Learning over Multitask Graphs-Part II: Performance Analysis

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

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 27 May 2020. The review of this article was arranged by Associate Editor Selin Aviyente.

Digital Object Identifier 10.1109/OJSP.2020.2989031

Learning Over Multitask Graphs—Part II: Performance 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, 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 Grants CCF-1524250 and ECCS-1407712, and in part by Swiss National Science Foundation under Grant 205121-184999.

ABSTRACT Part I of this paper formulated 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. A diffusion strategy was devised that responds to streaming data and employs stochastic approximations in place of actual gradient vectors, which are generally unavailable. The approach relied on minimizing a global cost consisting of the aggregate sum of individual costs regularized by a term that promotes smoothness. We examined the first-order, the second-order, and the fourth-order stability of the multitask learning algorithm. The results identified conditions on the step-size parameter, regularization strength, and data characteristics in order to ensure stability. This Part II examines steady-state performance of the strategy. The results reveal explicitly the influence of the network topology and the regularization strength on the network performance and 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, steady-state performance.

I. INTRODUCTION

As pointed out in Part I [2] of this work, most prior literature on distributed inference over networks 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 [3]–[13]. In this paper, and its accompanying Part I [2], we focus instead on multitask networks where the agents may need to estimate and track multiple objectives simultaneously [14]–[27]. 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.

In Part I [2], we considered 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 over the topology, as dictated by the graph Laplacian matrix. The smoothness property softens the transitions in the tasks among adjacent nodes and allows incorporating information about the graph structure into the solution of the inference problem. In order to exploit the smoothness prior, we formulated 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 [28], [29]. A diffusion strategy was proposed that

responds to streaming data and employs stochastic approximations in place of actual gradient vectors, which are generally unavailable.

The analysis from Part I [2] revealed how the regularization strength η can steer the convergence point of the network toward many modes of operation starting from the non-cooperative mode ($\eta = 0$) where each agent converges to the minimizer of its individual cost and ending with the single-task mode ($\eta \to \infty$) where all agents converge to a common parameter vector corresponding to the minimizer of the aggregate sum of individual costs. 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 carried out in Part I [2] a detailed stability analysis of the proposed strategy. We showed, 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. We also established the first and fourth-order moments stability of the network error process and showed that they tend asymptotically to bounded region on the order of $O(\mu)$ and $O(\mu^2)$, respectively.

Based on the results established in Part I [2], we shall derive in this paper 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 (through the eigenvalues and eigenvectors of the Laplacian matrix), 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 is also derived. This expression will provide insights on the design of effective multitask strategies for distributed inference over networks.

Notation: We adopt the same notation from Part I [2]. 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 symbols $\otimes$ and $\otimes_b$ denote the Kronecker product and the block Kronecker product, respectively. The symbol $\mathrm{vec}(\cdot)$ refers to the standard vectorization operator that stacks the columns of a matrix on top of each other and the symbol $\mathrm{bvec}(\cdot)$ refers to the block vectorization operation that vectorizes each block and stacks the vectors on top of each other.

II. DISTRIBUTED INFERENCE UNDER SMOOTHNESS PRIORS

A. PROBLEM FORMULATION AND ADAPTIVE STRATEGY

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 $v_1 = \frac{1}{\sqrt{N}}\mathbb{1}_N$ [30]. Thus, the Laplacian can be decomposed as:

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

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

Let $w_k \in \mathbb{R}^M$ denote some parameter vector at agent k and let $\mathcal{W} = \mathrm{col}\{w_1, \ldots, w_N\}$ 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. Let us recall the assumption on the risks $\{J_k(w_k)\}$ used in Part I [2].

