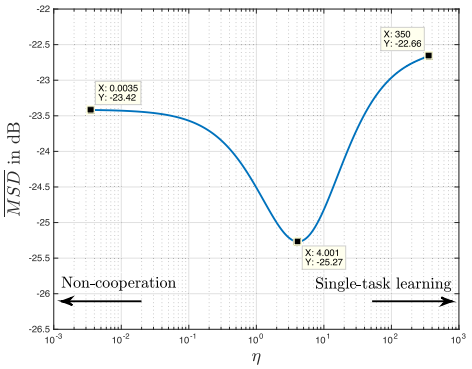


FIGURE 2. Network steady-state MSD relative to a smooth signal $\mathcal{W}^o$ as a function of the regularization strength $\eta \in [0, 350]$ at $\mu = 0.005$.

expression (82) in Part II [24] provide tools for characterizing the performance of multitask strategies in some great detail.

To illustrate the power of these results, consider a connected network where each agent is subjected to streaming data. The goal at each agent is to estimate a local parameter vector w_k^o from the observed data by minimizing a cost of the form $J_k(w_k) = \mathbb{E}\, Q_k(w_k; \boldsymbol{x}_k)$, where $\boldsymbol{x}_k$ denotes the random data. Consider network applications where the minimizers at neighboring agents tend to be similar [13], [27]. Although each agent is interested in estimating its own task w_k^o, cooperating neighboring agents can still benefit from their interactions because of this closeness. Given the graph Laplacian and data characteristics, one problem of interest would be to determine the optimal cooperation rule, i.e., the value of η that minimizes the network mean-square-error performance. Future expression (82) in Part II [24] can be used to solve this problem since it allows us to predict the network MSD relative to $\mathcal{W}^o$, namely, $\overline{\text{MSD}} = \limsup_{i\to\infty} \frac{1}{N}\sum_{k=1}^N \mathbb{E}\|w_k^o - \boldsymbol{w}_{k,i}\|^2$. By using expression (82) in Part II [24], for example, we will be able to construct curves of the form shown in Fig. 2, which illustrate how performance is dependent on the smoothness parameter η and how the nature of the limiting solution varies as a function of this parameter. As it can be seen from this figure, $\eta = 4$ gives the best network steady-state mean-square performance. Note that $\eta = 0$ corresponds to the non-cooperative scenario and that a large η induces a large bias in the estimation. In the sequel we will see that as η varies from $\eta = 0$ to $\eta \to \infty$, the network behavior moves from the non-cooperative mode of operation (where agents act independently) to the single-task mode of operation (where all agents focus on estimating a single parameter). For values of η in between, the network can operate in any of a multitude of multitask modes (where agents estimate their own local parameters under smoothness conditions to allow for some similarity between adjacent nodes). These limits are indicated in Fig. 2.

Finally, we would like to mention that one of the main tools used in the analysis in this work, and its accompanying Part II [24], is the linear transformation relative to the eigenspace of the graph Laplacian L from [6], which is also known as the graph Fourier transform [27], [30], [31]. Under some conditions on the data and costs profile, we show in Section VI-A in Part II [24] how the diffusion type algorithm (9) exhibits a low-pass graph filter behavior. Such filters are commonly used to reduce the network noise profile when the signal to be estimated is smooth with respect to the underlying topology [27], [32]–[34]. Interestingly, the theoretical results established in this Part I, and its accompanying Part II [24], reveal the reasons for performance improvements under localized cooperation.

C. NETWORK LIMIT POINT AND REGULARIZATION STRENGTH

Before examining the behavior and performance of strategy (9) with respect to the limiting point $\mathcal{W}_\eta^o$ in (4), we discuss the influence of η on $\mathcal{W}_\eta^o$. When $\eta = 0$, we have from (4) that $\mathcal{W}_\eta^o = \mathcal{W}^o$ and strategy (9) reduces to the single-agent mode of operation or the non-cooperative solution where each agent minimizes $J_k(w_k)$ locally without cooperation. When $\eta \to \infty$, we have from (4) that $\mathcal{W}_\eta^o = \mathbb{1}_N \otimes w^\star$ where

$$w^\star \triangleq \arg\min_w \sum_{k=1}^N J_k(w), \tag{12}$$

and we are in the single-task mode of operation where all agents seek to estimate a common parameter vector $w^\star$ corresponding to the minimizer of the aggregate sum of individual costs [6]–[10]. In order to study more closely the influence of (finite) $\eta > 0$ on the network output $\mathcal{W}_\eta^o$, we examine the influence of η on the transformed vector:

$$\overline{\mathcal{W}}_\eta^o \triangleq (V^\top \otimes I_M)\mathcal{W}_\eta^o = \text{col}\left\{\overline{w}_{m,\eta}^o\right\}_{m=1}^N, \tag{13}$$

with the m-th sub-vector $\overline{w}_{m,\eta}^o$ denoting the spectral content of $\mathcal{W}_\eta^o$ at the m-th eigenvalue λ_m of the Laplacian:

$$\overline{w}_{m,\eta}^o = (v_m^\top \otimes I_M)\overline{\mathcal{W}}_\eta^o. \tag{14}$$

From (1), the quadratic regularization term $S(\mathcal{W})$ in (3) can be written as:

$$S(\mathcal{W}) = \mathcal{W}^\top \mathcal{L}\mathcal{W} = \sum_{m=1}^N \lambda_m \|\overline{w}_m\|^2 = \sum_{m=2}^N \lambda_m \|\overline{w}_m\|^2, \tag{15}$$

where $\overline{w}_m = (v_m^\top \otimes I_M)\mathcal{W}$ and where we used the fact that $\lambda_1 = 0$. Intuitively, given that $\lambda_m > 0$ for $m = 2, \ldots, N$, the above expression shows that $\mathcal{W}$ is considered to be smooth if $\|\overline{w}_m\|^2$ corresponding to large λ_m is small. As a result, for a fixed $\lambda_m > 0$, and as the regularization strength $\eta > 0$ in (4) increases, one would expect $\|\overline{w}_{m,\eta}^o\|^2$ to decrease. Similarly, for a fixed $\eta \geq 0$, and as $\lambda_m > 0$ increases, one would expect $\|\overline{w}_{m,\eta}^o\|^2$ to decrease as well. However, as we will see in the sequel, this behavior does not always hold. We show in Section VI-A in Part II [24] that this is valid when the Hessian matrix function $H_k(w_k) \triangleq \nabla_{w_k}^2 J_k(w_k)$ is independent of w_k, i.e., the cost $J_k(w_k)$ is quadratic in w_k and is uniform

across the network. For more general scenarios, this is not necessarily the case. What is useful to note, however, is that as η moves from 0 towards ∞, a variety of solution points $\mathcal{W}_\eta^o$ can occur ranging from the non-cooperative to the single-task solution at both extremes.

From the optimality condition of (4), we have:

$$\text{col}\left\{\nabla_{w_k} J_k(w_{k,\eta}^o)\right\}_{k=1}^N = -\eta\mathcal{L}\mathcal{W}_\eta^o, \tag{16}$$

Using the mean value theorem [35, pp. 24], we can write:

$$\nabla_{w_k} J_k(w_{k,\eta}^o) - \underbrace{\nabla_{w_k} J_k(w_k^o)}_{=0} = H_{k,\eta}^o(w_{k,\eta}^o - w_k^o), \tag{17}$$

where

$$H_{k,\eta}^o \triangleq \int_0^1 \nabla_{w_k}^2 J_k(w_k^o + t(w_{k,\eta}^o - w_k^o))dt. \tag{18}$$

Let $\mathcal{H}_\eta^o \triangleq \text{diag}\{H_{k,\eta}^o\}_{k=1}^N$. Relation (16) can then be rewritten more compactly as:

$$\mathcal{W}_\eta^o = \left(\mathcal{H}_\eta^o + \eta\mathcal{L}\right)^{-1}\mathcal{H}_\eta^o\mathcal{W}^o. \tag{19}$$

Note that the inverse in (19) exists for all $\eta \geq 0$ since the matrix $\mathcal{L}$ is positive semi-definite and, under Assumption 1, the matrix $\mathcal{H}_\eta^o$ is positive definite. Pre-multiplying both sides of the above relation by $(V \otimes I_M)^\top$ gives:

$$\overline{\mathcal{W}}_\eta^o = (\overline{\mathcal{H}}_\eta^o + \eta\mathcal{J})^{-1}\overline{\mathcal{H}}_\eta^o\overline{\mathcal{W}}^o, \tag{20}$$

where $\overline{\mathcal{W}}_\eta^o$ is defined in (13) and

$$\overline{\mathcal{W}}^o \triangleq \mathcal{V}^\top\mathcal{W}^o, \tag{21}$$

$$\mathcal{V} \triangleq V \otimes I_M, \qquad \mathcal{J} \triangleq \Lambda \otimes I_M, \tag{22}$$

$$\overline{\mathcal{H}}_\eta^o \triangleq \mathcal{V}^\top\mathcal{H}_\eta^o\mathcal{V}. \tag{23}$$

Since L has a single eigenvalue at zero, Λ and V can be partitioned as follows:

$$\Lambda = \text{diag}\{0, \Lambda_o\}, V = [v_1, V_R], \text{ and } V^\top = \text{col}\{v_1^\top, V_R^\top\}. \tag{24}$$

Lemma 1: (Limiting point) Under Assumption 1, it can be shown that $\overline{\mathcal{W}}_\eta^o$ given by (20) satisfies:

$$\overline{\mathcal{W}}_\eta^o = \begin{bmatrix} I_M & \mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}\left(I_{M(N-1)} - \mathcal{K}\right) \\ 0 & \mathcal{K} \end{bmatrix}\begin{bmatrix} \overline{w}_1^o \\ [\overline{\mathcal{W}}^o]_{2:N} \end{bmatrix}, \tag{25}$$

where $\overline{w}_1^o = (v_1^\top \otimes I_M)\mathcal{W}^o$, $[\overline{\mathcal{W}}^o]_{2:N} = (V_R^\top \otimes I_M)\mathcal{W}^o$ and

$$\mathcal{Q}_{11} \triangleq (v_1^\top \otimes I_M)\mathcal{H}_\eta^o(v_1 \otimes I_M) = \frac{1}{N}\sum_{k=1}^N H_{k,\eta}^o, \tag{26}$$

$$\mathcal{Q}_{12} \triangleq (v_1^\top \otimes I_M)\mathcal{H}_\eta^o(V_R \otimes I_M), \tag{27}$$

$$\mathcal{Q}_{22} \triangleq (V_R^\top \otimes I_M)\mathcal{H}_\eta^o(V_R \otimes I_M) + \eta\Lambda_o \otimes I_M, \tag{28}$$

$$\mathcal{G} \triangleq (\mathcal{Q}_{22} - \mathcal{Q}_{12}^\top\mathcal{Q}_{11}^{-1}\mathcal{Q}_{12})^{-1}, \tag{29}$$

$$\mathcal{K} \triangleq I_{M(N-1)} - \eta\,\mathcal{G}\,(\Lambda_o \otimes I_M) \tag{30}$$

Proof: See Appendix A where we also show that:

$$\|\mathcal{K}\| \leq \left(\max_{1\leq k\leq N}\lambda_{k,\max}\right)\left(\eta\lambda_2(L) + \min_{1\leq k\leq N}\lambda_{k,\min}\right)^{-1} = \frac{O(1)}{(O(1) + O(\eta))}. \tag{31}$$

■

Consider the difference between $\mathcal{W}_\eta^o$ and $\mathcal{W}^o$. It turns out that the smoother $\mathcal{W}^o$ is, the smaller $\|\mathcal{W}^o - \mathcal{W}_\eta^o\|$ will be. To see this, let us subtract $\overline{\mathcal{W}}^o$ from both sides of equation (25). We obtain:

$$\overline{\mathcal{W}}_\eta^o - \overline{\mathcal{W}}^o = \begin{bmatrix} \mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}\left(I_{M(N-1)} - \mathcal{K}\right) \\ \mathcal{K} - I_{M(N-1)} \end{bmatrix}[\overline{\mathcal{W}}^o]_{2:N}. \tag{32}$$

The difference $\overline{\mathcal{W}}_\eta^o - \overline{\mathcal{W}}^o$ depends on $[\overline{\mathcal{W}}^o]_{2:N}$. Thus, from (15) and (32), we conclude that the smoother $\mathcal{W}^o$ is, the smaller $\|\mathcal{W}_\eta^o - \mathcal{W}^o\| = \|\overline{\mathcal{W}}_\eta^o - \overline{\mathcal{W}}^o\|$ will be.