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 $H_k(w_k) = \nabla^2_{w_k} J_k(w_k)$ is uniformly bounded from below and above, say, as:

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

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

In many situations, there is prior information available about $\mathcal{W}^o = \mathrm{col}\{w_1^o, \ldots, w_N^o\}$. In the current Part II, and its accompanying Part I [2], 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 [16], [17] provide variations for such problems for the special case of mean-square-error costs. 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 [28], [29], [31]–[33]:

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

where $\mathcal{N}_k$ is the set of neighbors of k, i.e., the set of nodes connected to agent k by an edge. 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:

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

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 $\eta \geq 0$ controls the trade-off between the two components of the objective function. 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:

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

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 $\boldsymbol{s}_{k,i}(\cdot)$:

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

Let $\boldsymbol{w}_{k,i}$ denote the estimate of $w_{k,\eta}^o$ at iteration i and node k. In order to solve (4) in a fully distributed and adaptive manner, we proposed in Part I [2] the following diffusion-type algorithm:

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

where $\mu > 0$ is a small step-size parameter and $\boldsymbol{\psi}_{k,i}$ is an intermediate variable.

B. SUMMARY OF MAIN RESULTS

One key observation that followed from the analysis in Part I [2] 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.

In Part I [2], we carried out a detailed stability analysis of the proposed strategy (7). We showed, under some conditions on the step-size parameter μ, that:

$$\limsup_{i\to\infty} \|\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)\| \quad = O(\mu), \quad \text{(see [2, Theorem 4])} \qquad (8)$$

$$\limsup_{i\to\infty} \mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^2 \quad = O(\mu), \quad \text{(see [2, Theorem 2])} \qquad (9)$$

$$\limsup_{i\to\infty} \mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^4 \quad = O(\mu^2), \quad \text{(see [2, Theorem 3])} \qquad (10)$$

where $\mathcal{W}_\eta^o$ is the solution of the regularized problem (4) and $\boldsymbol{\mathcal{W}}_i = \text{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\}$ denotes the network block weight vector at iteration i. Expression (9) indicates that the mean-square error $\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^2$ is on the order of μ. However, in this Part II, we are interested in characterizing how close the $\boldsymbol{\mathcal{W}}_i$ gets to the network limit point $\mathcal{W}_\eta^o$. In particular, we will be able to characterize the network mean-square deviation (MSD) (defined below in (51)) value in terms of the step-size μ, the regularization strength η, the network topology (captured by the eigenvalues λ_m and eigenvectors v_m of the Laplacian L), and the data characteristics (captured by the second-order properties of the costs $H_{k,\eta}$ and second-order moments of the gradient noise $R_{s,k,\eta}$) as follows:

$$\text{MSD} = \frac{\mu}{2N}\sum_{m=1}^{N} \text{Tr}\left(\left(\sum_{k=1}^{N}[v_m]_k^2 H_{k,\eta} + \eta\lambda_m I\right)^{-1}\left(\sum_{k=1}^{N}[v_m]_k^2 R_{s,k,\eta}\right)\right) + \frac{O(\mu)}{(O(1)+O(\eta))}, \qquad (11)$$

where $[v_m]_k$ denotes the k-th entry of the eigenvector v_m. The interpretation of (11) is explained in more detail in Section IV where it is shown, by coupling η and μ in an appropriate manner, that the term $\frac{O(\mu)}{(O(1)+O(\eta))}$ will be a strictly higher order term of μ. As we will explain later in Sections IV and V, by properly setting the parameters, expression (11) allows us to recover the mean-square-deviation of stand-alone adaptive agents ($\eta = 0$) and single-task diffusion networks ($\eta \to \infty$).