Lemma 1 will be useful in the sequel to establish Theorem 1 and to provide a low-pass graph filter interpretation for the uniform Hessian matrices scenario considered in Section VI-A in Part II [24].

III. NETWORK STABILITY

We examine the behavior of algorithm (9) under Assumption 2 on the gradient noise processes $\{\boldsymbol{s}_{k,i}(\cdot)\}$ defined in (6). As explained in [7], [10], these conditions are automatically satisfied by many risk functions of interest in learning and adaptation such as quadratic (considered in the simulation section of Part II [24]) and logistic risks (considered in the weather forecasting application of Section IV). Condition (33) essentially states that the gradient vector approximation should be unbiased conditioned on the past data, which is a reasonable condition to require. Condition (34) states that the second-order moment of the gradient noise process should get smaller for better estimates, since it is bounded by the squared norm of the iterate. Condition (35) states that the gradient noises across the agents are uncorrelated.

Assumption 2 (Gradient noise process): The gradient noise process defined in (6) satisfies for any $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$ and for all $k, \ell = 1, 2, \ldots, N$:

$$\mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}] = 0, \tag{33}$$

$$\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \leq \beta_k^2\|\boldsymbol{w}\|^2 + \sigma_{s,k}^2, \tag{34}$$

$$\mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w})\boldsymbol{s}_{\ell,i}^\top(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}] = 0, \quad k \neq \ell, \tag{35}$$

for some $\beta_k^2 \geq 0$, $\sigma_{s,k}^2 \geq 0$, and where $\boldsymbol{\mathcal{F}}_{i-1}$ denotes the filtration generated by the random processes $\{\boldsymbol{w}_{\ell,j}\}$ for all $\ell = 1, \ldots, N$ and $j \leq i-1$. □

In this section, we analyze how well the multitask strategy (9) approaches the optimal solution $\mathcal{W}_\eta^o$ of the regularized cost (4). We examine this performance in terms of the mean-square-error measure, $\mathbb{E}\|w_{k,\eta}^o - \boldsymbol{w}_{k,i}\|^2$, the fourth-order moment, $\mathbb{E}\|w_{k,\eta}^o - \boldsymbol{w}_{k,i}\|^4$, and the mean-error process, $\mathbb{E}(w_{k,\eta}^o - \boldsymbol{w}_{k,i})$. To establish mean-square error stability, we

extend the energy analysis framework of [6] to handle multitask distributed optimization. Then, following a similar line of reasoning as in [7, Chapter 9], we establish the stability of the first and fourth-order moments, which is necessary to arrive at an expression for the steady-state performance in Part II [24].

Let us introduce the network block vector $\boldsymbol{\mathcal{W}}_i = \text{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\}$. At each iteration, we can view (9) as a mapping from $\boldsymbol{\mathcal{W}}_{i-1}$ to $\boldsymbol{\mathcal{W}}_i$:

$$\boxed{\boldsymbol{\mathcal{W}}_i = (I_{MN} - \mu\eta\mathcal{L})\left(\boldsymbol{\mathcal{W}}_{i-1} - \mu\,\text{col}\left\{\widehat{\nabla_{w_k} J_k}(\boldsymbol{w}_{k,i-1})\right\}_{k=1}^{N}\right)} \tag{36}$$

We introduce the following condition on the combination matrix $(I_{MN} - \mu\eta\mathcal{L})$, which is necessary for studying the performance of (9). It can be easily verified that this requirement is always met by selecting μ and η to satisfy the bounds (37)–(38).

Assumption 3 (Combination matrix): The symmetric combination matrix $(I_{MN} - \mu\eta\mathcal{L})$ has nonnegative entries and its spectral radius is equal to one. Since L has an eigenvalue at zero, these conditions are satisfied when the step-size $\mu > 0$ and the regularization strength $\eta \geq 0$ satisfy:

$$0 \leq \mu\eta \leq \frac{2}{\lambda_{\max}(L)}, \tag{37}$$

$$0 \leq \mu\eta \leq \min_{1\leq k\leq N}\left\{\frac{1}{\sum_{\ell=1}^{N} a_{k\ell}}\right\}, \tag{38}$$

where condition (37) ensures stability and condition (38) ensures non-negative entries. □

A. STABILITY OF SECOND-ORDER ERROR MOMENT

We first show that algorithm (9), in the absence of gradient noise, converges and has a unique fixed-point. Then, we analyze the distance between this point and the vectors $w_{k,\eta}^o$ and $\boldsymbol{w}_{k,i}$ in the mean-square-sense.

1) EXISTENCE AND UNIQUENESS OF FIXED-POINT

Without gradient noise, relation (36) reduces to:

$$\mathcal{W}_i = (I_{MN} - \mu\eta\mathcal{L})\left(\mathcal{W}_{i-1} - \mu\,\text{col}\left\{\nabla_{w_k} J_k(w_{k,i-1})\right\}_{k=1}^{N}\right). \tag{39}$$

Let $\mathcal{X} \triangleq \text{col}\{x_1, \ldots, x_N\}$ denote an $N \times 1$ block vector, where x_k is $M \times 1$. The mapping (39) is equivalent to the deterministic mapping $\mathcal{X} \to \mathcal{Y}$ defined as:

$$\mathcal{Y} = (I_{MN} - \mu\eta\mathcal{L})\left(\mathcal{X} - \mu\,\text{col}\left\{\nabla_{w_k} J_k(x_k)\right\}_{k=1}^{N}\right). \tag{40}$$

Lemma 2: (Contractive mapping) Under Assumption 1 and condition (37), the deterministic mapping defined in (40) satisfies:

$$\|\mathcal{Y}^1 - \mathcal{Y}^2\| \leq \gamma\|\mathcal{X}^1 - \mathcal{X}^2\|, \tag{41}$$

with $\gamma \triangleq \max_{1\leq k\leq N}\{\gamma_k\}$ where:

$$\gamma_k \triangleq \max\{|1 - \mu\lambda_{k,\min}|, |1 - \mu\lambda_{k,\max}|\}. \tag{42}$$

This mapping is contractive when μ satisfies:

$$0 < \mu < \min_{1\leq k\leq N}\left\{\frac{2}{\lambda_{k,\max}}\right\}. \tag{43}$$

Proof: See Appendix B. ■

It then follows from Banach's fixed point theorem [36, pp. 299–303] that iteration (39) converges to a unique fixed point $\mathcal{W}_\infty = \lim_{i\to\infty}\mathcal{W}_i = \text{col}\{w_{1,\infty}, \ldots, w_{N,\infty}\}$ at an exponential rate given by γ. Observe that this fixed point is not $\mathcal{W}_\eta^o$. Since we wish to study $\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^2$, which can be decomposed as:

$$\begin{aligned}
&\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^2 \\
&= \limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o - \mathcal{W}_\infty + \mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i\|^2 \\
&\leq 2\|\mathcal{W}_\eta^o - \mathcal{W}_\infty\|^2 + 2\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i\|^2,
\end{aligned} \tag{44}$$

we shall first assess the size of $\|\mathcal{W}_\eta^o - \mathcal{W}_\infty\|^2$ and then examine the quantity $\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i\|^2$.

2) FIXED POINT BIAS ANALYSIS

Now we analyze how far this fixed point $\mathcal{W}_\infty$ is from the desired solution $\mathcal{W}_\eta^o$ when the step-size μ is small. We carry out the analysis in two steps. First, we derive an expression for $\widetilde{\mathcal{W}}_\infty \triangleq \mathcal{W}_\eta^o - \mathcal{W}_\infty$ and then we assess its size. Since $\mathcal{W}_\infty$ is the fixed point of (39), we have at convergence:

$$\boxed{\mathcal{W}_\infty = (I_{MN} - \mu\eta\mathcal{L})\left(\mathcal{W}_\infty - \mu\,\text{col}\left\{\nabla_{w_k} J_k(w_{k,\infty})\right\}_{k=1}^{N}\right)} \tag{45}$$

Let $\widetilde{w}_{k,\infty} \triangleq w_{k,\eta}^o - w_{k,\infty}$. Using the mean-value theorem [35, pp. 24],[7, Appendix D], we can write:

$$\nabla_{w_k} J_k(w_{k,\infty}) = \nabla_{w_k} J_k(w_{k,\eta}^o) - H_{k,\infty}\widetilde{w}_{k,\infty}, \tag{46}$$

where

$$H_{k,\infty} \triangleq \int_0^1 \nabla_{w_k}^2 J_k(w_{k,\eta}^o - t\widetilde{w}_{k,\infty})dt. \tag{47}$$

Subtracting the vector $(I_{MN} - \mu\eta\mathcal{L})\mathcal{W}_\eta^o$ from both sides of (45) and using relation (46), we obtain:

$$\begin{aligned}
\widetilde{\mathcal{W}}_\infty = (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\infty)\widetilde{\mathcal{W}}_\infty + \mu\eta\mathcal{L}\mathcal{W}_\eta^o + \\
\mu(I_{MN} - \mu\eta\mathcal{L})\text{col}\left\{\nabla_{w_k} J_k(w_{k,\eta}^o)\right\}_{k=1}^{N},
\end{aligned} \tag{48}$$

where $\mathcal{H}_\infty \triangleq \text{diag}\{H_{1,\infty}, \ldots, H_{N,\infty}\}$. From (16), recursion (48) can be written alternatively as:

$$\widetilde{\mathcal{W}}_\infty = (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\infty)\widetilde{\mathcal{W}}_\infty + \mu^2\eta^2\mathcal{L}^2\mathcal{W}_\eta^o, \tag{49}$$

so that:

$$\boxed{\widetilde{\mathcal{W}}_\infty = \mu^2\eta^2\left[I_{MN} - (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\infty)\right]^{-1}\mathcal{L}^2\mathcal{W}_\eta^o} \tag{50}$$

The inverse exists when $(I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\infty)$ is stable, i.e., its spectral radius is less than one. Since the spectral

radius of a matrix is upper bounded by any of its induced norms, we have:

$$\rho((I_{MN}-\mu\eta\mathcal{L})(I_{MN}-\mu\mathcal{H}_\infty)) \le \|I_{MN}-\mu\eta\mathcal{L}\|\|I_{MN}-\mu\mathcal{H}_\infty\|, \quad (51)$$

in terms of the 2−induced norm. Under condition (37) and since $\lambda_1(L)=0$, we have $\|I_{MN}-\mu\eta\mathcal{L}\|=1$. From Assumption 1, we have:

$$(1-\mu\lambda_{k,\max})I_M \le I_M-\mu H_{k,\infty} \le (1-\mu\lambda_{k,\min})I_M, \quad (52)$$

so that $\|I_{MN}-\mu\mathcal{H}_\infty\|_2 \le \max_{1\le k\le N}\gamma_k$ with γ_k given in (42). We conclude that when (37) and (43) are satisfied, the inverse exists.

From (50), we observe that $\widetilde{\mathcal{W}}_\infty$ is zero in two cases: i) when $\eta=0$; ii) when $w_k^o=w^o\ \forall k$, i.e., $\mathcal{W}^o=\mathbb{1}_N\otimes w^o$. In the second case, consider (25) and observe that $\overline{w}_1^o=\sqrt{N}w^o$, $[\overline{w}^o]_{2:N}=0$, and $\overline{\mathcal{W}}_\eta^o=\text{col}\{\overline{w}_1^o,0\}$. Thus, $\mathcal{W}_\eta^o=(V\otimes I_M)\overline{\mathcal{W}}_\eta^o=(v_1\otimes I_M)\overline{w}_1^o=\mathbb{1}_N\otimes w^o$ and $\mathcal{L}\mathcal{W}_\eta^o=\mathcal{L}(\mathbb{1}_N\otimes w^o)=0$.

Theorem 1: (Fixed point bias size) Under Assumption 1 and for small μ satisfying conditions (37) and (43), the steady-state bias $\widetilde{\mathcal{W}}_\infty=\mathcal{W}_\eta^o-\mathcal{W}_\infty$ of the mapping (39) satisfies:

$$\lim_{\mu\to0}\left(\frac{1}{\mu}\|\mathcal{W}_\eta^o-\mathcal{W}_\infty\|\right) \le \frac{O(\eta^2)}{(O(1)+O(\eta))^2}. \quad (53)$$

Proof: See Appendix C. ■

3) EVOLUTION OF THE STOCHASTIC RECURSION

We now examine how close the stochastic algorithm (9) approaches $\mathcal{W}_\eta^o$. First, we introduce the mean-square perturbation vector (MSP) at time i relative to $\mathcal{W}_\infty$:

$$\text{MSP}_i \triangleq \text{col}\left\{\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i}\|^2\right\}_{k=1}^N. \quad (54)$$

The k-th entry of MSP_i characterizes how far away the estimate $\boldsymbol{w}_{k,i}$ at agent k and time i is from $w_{k,\infty}$.

Theorem 2 (Network mean-square-error stability): Under Assumptions 1, 2, and 3, the *MSP* at time i can be recursively bounded as:

$$\text{MSP}_i \preceq (I_N-\mu\eta L)\,G\,\text{MSP}_{i-1}+\mu^2(I_N-\mu\eta L)b, \quad (55)$$

where:

$$G \triangleq \text{diag}\left\{\gamma_k^2+3\mu^2\beta_k^2\right\}_{k=1}^N, \quad (56)$$

$$b \triangleq \text{col}\left\{\sigma_{s,k}^2+3\beta_k^2\|w_{k,\eta}^o\|^2+3\beta_k^2\|w_{k,\eta}^o-w_{k,\infty}\|^2\right\}_{k=1}^N. \quad (57)$$

A sufficient condition for the stability of the above recursion is:

$$0<\mu<\min_{1\le k\le N}\left\{\min\left\{\frac{2\lambda_{k,\min}}{\lambda_{k,\min}^2+3\beta_k^2},\frac{2\lambda_{k,\max}}{\lambda_{k,\max}^2+3\beta_k^2}\right\}\right\}. \quad (58)$$

It follows that

$$\|\limsup_{i\to\infty}\text{MSP}_i\|_\infty = O(\mu), \quad (59)$$

and

$$\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o-\boldsymbol{\mathcal{W}}_i\|^2 = O(\mu)+\frac{O(\mu^2\eta^4)}{(O(1)+O(\eta))^4}=O(\mu). \quad (60)$$

Proof: See Appendix D. With regards to (60) note first that for fixed η, we have $O(\mu)+O(\mu^2)=O(\mu)$. When η and μ are coupled ($\eta=\mu^{-\epsilon}$), we obtain:

$$\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o-\boldsymbol{\mathcal{W}}_i\|^2 = O(\mu)+\frac{O(\mu^{2-4\epsilon})}{O(1)+O(\mu^{-4\epsilon})} \quad (61)$$

For $\epsilon<0$, $O(1)$ dominates $O(\mu^{-4\epsilon})$ in the denominator and we obtain $O(\mu)+O(\mu^{2-4\epsilon})=O(\mu)$. For $\epsilon>0$, $O(\mu^{-4\epsilon})$ dominates $O(1)$ in the denominator and we obtain $O(\mu)+O(\mu^2)=O(\mu)$. ■

B. STABILITY OF FOURTH-ORDER ERROR MOMENT

The results so far establish that the iterates $\boldsymbol{w}_{k,i}$ converge to a small $O(\mu)-$ neighborhood around the regularized solution $w_{k,\eta}^o$. We can be more precise and determine the size of this neighborhood, i.e., assess the size of the constant multiplying μ in the $O(\mu)-$term. To do so, we shall derive in Part II [24] an accurate first-order expression for the mean-square error (60); the expression will be accurate to first-order in μ. This expression will be useful because it will allow us to highlight several features of the limiting point of the network as a function of the parameter η.

To arrive at the desired expression, in Part II we first need to introduce a long-term approximation model and assess how close it is to the actual model. We then derive the performance for the long-term model and use this closeness to transform this result into an accurate expression for the performance of the original learning algorithm. When this argument is concluded we arrive at the desired performance expression, which we then use to comment on the behavior of the algorithm in a more informed manner. To derive the long-term model, we shall follow the approach developed in [7]. The first step is to establish the asymptotic stability of the fourth-order moment of the error vector, $\mathbb{E}\|\mathcal{W}_\eta^o-\boldsymbol{\mathcal{W}}_i\|^4$. This property is needed to justify the validity of the long-term approximate model that will be introduced in Part II [24].

To establish the fourth-order stability, we introduce the following assumption on the gradient noise process.

Assumption 4 (Fourth-order moment of the gradient noise): The gradient noise process defined in (6) satisfies for any $\boldsymbol{w}\in\boldsymbol{\mathcal{F}}_{i-1}$ and for all $k,\ell=1,2,\ldots,N$:

$$\mathbb{E}\left[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right] \le \overline{\beta}_k^4\|\boldsymbol{w}_k\|^4+\overline{\sigma}_{s,k}^4, \quad (62)$$

for some $\overline{\beta}_k^4\ge0$, and $\overline{\sigma}_{s,k}^4\ge0$.

As explained in [7], condition (62) implies (34) and, likewise, condition (62) holds for important cases of interest.

Exploiting the convexity of the norm functions $\|x\|^4$ and $\|x\|^2$ and using Jensen's inequality, we can write:

$$\mathbb{E}\|\mathcal{W}_\eta^o-\boldsymbol{\mathcal{W}}_i\|^4 \le 8\|\mathcal{W}_\eta^o-\mathcal{W}_\infty\|^4+8\mathbb{E}\|\mathcal{W}_\infty-\boldsymbol{\mathcal{W}}_i\|^4, \quad (63)$$

and

$$\begin{aligned}\mathbb{E}\|\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i\|^4 &= \mathbb{E}\left(\|\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i\|^2\right)^2 \\ &= \mathbb{E}\left(\sum_{k=1}^N \|w_{k,\infty} - \boldsymbol{w}_{k,i}\|^2\right)^2 \\ &= N^2\mathbb{E}\left(\sum_{k=1}^N \frac{1}{N}\|w_{k,\infty} - \boldsymbol{w}_{k,i}\|^2\right)^2 \\ &\le N\sum_{k=1}^N \mathbb{E}\|w_{k,\infty} - \boldsymbol{w}_{k,i}\|^4. \end{aligned} \tag{64}$$

Let us introduce the mean-fourth perturbation vector at time i relative to $\mathcal{W}_\infty$:

$$\mathrm{MFP}_i \triangleq \mathrm{col}\left\{\mathbb{E}\|w_{k,\infty} - \boldsymbol{w}_{k,i}\|^4\right\}_{k=1}^N. \tag{65}$$

Theorem 3 (Fourth-order error moment stability): Under Assumptions 1, 2, 3, and 4, the *MFP* at time i can be recursively bounded as:

$$\begin{aligned}\mathrm{MFP}_i \preceq\ & (I_N - \mu\eta L)G'\mathrm{MFP}_{i-1} + \\ & \mu^2(I_N - \mu\eta L)B\mathrm{MSP}_{i-1} + \mu^4(I_N - \mu\eta L)b,' \end{aligned} \tag{66}$$

where

$$G' \triangleq \mathrm{diag}\left\{\gamma_k^4 + 24\mu^2\gamma_k^2\beta_k^2 + 81\mu^4\overline{\beta}_k^4\right\}_{k=1}^N, \tag{67}$$

$$B \triangleq 8\gamma_k^2\mathrm{diag}\left\{\sigma_{s,k}^2 + 3\beta_k^2\|w_{k,\eta}^o\|^2 + 3\beta_k^2\|w_{k,\eta}^o - w_{k,\infty}\|^2\right\}_{k=1}^N \tag{68}$$

$$b' \triangleq \mathrm{col}\left\{3\overline{\sigma}_{s,k}^4 + 81\overline{\beta}_k^4\|w_{k,\eta}^o\|^4 + 81\overline{\beta}_k^4\|w_{k,\eta}^o - w_{k,\infty}\|^4\right\}_{k=1}^N \tag{69}$$

A sufficiently small μ ensures the stability of the above recursion. It follows that

$$\|\limsup_{i\to\infty}\mathrm{MFP}_i\|_\infty = O(\mu^2), \tag{70}$$

and

$$\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^4 = O(\mu^2) + \frac{O(\mu^4\eta^8)}{(O(1)+O(\eta))^8} = O(\mu^2). \tag{71}$$

Proof: See Appendix E. ■

C. STABILITY OF FIRST-ORDER ERROR MOMENT

We next need to examine the evolution of the mean-error vector $\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)$. To establish the mean-stability, we need to introduce a smoothness condition on the Hessian matrices of the individual costs. This smoothness condition will be adopted in the next Part II [24] when we study the long term behavior of the network.

Assumption 5 (Smoothness condition on individual cost functions): It is assumed that each $J_k(w_k)$ satisfies a smoothness condition close to $w_{k,\eta}^o$, in that the corresponding Hessian matrix is Lipchitz continuous in the proximity of $w_{k,\eta}^o$ with some parameter $\kappa_d \ge 0$, i.e.,

$$\|\nabla_{w_k}^2 J_k(w_{k,\eta}^o + \Delta w_k) - \nabla_{w_k}^2 J_k(w_{k,\eta}^o)\| \le \kappa_d\|\Delta w_k\|, \tag{72}$$

for small perturbations $\|\Delta w_k\| \le \epsilon$. □

From the triangle inequality, we have:

$$\|\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)\| \le \|\mathcal{W}_\eta^o - \mathcal{W}_\infty\| + \|\mathbb{E}(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i)\|. \tag{73}$$

Let us introduce the square-mean perturbation (SMP) vector at time i relative to $\mathcal{W}_\infty$:

$$\mathrm{SMP}_i \triangleq \mathrm{col}\left\{\|\mathbb{E}(w_{k,\infty} - \boldsymbol{w}_{k,i})\|^2\right\}_{k=1}^N. \tag{74}$$

Theorem 4 (First-order error moment stability): Under Assumptions 1, 2, 3, and 5, the *SMP* at time i can be recursively bounded as:

$$\begin{aligned}\mathrm{SMP}_i \preceq\ & (I_N - \mu\eta L)G''\mathrm{SMP}_{i-1} + \\ & \mu^2(I_N - \mu\eta L)(I_N - G'')^{-1}B'\mathrm{MSP}_{i-1} + \\ & \mu^2\frac{1}{2}(\kappa_d')^2(I_N - \mu\eta L)(I_N - G'')^{-1}\mathrm{MFP}_{i-1}. \end{aligned} \tag{75}$$

where

$$G'' \triangleq \mathrm{diag}\{\gamma_k\}_{k=1}^N, \tag{76}$$

$$B' \triangleq 2(\kappa_d')^2\mathrm{diag}\left\{\|w_{k,\eta}^o - w_{k,\infty}\|^2\right\}_{k=1}^N, \tag{77}$$

with $\kappa_d' = \max\{\kappa_d, \frac{\lambda_{k,\max} - \lambda_{k,\min}}{\epsilon}\}$. Under Assumption 4, a sufficiently small μ ensures the stability of the above recursion. It follows that

$$\|\limsup_{i\to\infty}\mathrm{SMP}_i\|_\infty = O(\mu^2), \tag{78}$$

and that

$$\limsup_{i\to\infty}\|\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)\| = O(\mu) + \frac{O(\mu\eta^2)}{(O(1)+O(\eta))^2}. \tag{79}$$

Proof: See Appendix F. ■

We have established so far the stability of the mean-error process, $\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)$, the mean-square-error $\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^2$, and the fourth order moment $\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^4$. Building on these results, we will derive in Part II [24] closed form expressions for the steady-state performance of algorithm (9). Section VI in Part II [24] will provide illustration for the theoretical results in this part (Theorems 1, 2, and 4), and its accompanying Part II.

IV. SIMULATION RESULTS WITH REAL DATASET

In this section, we test algorithm (9) on a weather dataset corresponding to a collection of daily measurements (mean temperature, mean dew point, mean visibility, mean wind speed, maximum sustained wind speed, and rain or snow occurrence) taken from 2004 to 2017 at $N = 139$ weather stations located around the continental United States [37]. We construct a representation graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, A)$ for the stations using geographical distances between sensors. Each sensor corresponds to a node k and is connected to $|\mathcal{N}_k|$

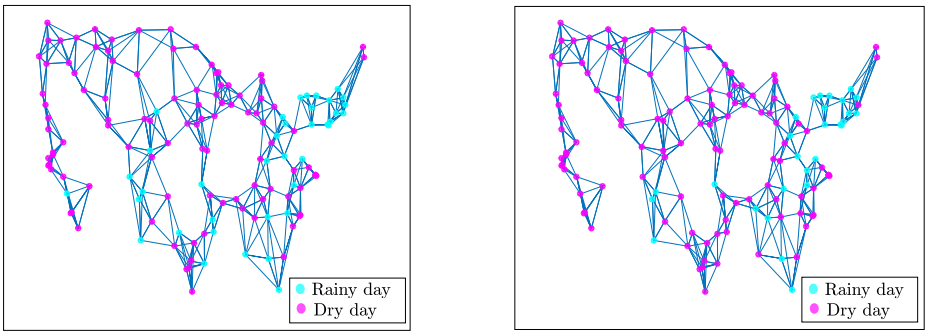


FIGURE 3. ***(Left)*** **Occurrence of rain reported by 139 weather stations across the US on July 30, 2015.** ***(Right)*** **Prediction of rain occurrence from weather data based on logistic regression and multitask learning.**

neighbor nodes with undirected edges weighted according to $a_{k\ell} = \frac{1}{2}(p_{k\ell} + p_{\ell k})$ with [32]:

$$p_{k\ell} = \frac{e^{-d_{k\ell}^2}}{\sqrt{\sum_{m\in\mathcal{N}_{k,0}} e^{-d_{km}^2} \sum_{n\in\mathcal{N}_{\ell,0}} e^{-d_{\ell n}^2}}}, \quad \ell \in \mathcal{N}_{k,0}, \quad (80)$$

where $\mathcal{N}_{k,0}$ is the set of 4-nearest neighbors of node k and $d_{k\ell}$ denotes the geodesical distance between the k-th and ℓ-sensors – see Fig. 3 (left). Let $h_{k,i} \in \mathbb{R}^M$ denote the feature vector at sensor k and day i composed of $M = 5$ entries corresponding to the mean temperature, mean dew point, mean visibility, mean wind speed, and maximum sustained wind speed reported at day i at sensor k. Let $\gamma_k(i)$ denote a binary variable associated with the occurrence of rain (or snow) at node k and day i, i.e, $\gamma_k(i) = 1$ if rain (or snow) occurred and $\gamma_k(i) = -1$ otherwise. We would like to construct a classifier that allows us to predict whether it will rain (or snow) or not based on the knowledge of the feature vector $h_{k,i}$. In principle, each station could use an individual logistic regression machine [7], [38], [39], that seeks a vector w_k^o, such that $\widehat{\boldsymbol{\gamma}}_k(i) = \text{sign}(\boldsymbol{h}_{k,i}^\top w_k^o)$ and

$$w_k^o \triangleq \arg\min_{w_k} \mathbb{E} \ln\left(1 + e^{-\boldsymbol{\gamma}_k(i)\boldsymbol{h}_{k,i}^\top w_k}\right) + \rho\|w_k\|^2. \quad (81)$$

In this application, however, it is expected that the decision rules $\{w_k^o\}$ at neighboring stations will be similar. In the experiment, the dataset is split into a training set used to learn the decision rule w_k^o, and a test set from which $\widehat{\boldsymbol{\gamma}}_k(i)$ are generated for performance evaluation. The first dataset comprises daily weather data recorded at the stations in the interval $2004 - 2012$ (a total number of $D_a = 3288$ days) and the test set contains data recorded in the interval $2012 - 2017$ (a total number of $D_t = 1826$ days). We set $\mu = 3 \cdot 10^{-4}$ and $\rho = 10^{-5}$. We generate the first iterate $w_{k,0}$ from the Gaussian distribution $\mathcal{N}(0, I_M)$ and we run strategy (9) over the training set ($i = 1, \ldots, D_a$) for different values of η. For each value of η, we report in Table 2 the prediction error over the test set defined as:

$$\frac{1}{N}\sum_{k=1}^{N}\frac{1}{D_t}\sum_{i=1}^{D_t=1826} \mathbb{I}[\text{sign}(\boldsymbol{h}_{k,i}^\top \widehat{w}_{k,\infty}) \neq \gamma_k(i)], \quad (82)$$

TABLE 2. Rain Prediction Error (82) in Weather Sensor Networks for Different Values of Regularization Strength η

	$\eta = 0$	$\eta = 10$	$\eta = 45$	$\eta = 100$	$\eta = 1000$	$\eta = \mu^{-1}$
prediction error	0.309	0.232	**0.225**	0.226	0.228	0.232

where $N = 139$ is the number of nodes, $\widehat{w}_{k,\infty}$ is the average of the last 200 iterates generated by the algorithm at agent k, and $\mathbb{I}[x]$ is the indicator function at x, namely, $\mathbb{I}[x] = 1$ if x is true and 0 otherwise. Table 2 shows that through cooperation, the agents improve performance. This is due to the fact that the non-cooperative solution ($\eta = 0$) may suffer from a slow convergence rate [1, Section V-B] in which case some nodes may not be able to converge in the finite dataset scenario. By increasing η, the convergence rate improves. However, a large value of η (such as $\eta = \mu^{-1}$) yields a deterioration in the accuracy since in this case all agents converge approximately to the same classifier. By setting $\eta = 45$, we obtain the smallest prediction error. We show in Fig. 3 (right) the results of the prediction on July 30, 2015 across the US for $\eta = 45$.

V. CONCLUSION

In this work, we considered multitask inference problems where agents in the network have individual parameter vectors to estimate subject to a smoothness condition over the graph. Based on diffusion adaptation, we proposed a strategy that allows the network to minimize a global cost consisting of the aggregate sum of the individual costs regularized by a term promoting smoothness. We showed that, for small step-size parameter, the network is able to approach the minimizer of the regularized problem to arbitrarily good accuracy levels. Furthermore, we showed how the regularization strength can steer the convergence point of the network toward many modes starting from the non-cooperative mode and ending with the single-task mode.

APPENDIX A PROOF OF LEMMA 1

Consider the matrix inversion identity [40]:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}, \quad (83)$$

which allows us to write:

$$(U+W)^{-1}U = I - U^{-1}(I+WU^{-1})^{-1}$$
$$W = I - (U+W)^{-1}W, \tag{84}$$

for any invertible matrix U. Using (84), we write (20) alternatively as:

$$\overline{\mathcal{W}}_{\eta}^{o} = \left(I - \eta\left(\overline{\mathcal{H}}_{\eta}^{o} + \eta\mathcal{J}\right)^{-1}\mathcal{J}\right)\overline{\mathcal{W}}^{o}. \tag{85}$$

Let

$$\mathcal{Q} \triangleq \overline{\mathcal{H}}_{\eta}^{o} + \eta\mathcal{J}. \tag{86}$$

Using the definitions (22) and (23), we can partition $\mathcal{Q}$ into blocks:

$$\mathcal{Q} = \begin{bmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{12}^{\top} & \mathcal{Q}_{22} \end{bmatrix}, \tag{87}$$

with Q_{11}, $\mathcal{Q}_{12}$, and $\mathcal{Q}_{22}$ defined in (26), (27), and (28), respectively. Since $v_1 = \frac{1}{\sqrt{N}}\mathbb{1}_N$, we have $\mathcal{Q}_{11} = \frac{1}{N}\sum_{k=1}^{N} H_{k,\eta}^{o}$ which is positive definite from Assumption 1. Observe that $\mathcal{Q}$ is invertible since it is similar to $\mathcal{H}_{\eta}^{o} + \eta\mathcal{L}$ which is positive definite under Assumption 1. Now, by applying the block inversion formula to $\mathcal{Q}$, we obtain:

$$\mathcal{Q}^{-1} = (\overline{\mathcal{H}}_{\eta}^{o} + \eta\mathcal{J})^{-1}$$
$$= \begin{bmatrix} \mathcal{Q}_{11}^{-1} + \mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}\mathcal{G}\mathcal{Q}_{12}^{\top}\mathcal{Q}_{11}^{-1} & -\mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}\mathcal{G} \\ -\mathcal{G}\mathcal{Q}_{12}^{\top}\mathcal{Q}_{11}^{-1} & \mathcal{G} \end{bmatrix}, \tag{88}$$

where

$$\mathcal{G} \triangleq (\mathcal{Q}_{22} - \mathcal{Q}_{12}^{\top}\mathcal{Q}_{11}^{-1}\mathcal{Q}_{12})^{-1}. \tag{89}$$

Replacing (88) into (85) and using (22), we arrive at:

$$\overline{\mathcal{W}}_{\eta}^{o} = \begin{bmatrix} I_M & \eta\mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}\mathcal{G}\,(\Lambda_o \otimes I_M) \\ 0 & I_{M(N-1)} - \eta\,\mathcal{G}\,(\Lambda_o \otimes I_M) \end{bmatrix}\overline{\mathcal{W}}^{o}. \tag{90}$$

Using definition (30) into (90), we conclude (25).

Now, we establish (31). Let us first introduce the matrix $\mathcal{G}'$:

$$\mathcal{G}' \triangleq (V_R^{\top} \otimes I_M)\mathcal{H}_{\eta}^{o}(V_R \otimes I_M) - \mathcal{Q}_{12}^{\top}\mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}. \tag{91}$$

Using the above definition and expressions (29) and (28), we can re-write the matrix $\mathcal{K}$ in (30) alternatively as:

$$\mathcal{K} = I_{M(N-1)} - \eta\,[(V_R^{\top} \otimes I_M)\mathcal{H}_{\eta}^{o}(V_R \otimes I_M) + \eta\Lambda_o \otimes I_M - \mathcal{Q}_{12}^{\top}\mathcal{Q}_{11}^{-1}\mathcal{Q}_{12}]^{-1}\,(\Lambda_o \otimes I_M)$$
$$= I_{M(N-1)} - \eta(\mathcal{G}' + \eta\Lambda_o \otimes I_M)^{-1}(\Lambda_o \otimes I_M)$$
$$\overset{(84)}{=} (\mathcal{G}' + \eta\Lambda_o \otimes I_M)^{-1}\mathcal{G}'. \tag{92}$$

The matrix $\mathcal{G}'$ in (91) is the Schur complement of $\overline{\mathcal{H}}_{\eta}^{o}$ in (23) which can be partitioned as:

$$\overline{\mathcal{H}}_{\eta}^{o} = \begin{bmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} \\ \mathcal{Q}_{12}^{\top} & (V_R^{\top} \otimes I_M)\mathcal{H}_{\eta}^{o}(V_R \otimes I_M) \end{bmatrix}. \tag{93}$$

Thus, $\mathcal{G}'$ is positive definite since it is the Schur complement of the positive definite matrix $\overline{\mathcal{H}}_{\eta}^{o}$ [41, pp. 651]. Since $\mathcal{G}'$ is symmetric, from Weyl's inequality [42, pp. 239] we have:

$$0 < \eta\lambda_2(L) + \lambda_{\min}(\mathcal{G}') \leq \lambda_{\min}(\mathcal{G}' + \eta\Lambda_o \otimes I_M) \leq \eta\lambda_2(L) + \lambda_{\max}(\mathcal{G}'). \tag{94}$$

Furthermore, since $\mathcal{G}'$ is the Schur complement of the positive definite matrix $\overline{\mathcal{H}}_{\eta}^{o}$, we have [43, Theorem 5]:

$$\lambda_{\min}(\mathcal{G}') \geq \lambda_{\min}(\overline{\mathcal{H}}_{\eta}^{o}) = \lambda_{\min}(\mathcal{H}_{\eta}^{o}) \geq \min_{1\leq k\leq N}\lambda_{k,\min}, \tag{95}$$

$$\lambda_{\max}(\mathcal{G}') \leq \lambda_{\max}(\overline{\mathcal{H}}_{\eta}^{o}) = \lambda_{\max}(\mathcal{H}_{\eta}^{o}) \leq \max_{1\leq k\leq N}\lambda_{k,\max}. \tag{96}$$

Therefore, from (94) and (95), we get:

$$\lambda_{\min}(\mathcal{G}' + \eta\Lambda_o \otimes I_M) \geq \eta\lambda_2(L) + \min_{1\leq k\leq N}\lambda_{k,\min}, \tag{97}$$

and

$$\lambda_{\max}((\mathcal{G}' + \eta\Lambda_o \otimes I_M)^{-1}) = \frac{1}{\lambda_{\min}(\mathcal{G}' + \eta\Lambda_o \otimes I_M)} \leq \left(\eta\lambda_2(L) + \min_{1\leq k\leq N}\lambda_{k,\min}\right)^{-1}. \tag{98}$$

Since the 2−induced norm of a positive definite matrix is equal to its maximum eigenvalue, we obtain:

$$\|(\mathcal{G}' + \eta\Lambda_o \otimes I_M)^{-1}\| \leq \left(\eta\lambda_2(L) + \min_{1\leq k\leq N}\lambda_{k,\min}\right)^{-1}. \tag{99}$$

From the sub-multiplicative property of the 2−induced norm and from (92), (99), and (96), we obtain:

$$\|\mathcal{K}\| \leq \|(\mathcal{G}' + \eta\Lambda_o \otimes I_M)^{-1}\| \cdot \|\mathcal{G}'\| \leq \left(\max_{1\leq k\leq N}\lambda_{k,\max}\right)\left(\eta\lambda_2(L) + \min_{1\leq k\leq N}\lambda_{k,\min}\right)^{-1}. \tag{100}$$

APPENDIX B PROOF OF LEMMA 2

Given any two input vectors x^1 and x^2 with corresponding updated vectors y^1 and y^2, we have from (40):

$$y^1 - y^2 = (I_{MN} - \mu\eta\mathcal{L}) \cdot \left(x^1 - x^2 - \mu\,\mathrm{col}\left\{\nabla_{w_k}J_k(x_k^1) - \nabla_{w_k}J_k(x_k^2)\right\}_{k=1}^{N}\right). \tag{101}$$

From the mean-value theorem [35, pp. 24], we have:

$$\nabla_{w_k}J_k(x_k^1) - \nabla_{w_k}J_k(x_k^2) = \left(\int_0^1 \nabla_{w_k}^2 J_k(x_k^2 + t(x_k^1 - x_k^2))dt\right)(x_k^1 - x_k^2). \tag{102}$$

Using (102) into (101), and the sub-multiplicative property of the 2−induced norm [7], we obtain:

$$\|y^1 - y^2\| \leq \|I_{MN} - \mu\eta\mathcal{L}\|\|\mathcal{D}\|\|x^1 - x^2\|, \tag{103}$$

where

$$\mathcal{D} \triangleq \operatorname{diag}\left\{I_M - \mu \int_0^1 \nabla^2_{w_k} J_k(x_k^2 + t(x_k^1 - x_k^2))dt\right\}_{k=1}^N. \tag{104}$$

We have

$$\|I_{MN} - \mu\eta\mathcal{L}\| = \|(I_N - \mu\eta L) \otimes I_M\| = \|I_N - \mu\eta L\|. \tag{105}$$

Let $\rho(\cdot)$ denote the spectral radius of its matrix argument. Since L is symmetric, we have $\|I_N - \mu\eta L\| = \rho(I_N - \mu\eta L)$. Since L has one eigenvalue at zero, $\rho(I_N - \mu\eta L)$ is guaranteed to be equal to 1 if $\mu\eta$ satisfies condition (37). For the block diagonal symmetric matrix $\mathcal{D}$ in (104), we have:

$$\|\mathcal{D}\| = \max_{1\le k\le N} \left\|I_M - \mu \int_0^1 \nabla^2_{w_k} J_k(x_k^2 + t(x_k^1 - x_k^2))dt\right\|. \tag{106}$$

Due to Assumption 1, we have:

$$0 < \lambda_{k,\min} I_M \le \int_0^1 \nabla^2_{w_k} J_k(x_k^2 + t(x_k^1 - x_k^2))dt \le \lambda_{k,\max} I_M. \tag{107}$$

It follows that $\|\mathcal{D}\| \le \gamma$ where $\gamma \triangleq \max_{1\le k\le N}\{\gamma_k\}$ and γ_k is given in (42). It holds that $0 < \gamma_k < 1$ when μ is chosen according to (43). Combining the previous results, we arrive at:

$$\|\mathcal{Y}^1 - \mathcal{Y}^2\|_2 \le \gamma \|x^1 - x^2\|_2, \tag{108}$$

for $\gamma < 1$ when (37) and (43) are satisfied and, in this case, the deterministic mapping (40) is a contraction.

APPENDIX C PROOF OF THEOREM 1

From (50), we obtain the following expression for $\widetilde{\mathcal{W}}_\infty$:

$$\widetilde{\mathcal{W}}_\infty = \mu\eta^2[\mathcal{H}_\infty + \eta\mathcal{L} - \mu\eta\mathcal{L}\mathcal{H}_\infty]^{-1}\mathcal{L}^2\mathcal{W}^o_\eta. \tag{109}$$

Pre-multiplying both sides of (109) by $\mathcal{V}^\top = V^\top \otimes I_M$ gives:

$$\overline{\mathcal{W}}_\infty = \mu\eta^2\left[\overline{\mathcal{H}}_\infty + \eta\mathcal{J} - \mu\eta\mathcal{J}\overline{\mathcal{H}}_\infty\right]^{-1}\mathcal{J}^2\,\overline{\mathcal{W}}^o_\eta, \tag{110}$$

where $\overline{\mathcal{W}}_\infty \triangleq \mathcal{V}^\top\widetilde{\mathcal{W}}_\infty$, $\overline{\mathcal{W}}^o_\eta \triangleq \mathcal{V}^\top\mathcal{W}^o_\eta$,

$$\overline{\mathcal{H}}_\infty \triangleq \mathcal{V}^\top\mathcal{H}_\infty\mathcal{V}, \tag{111}$$

and $\mathcal{J}$ is given by (22).

In the following we show that $\overline{\mathcal{W}}_\infty$ can be written as:

$$\overline{\mathcal{W}}_\infty = \mu\eta^2\begin{bmatrix} -\mathcal{P}_{11}^{-1}\mathcal{P}_{12}\mathcal{T} \\ \mathcal{T}\end{bmatrix}(\Lambda_o^2 \otimes I_M)\mathcal{K}[\overline{\mathcal{W}}^o]_{2:N}, \tag{112}$$

where $\mathcal{K}$ is defined in (30) and:

$$\mathcal{T} \triangleq (\mathcal{P}_{22} - \mathcal{P}_{21}\mathcal{P}_{11}^{-1}\mathcal{P}_{12})^{-1}, \tag{113}$$

$$\mathcal{P}_{11} \triangleq (v_1^\top \otimes I_M)\mathcal{H}_\infty(v_1 \otimes I_M) = \frac{1}{N}\sum_{k=1}^N H_{k,\infty}, \tag{114}$$

$$\mathcal{P}_{12} \triangleq (v_1^\top \otimes I_M)\mathcal{H}_\infty(V_R \otimes I_M), \tag{115}$$

$$\mathcal{P}_{21} \triangleq ((I_{N-1} - \mu\eta\Lambda_o) \otimes I_M)(V_R^\top \otimes I_M)\mathcal{H}_\infty(v_1 \otimes I_M), \tag{116}$$

$$\mathcal{P}_{22} \triangleq \eta\Lambda_o \otimes I_M + ((I_{N-1} - \mu\eta\Lambda_o) \otimes I_M)(V_R^\top \otimes I_M)\mathcal{H}_\infty(V_R \otimes I_M), \tag{117}$$

We introduce the following matrix, which appears in (110):

$$\mathcal{P} \triangleq (I_{MN} - \mu\eta\mathcal{J})\overline{\mathcal{H}}_\infty + \eta\mathcal{J} = \begin{bmatrix}\mathcal{P}_{11} & \mathcal{P}_{12} \\ \mathcal{P}_{21} & \mathcal{P}_{22}\end{bmatrix}, \tag{118}$$

where the blocks $\{\mathcal{P}_{ij}\}$ are given by (114)–(117). Note that, under Assumption 1, $\mathcal{P}_{11}$ in (114) is invertible since it can be bounded as follows:

$$0 < \frac{1}{N}\left(\sum_{k=1}^N \lambda_{k,\min}\right) I_M \le \mathcal{P}_{11} \le \frac{1}{N}\left(\sum_{k=1}^N \lambda_{k,\max}\right) I_M. \tag{119}$$

Applying the block inversion formula to $\mathcal{P}$, we obtain:

$$\mathcal{P}^{-1} = \begin{bmatrix}\mathcal{P}_{11}^{-1} + \mathcal{P}_{11}^{-1}\mathcal{P}_{12}\mathcal{T}\mathcal{P}_{21}\mathcal{P}_{11}^{-1} & -\mathcal{P}_{11}^{-1}\mathcal{P}_{12}\mathcal{T} \\ -\mathcal{T}\mathcal{P}_{21}\mathcal{P}_{11}^{-1} & \mathcal{T}\end{bmatrix}, \tag{120}$$

with $\mathcal{T}$ defined in (113). Replacing (120) into (110), and using (22) and (25), we conclude (112).

Our goal now is to show that

$$\lim_{\mu\to 0} \frac{\|\mathcal{W}^o_\eta - \mathcal{W}_\infty\|}{\mu} = c, \tag{121}$$

for some constant c that may depend on η (the regularization strength), but not on μ (the step-size parameter). From (112), we have:

$$\lim_{\mu\to 0} \frac{\|\mathcal{W}^o_\eta - \mathcal{W}_\infty\|}{\mu} = \eta^2 \lim_{\mu\to 0}\left\|\begin{bmatrix} -\mathcal{P}_{11}^{-1}\mathcal{P}_{12}\mathcal{T} \\ \mathcal{T}\end{bmatrix}(\Lambda_o^2 \otimes I_M)\mathcal{K}[\overline{\mathcal{W}}^o]_{2:N}\right\|. \tag{122}$$

Since the Euclidean norm is continuous, we have $\lim_{\mu\to 0}\|g(\mu)\| = \|\lim_{\mu\to 0} g(\mu)\|$. In the following we show that

$$\eta^4\left\|\lim_{\mu\to 0}\mathcal{T}(\Lambda_o^2 \otimes I_M)\mathcal{K}[\overline{\mathcal{W}}^o]_{2:N}\right\|^2 \le O(\eta^4)(O(1) + O(\eta))^{-4}, \tag{123}$$

and

$$\eta^4\left\|\lim_{\mu\to 0}\mathcal{P}_{11}^{-1}\mathcal{P}_{12}\mathcal{T}(\Lambda_o^2 \otimes I_M)\mathcal{K}[\overline{\mathcal{W}}^o]_{2:N}\right\|^2 \le O(\eta^4)(O(1) + O(\eta))^{-4}. \tag{124}$$

From (122), (123), and (124), we can conclude (53).

Let us first establish (123). We have:

$$\|\mathcal{T}(\Lambda_o^2 \otimes I_M)\mathcal{K}[\overline{\mathcal{W}}^o]_{2:N}\|^2 \le \|\mathcal{T}\|^2\|\Lambda_o^2\|^2\|\mathcal{K}\|^2\|[\overline{\mathcal{W}}^o]_{2:N}\|^2. \tag{125}$$

From (31), we have $\|\mathcal{K}\|^2 \le (O(1) + O(\eta))^{-2}$. For sufficiently small step-sizes, we have:

$$\lim_{\mu\to 0}\mathcal{P}_{21} = (V_R^\top \otimes I_M)\mathcal{H}_\infty(v_1 \otimes I_M), \tag{126}$$

$$\lim_{\mu\to 0}\mathcal{P}_{22} = \eta\Lambda_o \otimes I_M + (V_R^\top \otimes I_M)\mathcal{H}_\infty(V_R \otimes I_M). \tag{127}$$

Following the same line of reasoning as in (91)–(99), we can show that, when $\mu \to 0$, we have:

$$\|\mathcal{T}\|^2 \leq \left(\eta\lambda_2(L) + \min_{1\leq k\leq N} \lambda_{k,\min}\right)^{-2} = (O(1) + O(\eta))^{-2}. \tag{128}$$

Thus, we conclude (123).

Now, we establish (124). From (119), we have $\mathcal{P}_{11} = O(1)$ and $\|\mathcal{P}_{11}^{-1}\|^2 = O(1)$. Similarly, we can conclude from (115) that $\|\mathcal{P}_{12}\|^2 \leq O(1)$. Thus, using (123), we arrive at (124).

APPENDIX D PROOF OF THEOREM 2

From (6), (36), and (45), we have:

$$\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i = (I_{MN} - \mu\eta\mathcal{L})\,(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_{i-1} -$$
$$\mu\ \operatorname{col}\left\{\nabla_{w_k} J_k(w_{k,\infty}) - \nabla_{w_k} J_k(\boldsymbol{w}_{k,i-1}) + \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\}_{k=1}^N\Big). \tag{129}$$

Using the mean-value theorem (102), the above relation can be written as:

$$\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i = (I_{MN} - \mu\eta\mathcal{L})$$
$$\cdot\left((I_{MN} - \mu\boldsymbol{\mathcal{H}}_{i-1})(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_{i-1}) - \mu\operatorname{col}\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\}_{k=1}^N\right), \tag{130}$$

where $\boldsymbol{\mathcal{H}}_{i-1} \triangleq \operatorname{diag}\{\boldsymbol{H}_{1,i-1}, \ldots, \boldsymbol{H}_{N,i-1}\}$ with:

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla^2_{w_k} J_k(w_{k,\infty} - t(w_{k,\infty} - \boldsymbol{w}_{k,i-1}))dt \tag{131}$$

Let

$$\boldsymbol{\phi}_i \triangleq (I_{MN} - \mu\boldsymbol{\mathcal{H}}_{i-1})(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_{i-1}) - \mu\operatorname{col}\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\}_{k=1}^N \tag{132}$$

$$C \triangleq I_N - \mu\eta L. \tag{133}$$

From the Laplacian matrix definition, it can be verified that the off-diagonal entries of the matrix C are non-negative and that its diagonal entries are non-negative under condition (38). Furthermore, since we have $L\mathbb{1}_N = 0$, the entries on each row of C will add up to one. Thus, applying Jensen's inequality [41, pp. 77] to the convex function $\|\cdot\|^2$, we obtain from (130) and (132):

$$\mathbb{E}\|w_{k,\infty} - \boldsymbol{w}_{k,i}\|^2 \leq \sum_{\ell=1}^N [C]_{k\ell}\mathbb{E}\|\boldsymbol{\phi}_{\ell,i}\|^2, \tag{134}$$

where $\boldsymbol{\phi}_{k,i}$ is the k-th sub-vector of $\boldsymbol{\phi}_i$ given by:

$$\boldsymbol{\phi}_{k,i} = (I_M - \mu\boldsymbol{H}_{k,i-1})(w_{k,\infty} - \boldsymbol{w}_{k,i-1}) - \mu\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}). \tag{135}$$

Squaring both sides of (135), conditioning on $\boldsymbol{\mathcal{F}}_{i-1}$, and taking expectations we obtain:

$$\mathbb{E}[\|\boldsymbol{\phi}_{k,i}\|^2|\boldsymbol{\mathcal{F}}_{i-1}] = \|w_{k,\infty} - \boldsymbol{w}_{k,i-1}\|^2_{\boldsymbol{\Sigma}_{k,i-1}} +$$
$$\mu^2\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}]. \tag{136}$$

where $\boldsymbol{\Sigma}_{k,i-1} \triangleq (I_M - \mu\boldsymbol{H}_{k,i-1})^2$ and where the cross term is zero because of the zero-mean condition (33). Due to Assumption 1, $\boldsymbol{\Sigma}_{k,i-1}$ can be bounded as follows:

$$0 < \boldsymbol{\Sigma}_{k,i-1} \leq \gamma_k^2 I_M, \tag{137}$$

where γ_k is given by (42). From Assumption 2, $\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}]$ can be bounded as follows:

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \\
&\quad\leq \beta_k^2\|\boldsymbol{w}_{k,i-1}\|^2 + \sigma_{s,k}^2 \\
&\quad= \beta_k^2\|w^o_{k,\eta} - w_{k,\infty} + w_{k,\infty} - \boldsymbol{w}_{k,i-1} - w^o_{k,\eta}\|^2 + \sigma_{s,k}^2 \\
&\quad\leq 3\beta_k^2\|w^o_{k,\eta} - w_{k,\infty}\|^2 + 3\beta_k^2\|w_{k,\infty} - \boldsymbol{w}_{k,i-1}\|^2 \\
&\qquad + 3\beta_k^2\|w^o_{k,\eta}\|^2 + \sigma_{s,k}^2.
\end{aligned} \tag{138}$$

Taking expectation again in (136), and using the bounds (137) and (138), we obtain:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\phi}_{k,i}\|^2 &= \mathbb{E}\|w_{k,\infty} - \boldsymbol{w}_{k,i-1}\|^2_{\boldsymbol{\Sigma}_{k,i-1}} + \mu^2\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2 \\
&\leq (\gamma_k^2 + 3\mu^2\beta_k^2)\mathbb{E}\|w_{k,\infty} - \boldsymbol{w}_{k,i-1}\|^2 \\
&\quad + \mu^2\Big(3\beta_k^2\|w^o_{k,\eta} - w_{k,\infty}\|^2 + 3\beta_k^2\|w^o_{k,\eta}\|^2 + \sigma_{s,k}^2\Big).
\end{aligned} \tag{139}$$

Now, combining (139) and (134), we obtain (55).

Iterating (55) starting from $i = 1$, we get:

$$\mathrm{MSP}_i \preceq (CG)^i\mathrm{MSP}_0 + \mu^2\sum_{j=0}^{i-1}(CG)^jCb. \tag{140}$$

Under Assumption 3 and condition (58), the matrix CG can be guaranteed to be stable. To see this, we upper bound the spectral radius as follows:

$$\begin{aligned}
\rho(CG) &\leq \|CG\|_\infty \leq \|C\|_\infty\|G\|_\infty \\
&= \|G\|_\infty = \max_{1\leq k\leq N} \gamma_k^2 + 3\mu^2\beta_k^2,
\end{aligned} \tag{141}$$

where we used the fact that, under condition (38), the matrix C is a right-stochastic matrix. We have:

$$\begin{aligned}
\gamma_k^2 + 3\mu^2\beta_k^2 = \max\{&1 - 2\mu\lambda_{k,\min} + \mu^2\lambda^2_{k,\min} + 3\mu^2\beta_k^2, \\
&1 - 2\mu\lambda_{k,\max} + \mu^2\lambda^2_{k,\max} + 3\mu^2\beta_k^2\},
\end{aligned} \tag{142}$$

which is guaranteed to be less than one when:

$$0 < \mu < \min\left\{\frac{2\lambda_{k,\min}}{\lambda^2_{k,\min} + 3\beta_k^2}, \frac{2\lambda_{k,\max}}{\lambda^2_{k,\max} + 3\beta_k^2}\right\}. \tag{143}$$

Then we conclude that the matrix CG is stable under condition (58). In this case, we have:

$$\limsup_{i\to\infty} \mathrm{MSP}_i \preceq \mu^2\sum_{j=0}^{\infty}(CG)^jCb. \tag{144}$$

Using the submultiplicative property of the induced infinity norm, we obtain:

$$\begin{aligned}\|\limsup_{i\to\infty}\mathrm{MSP}_i\|_\infty &\le \mu^2\left\|\sum_{j=0}^{\infty}(CG)^j\right\|_\infty\|C\|_\infty\|b\|_\infty\\ &\le \mu^2\sum_{j=0}^{\infty}\|(CG)^j\|_\infty\|b\|_\infty,\\ &\le \mu^2\sum_{j=0}^{\infty}\|C\|_\infty^j\|G\|_\infty^j\|b\|_\infty=\frac{\mu^2\|b\|_\infty}{1-\|G\|_\infty},\end{aligned}\tag{145}$$

where we used the fact that $\|C\|_\infty=1$ and where $\|G\|_\infty=\max_{1\le k\le N}\gamma_k^2+3\mu^2\beta_k^2$. From (142), we have:

$$\gamma_k^2+3\mu^2\beta_k^2=1-\mu\zeta_k,\tag{146}$$

where

$$\begin{aligned}\zeta_k\triangleq\min\{&2\lambda_{k,\min}-\mu\lambda_{k,\min}^2-3\mu\beta_k^2,\\ &2\lambda_{k,\max}-\mu\lambda_{k,\max}^2-3\mu\beta_k^2\}.\end{aligned}\tag{147}$$

Thus,

$$\|G\|_\infty=\max_{1\le k\le N}\{1-\mu\zeta_k\}=1-\mu\min_{1\le k\le N}\zeta_k.\tag{148}$$

Substituting into (145), we obtain:

$$\|\limsup_{i\to\infty}\mathrm{MSP}_i\|_\infty\le\frac{\mu\|b\|_\infty}{\min_{1\le k\le N}\zeta_k}.\tag{149}$$

For sufficiently small μ, we have from (57) and Theorem 1 that $\|b\|=O(1)+O(\mu^2\eta^4)(O(1)+O(\eta))^{-4}$. We conclude that $\|\limsup_{i\to\infty}\mathrm{MSP}_i\|_\infty\le O(\mu)$.

From (44), we have:

$$\begin{aligned}\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}_\eta^o-\boldsymbol{\mathcal{W}}_i\|^2&\le 2\|\mathcal{W}_\eta^o-\mathcal{W}_\infty\|^2\\ &+2\limsup_{i\to\infty}\mathbb{1}_N^\top\cdot\mathrm{MSP}_i\end{aligned}\tag{150}$$

Therefore, from Theorem 1 and (59), we conclude (60).

APPENDIX E PROOF OF THEOREM 3

Applying Jensen's inequality [41, pp. 77] to the convex function $\|\cdot\|^4$, we obtain from (130) and (132):

$$\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i}\|^4\le\sum_{\ell=1}^{N}[C]_{k\ell}\mathbb{E}\|\boldsymbol{\phi}_{\ell,i}\|^4,\tag{151}$$

where C and $\boldsymbol{\phi}_{k,i}$ are given by (133) and (135), respectively. Using the inequality [7, pp. 523]:

$$\|a+b\|^4\le\|a\|^4+3\|b\|^4+8\|a\|^2\|b\|^2+4\|a\|^2(a^\top b),\tag{152}$$

we obtain from (135) under Assumption 2 on the gradient noise:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\phi}_{k,i}\|^4&\le\mathbb{E}\|(I_M-\mu\boldsymbol{H}_{k,i-1})(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|^4\\ &+3\mu^4\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4+8\mu^2\cdot\\ &\left(\mathbb{E}\|(I_M-\mu\boldsymbol{H}_{k,i-1})(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|^2\right)\left(\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2\right)\end{aligned}\tag{153}$$

From Assumption 1, the matrices $(I_M-\mu\boldsymbol{H}_{k,i-1})^2$ and $(I_M-\mu\boldsymbol{H}_{k,i-1})^4$ can be bounded as follows:

$$0<(I_M-\mu\boldsymbol{H}_{k,i-1})^2\le\gamma_k^2I_M,\tag{154}$$

$$0<(I_M-\mu\boldsymbol{H}_{k,i-1})^4\le\gamma_k^4I_M,\tag{155}$$

where γ_k is given by (42). Thus, we obtain:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{\phi}_{k,i}\|^4&\le\gamma_k^4\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^4+3\mu^4\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4\\ &+8\mu^2\gamma_k^2\left(\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\right)\left(\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2\right).\end{aligned}\tag{156}$$

Under condition (62), we have:

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right]\\ &\le\overline{\beta}_k^4\|\boldsymbol{w}_{k,i-1}\|^4+\overline{\sigma}_{s,k}^4\\ &=\overline{\beta}_k^4\|w_{k,\eta}^o-w_{k,\infty}+w_{k,\infty}-\boldsymbol{w}_{k,i-1}-w_{k,\eta}^o\|^4+\overline{\sigma}_{s,k}^4\\ &\le 27\overline{\beta}_k^4\|w_{k,\eta}^o-w_{k,\infty}\|^4+27\overline{\beta}_k^4\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^4\\ &+27\overline{\beta}_k^4\|w_{k,\eta}^o\|^4+\overline{\sigma}_{s,k}^4,\end{aligned}\tag{157}$$

where we applied Jensen's inequality to the function $\|\cdot\|^4$. Furthermore, from (138), the last term on the RHS of (156) can be bounded as follows:

$$\begin{aligned}&\left(\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\right)\left(\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2\right)\\ &\le 3\beta_k^2\left(\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\right)^2\\ &+(3\beta_k^2\|w_{k,\eta}^o-w_{k,\infty}\|^2+3\beta_k^2\|w_{k,\eta}^o\|^2+\sigma_{s,k}^2)\\ &\cdot\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\\ &\le 3\beta_k^2\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^4\\ &+(3\beta_k^2\|w_{k,\eta}^o-w_{k,\infty}\|^2+3\beta_k^2\|w_{k,\eta}^o\|^2+\sigma_{s,k}^2)\\ &\cdot\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\end{aligned}\tag{158}$$

where we used the fact that for any random variable $\boldsymbol{a}$, we have $(\mathbb{E}\boldsymbol{a})^2\le\mathbb{E}\boldsymbol{a}^2$. Replacing (157) and (158) into (156), we obtain:

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{\phi}_{k,i}\|^4\le\\ &(\gamma_k^4+81\mu^4\overline{\beta}_k^4+24\mu^2\gamma_k^2\beta_k^2)\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^4+8\mu^2\gamma_k^2\cdot\\ &(3\beta_k^2\|w_{k,\eta}^o-w_{k,\infty}\|^2+3\beta_k^2\|w_{k,\eta}^o\|^2+\sigma_{s,k}^2)\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\\ &+81\mu^4\overline{\beta}_k^4\|w_{k,\eta}^o-w_{k,\infty}\|^4+81\mu^4\overline{\beta}_k^4\|w_{k,\eta}^o\|^4+3\mu^4\overline{\sigma}_{s,k}^4.\end{aligned}\tag{159}$$

Now, combining (159) and (151), we arrive at (66).

Iterating (66) starting from $i = 1$, we get:

$$\begin{aligned}\mathrm{MFP}_i &\preceq (CG')^i \mathrm{MFP}_0 \\ &\quad + \mu^2 \sum_{j=0}^{i-1} (CG')^j CB\mathrm{MSP}_{i-1-j} + \mu^4 \sum_{j=0}^{i-1} (CG')^j Cb'. \end{aligned} \tag{160}$$

Under Assumption 3 and for sufficiently small μ, the matrix CG' can be guaranteed to be stable. To see this, we upper bound its spectral radius as follows:

$$\rho(CG') \le \|CG'\|_\infty \le \|C\|_\infty \|G'\|_\infty = \|G'\|_\infty, \tag{161}$$

since under condition (38), C is a right-stochastic matrix. The $\infty-$norm of G' is given by:

$$\begin{aligned}&\|G'\|_\infty \\ &= \max_{1\le k\le N} \left\{\gamma_k^4 + 24\mu^2\gamma_k^2\beta_k^2 + 81\mu^4\overline{\beta}_k^4\right\} \\ &= \max_{1\le k\le N} \big\{\max\{1 - 4\mu\lambda_{k,\min} + 6\mu^2\lambda_{k,\min}^2 - 4\mu^3\lambda_{k,\min}^3 \\ &\quad + \mu^4\lambda_{k,\min}^4 + 24\mu^2\gamma_k^2\beta_k^2 + 81\mu^4\overline{\beta}_k^4, \\ &\quad 1 - 4\mu\lambda_{k,\max} + 6\mu^2\lambda_{k,\max}^2 - 4\mu^3\lambda_{k,\max}^3 \\ &\quad + \mu^4\lambda_{k,\max}^4 + 24\mu^2\gamma_k^2\beta_k^2 + 81\mu^4\overline{\beta}_k^4\}\big\} \\ &= 1 - \mu \min_{1\le k\le N} \big\{\min\{4\lambda_{k,\min} - 6\mu\lambda_{k,\min}^2 + 4\mu^2\lambda_{k,\min}^3 \\ &\quad - \mu^3\lambda_{k,\min}^4 - 24\mu\gamma_k^2\beta_k^2 - 81\mu^3\overline{\beta}_k^4, \\ &\quad 4\lambda_{k,\max} - 6\mu\lambda_{k,\max}^2 + 4\mu^2\lambda_{k,\max}^3 \\ &\quad - \mu^3\lambda_{k,\max}^4 - 24\mu\gamma_k^2\beta_k^2 - 81\mu^3\overline{\beta}_k^4\}\big\}. \end{aligned} \tag{162}$$

A sufficiently small μ ensures $\|G'\|_\infty < 1$ and, thus, ensures the stability of CG'.

We have established in Theorem 2 that, for small μ, after sufficient iterations have passed, MSP_j converges to a bounded region on the order of μ. This implies that, there exists a j_o large enough such that for all $j \ge j_o$ it holds that:

$$\|\mathrm{MSP}_j\|_\infty \le s_{\max} = O(\mu). \tag{163}$$

In this case, we have from (160):

$$\begin{aligned}&\limsup_{i\to\infty} \mathrm{MFP}_i \\ &\preceq \mu^4 \sum_{j=0}^{\infty} (CG')^j Cb' + \mu^2 \limsup_{i\to\infty} \sum_{j=0}^{i-1} (CG')^j CB\mathrm{MSP}_{i-1-j} \\ &= \mu^4 \sum_{j=0}^{\infty} (CG')^j Cb' + \mu^2 \limsup_{i\to\infty} \sum_{j=0}^{i-1} (CG')^{i-1-j} CB\mathrm{MSP}_j \\ &= \mu^4 \sum_{j=0}^{\infty} (CG')^j Cb' + \mu^2 \limsup_{i\to\infty} \left(\sum_{j=0}^{j_o} (CG')^{i-1-j} CB\mathrm{MSP}_j \right. \\ &\qquad \left. + \sum_{j=j_o+1}^{i-1} (CG')^{i-1-j} CB\mathrm{MSP}_j \right) \\ &= \mu^4 \sum_{j=0}^{\infty} (CG')^j Cb' + \mu^2 \limsup_{i\to\infty} \left((CG')^i \sum_{j=0}^{j_o} (CG')^{-1-j} CB\mathrm{MSP}_j \right. \\ &\qquad \left. + \sum_{j=j_o+1}^{i-1} (CG')^{i-1-j} CB\mathrm{MSP}_j \right) \\ &= \mu^4 \sum_{j=0}^{\infty} (CG')^j Cb' + \mu^2 \limsup_{i\to\infty} \sum_{j=j_o+1}^{i-1} (CG')^{i-1-j} CB\mathrm{MSP}_j \end{aligned} \tag{164}$$

Using the submultiplicative and sub-additive properties of the induced infinity norm, we obtain:

$$\begin{aligned}&\|\limsup_{i\to\infty} \mathrm{MFP}_i\|_\infty \\ &\le \mu^4 \left\| \sum_{j=0}^{\infty} (CG')^j \right\|_\infty \|b'\|_\infty \\ &\quad + \mu^2 \limsup_{i\to\infty} \sum_{j=j_o+1}^{i-1} \|(CG')^{i-1-j} CB\mathrm{MSP}_j\|_\infty \\ &\le \mu^4 \sum_{j=0}^{\infty} \|G'\|_\infty^j \|b'\|_\infty + \mu^2 \limsup_{i\to\infty} \sum_{j=0}^{i-j_o-2} \|G'\|_\infty^j \|B\|_\infty s_{\max} \\ &= \mu^4 \frac{\|b'\|_\infty}{1 - \|G'\|_\infty} + \mu^2 \frac{\|B\|_\infty s_{\max}}{1 - \|G'\|_\infty} \end{aligned} \tag{165}$$

where in the second line we used (163) and where $\|G'\|_\infty$ is given by (162). Since $\|b'\|_\infty = O(1)$, $\|B\|_\infty = O(1)$, $s_{\max} = O(\mu)$, and $1 - \|G'\|_\infty = O(\mu)$, we conclude (70).

From (63) and (64), we have:

$$\begin{aligned}\limsup_{i\to\infty} \mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^4 &\le 8\|\mathcal{W}_\eta^o - \mathcal{W}_\infty\|^4 \\ &\quad + 8N \limsup_{i\to\infty} \mathbb{1}_N \cdot \mathrm{MFP}_i. \end{aligned} \tag{166}$$

Therefore, from Theorem 1 and (70), we conclude (71).

APPENDIX F PROOF OF THEOREM 4

Conditioning both sides of (130), invoking the conditions on the gradient noise from Assumption 2, and computing the conditional expectations, we obtain:

$$\begin{aligned}&\mathbb{E}[(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i)|\boldsymbol{\mathcal{F}}_{i-1}] \\ &= (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\boldsymbol{\mathcal{H}}_{i-1})(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_{i-1}). \end{aligned} \tag{167}$$

Taking expectations again, we arrive at:

$$\begin{aligned}&\mathbb{E}(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_i) = (I_{MN} - \mu\eta\mathcal{L}) \\ &\cdot \mathbb{E}[(I_{MN} - \mu\boldsymbol{\mathcal{H}}_{i-1})(\mathcal{W}_\infty - \boldsymbol{\mathcal{W}}_{i-1})]. \end{aligned} \tag{168}$$

Applying Jensen's inequality [41, pp. 77] to the convex function $\|\cdot\|^2$, we obtain from the above relation:

$$\|\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i})\|^2 \le \sum_{\ell=1}^{N}[C]_{k\ell}\left\|\mathbb{E}\left[(I_M-\mu\boldsymbol{H}_{\ell,i-1})(w_{\ell,\infty}-\boldsymbol{w}_{\ell,i-1})\right]\right\|^2, \tag{169}$$

where C and $\boldsymbol{H}_{k,i-1}$ are given by (133) and (131), respectively. Let

$$\widetilde{\boldsymbol{H}}_{k,i-1} \triangleq H_{k,\eta}-\boldsymbol{H}_{k,i-1}, \tag{170}$$

where

$$H_{k,\eta}=\nabla^2_{w_k}J_k(w^o_{k,\eta}). \tag{171}$$

Then, we can write:

$$\mathbb{E}\left[(I_M-\mu\boldsymbol{H}_{k,i-1})(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\right] = (I_M-\mu H_{k,\eta})\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})+\mu\boldsymbol{c}_{k,i-1}, \tag{172}$$

in terms of a deterministic perturbation sequence defined by

$$\boldsymbol{c}_{k,i-1} \triangleq \mathbb{E}[\widetilde{\boldsymbol{H}}_{k,i-1}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})]. \tag{173}$$

By applying Jensen's inequality to the convex function $\|\cdot\|^2$, we obtain:

$$\begin{aligned}
&\|\mathbb{E}(I_M-\mu\boldsymbol{H}_{k,i-1})(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|^2\\
&=\|(I_M-\mu H_{k,\eta})\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})+\mu\boldsymbol{c}_{k,i-1}\|^2\\
&=\left\|t\frac{1}{t}(I_M-\mu H_{k,\eta})\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\right.\\
&\quad\left.+\mu(1-t)\frac{1}{1-t}\boldsymbol{c}_{k,i-1}\right\|^2\\
&\le t\left\|\frac{1}{t}(I_M-\mu H_{k,\eta})\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\right\|^2\\
&\quad+\mu^2(1-t)\left\|\frac{1}{1-t}\boldsymbol{c}_{k,i-1}\right\|^2\\
&=\frac{1}{t}\|(I_M-\mu H_{k,\eta})\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|^2+\mu^2\frac{1}{1-t}\|\boldsymbol{c}_{k,i-1}\|^2
\end{aligned} \tag{174}$$

for any arbitrary positive number $t\in(0,1)$. We select $t=\gamma_k$ where γ_k is given by (42), which is guaranteed to be less than one under condition (43). From Assumption 1, we have $\|I_M-\mu H_{k,\eta}\|^2\le\gamma_k^2$. Thus, we obtain:

$$\|\mathbb{E}(I_M-\mu\boldsymbol{H}_{k,i-1})(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|^2 \le \gamma_k\|\mathbb{E}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|^2+\mu^2\frac{1}{1-\gamma_k}\|\boldsymbol{c}_{k,i-1}\|^2. \tag{175}$$

As shown in [7, Appendix E], the Hessian of a twice differentiable strongly convex function $J_k(w_k)$ satisfying Assumptions 1 and 5 is globally Lipschitz relative to $w^o_{k,\eta}$, namely, it satisfies:

$$\|\nabla^2_{w_k}J_k(w_k)-\nabla^2_{w_k}J_k(w^o_{k,\eta})\|\le\kappa'_d\|w_k-w^o_{k,\eta}\|,\quad\forall w_k, \tag{176}$$

where $\kappa'_d=\max\{\kappa_d,\frac{\lambda_{k,\max}-\lambda_{k,\min}}{\epsilon}\}$. Then, for each agent k we obtain:

$$\begin{aligned}
&\|\widetilde{\boldsymbol{H}}_{k,i-1}\|\\
&\triangleq\|H_{k,\eta}-\boldsymbol{H}_{k,i-1}\|\\
&\le\int_0^1\left\|\nabla^2_{w_k}J_k(w^o_{k,\eta})-\nabla^2_{w_k}J_k(w_{k,\infty}-t(w_{k,\infty}-\boldsymbol{w}_{k,i-1}))\right\|dt\\
&\le\int_0^1\kappa'_d\|w^o_{k,\eta}-w_{k,\infty}+t(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|dt\\
&\le\int_0^1\kappa'_d\|w^o_{k,\eta}-w_{k,\infty}\|dt+\int_0^1\kappa'_d\|t(w_{k,\infty}-\boldsymbol{w}_{k,i-1})\|dt\\
&=\kappa'_d\|w^o_{k,\eta}-w_{k,\infty}\|+\frac{1}{2}\kappa'_d\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|,
\end{aligned} \tag{177}$$

and, hence,

$$\begin{aligned}
\|\boldsymbol{c}_{k,i-1}\|&=\|\mathbb{E}[\widetilde{\boldsymbol{H}}_{k,i-1}(w_{k,\infty}-\boldsymbol{w}_{k,i-1})]\|\\
&\le\mathbb{E}[\|\widetilde{\boldsymbol{H}}_{k,i-1}\|\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|]\\
&\le\kappa'_d\|w^o_{k,\eta}-w_{k,\infty}\|\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|\\
&\quad+\frac{1}{2}\kappa'_d\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2.
\end{aligned} \tag{178}$$

where we used the stochastic version of Jensen's inequality:

$$f(\mathbb{E}\boldsymbol{a})\le\mathbb{E}(f(\boldsymbol{a})) \tag{179}$$

when $f(x)\in\mathbb{R}$ is convex. Applying Jensen's inequality to the convex function $\|\cdot\|^2$ and using the fact that $(\mathbb{E}\boldsymbol{a})^2\le\mathbb{E}\boldsymbol{a}^2$ for any real-valued random variable $\boldsymbol{a}$, we obtain from (178):

$$\begin{aligned}
\|\boldsymbol{c}_{k,i-1}\|^2&\le 2(\kappa'_d)^2\|w^o_{k,\eta}-w_{k,\infty}\|^2\left(\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|\right)^2\\
&\quad+2\frac{1}{4}(\kappa'_d)^2\left(\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\right)^2\\
&\le 2(\kappa'_d)^2\|w^o_{k,\eta}-w_{k,\infty}\|^2\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\\
&\quad+\frac{1}{2}(\kappa'_d)^2\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^4.
\end{aligned} \tag{180}$$

From (169) and using the above bound in (175), we conclude (75).

Iterating (75) starting from $i=1$, we obtain:

$$\begin{aligned}
\mathrm{SMP}_i&\le(CG'')^i\mathrm{SMP}_0\\
&\quad+\mu^2\frac{1}{2}(\kappa'_d)^2\sum_{j=0}^{i-1}(CG'')^jC(I-G'')^{-1}\mathrm{MFP}_{i-1-j}+\\
&\quad\mu^2\sum_{j=0}^{i-1}(CG'')^jC(I-G'')^{-1}B'\mathrm{MSP}_{i-1-j}.
\end{aligned} \tag{181}$$

Under Assumption 3 and condition (43), the matrix CG'' is guaranteed to be stable. From (59), (70), and following similar arguments as the ones used to establish (70) in Appendix E, we conclude that

$$\|\limsup_{i\to\infty} \text{SMP}_i\|_\infty = O(\mu^2) + \frac{O(\mu^3\eta^4)}{(O(1)+O(\eta))^4} = O(\mu^2), \tag{182}$$

where we used the fact that $\|B'\|_\infty \leq O(\mu^2\eta^4)/(O(1)+O(\eta))^4$ from Theorem 1 and $\|(I_N - G'')^{-1}\|_\infty \leq O(\mu^{-1})$.

Using (73) and since $\|\mathbb{E}(\mathcal{W}_\infty - \mathcal{W}_i)\|^2 = \mathbb{1}_N \cdot \text{SMP}_i$, we conclude (79) from Theorem 1 and (78).

REFERENCES

[1] R. Nassif, S. Vlaski, and A. H. Sayed, "Distributed inference over multitask graphs under smoothness," in *Proc. IEEE Int. Workshop Signal Process. Advances Wireless Commun.*, Kalamata, Greece, Jun. 2018.

[2] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optim.*, vol. 7, no. 4, pp. 913–926, 1997.

[3] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[4] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[5] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010.

[6] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 205–220, Apr. 2013.

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

[8] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks—Part I: Transient analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3487–3517, Jun. 2015.

[9] J. Chen and A. H. Sayed, "On the learning behavior of adaptive networks—Part II: Performance analysis," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3518–3548, Jun. 2015.

[10] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[11] S. Vlaski, L. Vandenberghe, and A. H. Sayed, "Diffusion stochastic optimization with non-smooth regularizers," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 4149–4153.

[12] J. Plata-Chaves, A. Bertrand, M. Moonen, S. Theodoridis, and A. M. Zoubir, "Heterogeneous and multitask wireless sensor networks—Algorithms, applications, and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 450–465, Apr. 2017.

[13] 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.

[14] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Proximal multitask learning over networks with sparsity-inducing coregularization," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6329–6344, Dec. 2016.

[15] X. Cao and K. J. R. Liu, "Decentralized sparse multitask RLS over networks," *IEEE Trans. Signal Process.*, vol. 65, no. 23, pp. 6217–6232, Dec. 2017.

[16] C. Eksin and A. Ribeiro, "Distributed network optimization with heuristic rational agents," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5396–5411, Oct. 2012.

[17] D. Hallac, J. Leskovec, and S. Boyd, "Network Lasso: Clustering and optimization in large graphs," in *Proc. ACM SIGKDD*, Sydney, Australia, Aug. 2015, pp. 387–396.

[18] C. Li, S. Huang, Y. Liu, and Y. Liu, "Distributed TLS over multitask networks with adaptive intertask cooperation," *IEEE Trans. Aerospace Electron. Syst.*, vol. 52, no. 6, pp. 3036–3052, Dec. 2016.

[19] V. Kekatos and G. B. Giannakis, "Distributed robust power system state estimation," *IEEE Trans. Signal Process.*, vol. 28, no. 2, pp. 1617–1626, May 2013.

[20] J. Plata-Chaves, N. Bogdanović, and K. Berberidis, "Distributed diffusion-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3448–3460, Jul. 2015.

[21] S. A. Alghunaim, K. Yuan, and A. H. Sayed, "Decentralized exact coupled optimization," in *Proc. Ann. Allerton Conf. Commun., Control, Comput.*, 2017, pp. 338–345.

[22] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Diffusion LMS for multitask problems with local linear equality constraints," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 4979–4993, Oct. 2017.

[23] J. Chen, C. Richard, A. O. Hero, and A. H. Sayed, "Diffusion LMS for multitask problems with overlapping hypothesis subspaces," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Reims, France, Sep. 2014, pp. 1–6.

[24] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "Learning over multitask graphs—Part II: Performance analysis," Nov. 2019, *arXiv:1805.08547*.

[25] J. Tuck, S. Barratt, and S. Boyd, "A distributed method for fitting Laplacian regularized stratified models," Apr. 2019, *arXiv:1904.12017*.

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

[27] 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.

[28] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[29] K. Q. Weinberger, F. Sha, Q. Zhu, and L. K. Saul, "Graph Laplacian regularization for large-scale semidefinite programming," in *Proc. Advances Neural Inf. Process. Syst.*, Dec. 2007, pp. 1489–1496.

[30] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, May 2018.

[31] 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.

[32] 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.

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

[34] D. I. Shuman, P. Vandergheynst, and P. Frossard, "Chebyshev polynomial approximation for distributed signal processing," in *Proc. IEEE Int. Conf. Dist. Comp. Sensor Syst.*, 2011, pp. 1–8.

[35] B. T. Polyak, *Introduction to Optimization*. New York, NY, USA: Optimization Software, 1987.

[36] E. Kreyszig, *Introductory Functional Analysis With Applications*, Hoboken, NJ, USA: Wiley, 1989.

[37] J. H. Lawrimore, M. J. Menne, B. E. Gleason, C. N. Williams, and D. B. Wuertz, *Global Historical Climatology Network–Monthly (GHCN-M)*, NOAA National Climatic Data Center. [Online]. Available: ftp:/ftp.ncdc.noaa.gov/pub/data/gsod

[38] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. Hoboken, NJ, USA: Wiley, 2000.

[39] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. New York, NY, USA: Academic, 2008.

[40] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1980.

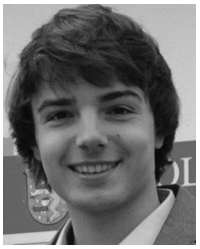
[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[42] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[43] R. L. Smith, "Some interlacing properties of the Schur complement of a Hermitian matrix," *Linear Algebra Appl.*, vol. 177, pp. 137–144, 1992.



ROULA NASSIF (Member, IEEE) received the Dipl.-Ing. degree from Lebanese University, Lebanon, and the M.S. degrees from the Compiègne University of Technology, France, in 2013, both in industrial control and electrical engineering, and the Ph.D. degree from the University of Côte d'Azur, France, in 2016. She is an Assistant Professor with the American University of Beirutc (AUB), Lebanon. Prior to joining AUB, she was a Postdoctoral Scholar with Adaptive Systems Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland. Her current research interests include distributed inference over multitask networks and graph signal processing.



STEFAN VLASKI (Member, IEEE) received the B.Sc. degree in electrical engineering from Technical University Darmstadt, Germany, in 2013, and the M.S. degree in electrical engineering and the Ph.D. degree in electrical and computer engineering from the University of California, Los Angeles, in 2014 and 2019, respectively. He is currently a Postdoctoral Researcher with the Adaptive Systems Laboratory, EPFL, Switzerland. His research interests are in machine learning, signal processing, and optimization. His current focus is on the development and study of learning algorithms with a particular emphasis on adaptive and decentralized solutions. He was a recipient of the German National Scholarship at TU Darmstadt and the Graduate Division Fellowship at UCLA. His work was recognized as Best Student Paper in the area "Signal Processing Theory and Methods" at IEEE ICASSP 2016 and received the 2nd Place Award at IEEE CAMSAP 2019.



CÉDRIC RICHARD (Senior Member, IEEE) received the Dipl.-Ing., and master's and Ph.D. degree in electrical and computer engineering from the Compiègne University of Technology, France, in 1994 and 1998, respectively. He is a Full Professor with the Université Côte d'Azur, France.

His current research interests include statistical signal processing and machine learning. He is the author of over 300 papers.

Prof. Richard is the Director-at-Large of Region 8 (Europe, Middle East, Africa) of the IEEE Signal Processing Society (IEEE-SPS), and a Member of the Board of Governors of the IEEE-SPS. He is also the Director of the French Federal CNRS Research Association ISIS (Information, Signal, Image, Vision). In 2010–2015, Prof. Richard was distinguished as a member of the Institut Universitaire de France (IUF). Since 2019, he serves as an Associate Editor of the IEEE OPEN JOURNAL OF SIGNAL PROCESSING (OJSP), and since 2009, as an Associate Editor of *Signal Processing* (Elsevier). In 2015–2018, he was a Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS. He was also an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2006–2010). He is an Elected Member of the IEEE Signal Processing Theory and Methods Technical Committee (2009–2014, and 2018–present), and was an Elected Member of the IEEE Machine Learning for Signal Processing Technical Committee (2012–2018). Prof. Richard was the General Co-Chair of the IEEE SSP'11 Workshop that was held in Nice, France. He was the Technical Co-Chair of EUSIPCO'15 that was held in Nice, France, and of the IEEE CAMSAP'15 Workshop that was held in Cancun, Mexico. He was also the Special Session Chair of the IEEE SAM'16 Workshop that was held in Rio de Janeiro, Brazil, and the Local Co-Chair of the IEEE CAMSAP'19 Workshop. He will be the General Co-Chair of EUSIPCO'20 that will take place in Amsterdam, The Netherlands.



ALI H. SAYED (Fellow, IEEE) is the Dean of Engineering at EPFL, Switzerland. He has served as a Distinguished Professor and former Chairman of electrical engineering with the University of California, Los Angeles. He is a member of the US National Academy of Engineering and recognized as a Highly-Cited author. An author of over 550 scholarly publications and six books, his research involves several areas including adaptation and learning theories, data and network sciences, and statistical inference. His work has been recognized with several awards including the 2014 Athanasios Papoulis Award from the European Association for Signal Processing, the 2015 Education Award, the 2013 Meritorious Service Award, and the 2012 Technical Achievement Award from the IEEE Signal Processing Society, the 2005 Terman Award from the American Society for Engineering Education, the 2003 Kuwait Prize, and the 1996 IEEE Donald G. Fink Prize. He served as Distinguished Lecturer for the IEEE Signal Processing Society in 2005 and as Editor-in-Chief of the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2003–2005). His articles received several Best Paper Awards from the IEEE Signal Processing Society in 2002, 2005, 2012, and 2014. He is a Fellow of the IEEE, EURASIP, and the American Association for the Advancement of Science (AAAS). He served as President of the IEEE Signal Processing Society in 2018 and 2019.