Recall that the objective of the multitask strategy (7) 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. Section V in this paper is devoted to quantify the *benefit* of cooperation, namely, the objective of improving the mean-square deviation around the limiting point of the algorithm. In particular, we will be able to characterize the mean-square-deviation ($\overline{\text{MSD}}$) value relative to the multitask objective $\mathcal{W}^o$ in terms of the MSD in (11) and the mismatch

$\mathcal{W}_\eta^o - \mathcal{W}^o$ as follows:

$$\overline{\text{MSD}} = \underbrace{\text{MSD}}_{O(\mu),\eta} + \underbrace{\|\mathcal{W}_\eta^o - \mathcal{W}^o\|^2}_{\text{smoothness},\eta} + 2(\mathcal{W}_\eta^o - \mathcal{W}^o)^\top \underbrace{\text{bias}}_{O(\mu),\eta}, \quad (12)$$

where "bias" is the bias of algorithm (7) relative to $\mathcal{W}_\eta^o$ given in future expression (40). By increasing η, the MSD in (11) is more likely to decrease. However, by increasing η, from expression (31) in Part I [2], $\|\mathcal{W}_\eta^o - \mathcal{W}^o\|^2$ is more likely to increase and the size of this increase is determined by the smoothness of $\mathcal{W}^o$. From future Lemma 2, it turns out that the third term on the RHS in (12) is a function of μ, η, and the smoothness of the multitask objective $\mathcal{W}^o$. By increasing η, this term is more likely to increase. The key conclusion will be that, while the second and third terms on the RHS in (12) 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 these terms will be. This implies that as long as $\mathcal{W}^o$ is sufficiently smooth, moderate regularization strengths η in the range $]0, \infty[$ exist such that $\overline{\text{MSD}}$ at these values of η will be less than $\overline{\text{MSD}}$ at $\eta = 0$ which corresponds to the non-cooperative mode of operation. The best choice for η would be the one minimizing $\overline{\text{MSD}}$ in (12). We refer the reader to Fig. 2 in [2, Section II-B] for an illustration of this concept of multitask learning benefit. This example will be considered further in the numerical experiments section.

C. MODELING ASSUMPTIONS FROM PART I [2]

In this section, we recall the assumptions used in Part I [2] to establish the network mean-square error stability (9).

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 $k, \ell = 1, 2, \ldots, N$:

$$\mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}] = 0, \quad (13)$$

$$\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \le \beta_k^2\|\boldsymbol{w}\|^2 + \sigma_{s,k}^2, \quad (14)$$

$$\mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w})\boldsymbol{s}_{\ell,i}^\top(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}] = 0, \quad k \neq \ell, \quad (15)$$

for some $\beta_k^2 \ge 0$, $\sigma_{s,k}^2 \ge 0$, 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 $j \le i-1$. □

Let us introduce the network block vector $\mathcal{W}_i = \text{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\}$. Recall from Part I [2] that at each iteration, we can view (7) as a mapping from $\mathcal{W}_{i-1}$ to $\mathcal{W}_i$:

$$\boxed{\mathcal{W}_i = (I_{MN} - \mu\eta\mathcal{L})\left(\mathcal{W}_{i-1} - \mu\,\text{col}\left\{\widehat{\nabla_{w_k} J_k}(\boldsymbol{w}_{k,i-1})\right\}_{k=1}^N\right)} \quad (16)$$

We introduced the following condition on the combination matrix $(I_{MN} - \mu\eta\mathcal{L})$.

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 $\eta \ge 0$ satisfy:

$$0 \le \mu\eta \le \frac{2}{\lambda_{\max}(L)}, \quad (17)$$

$$0 \le \mu\eta \le \min_{1\le k\le N}\left\{\frac{1}{\sum_{\ell=1}^N a_{k\ell}}\right\}, \quad (18)$$

where condition (17) ensures stability and condition (18) ensures non-negative entries. □

The results in Part I [2] established that the iterates $\boldsymbol{w}_{k,i}$ converge in the mean-square-error sense to a small $O(\mu)-$neighborhood around the regularized solution $w_{k,\eta}^o$. In this part of the work, we will 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 an accurate first-order expression for the mean-square error (9); the expression will be accurate to first-order in μ.

To arrive at the desired expression, 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. To derive the long-term model, we follow the approach developed in [9]. The first step is to establish the asymptotic stability of the fourth-order moment of the error vector, $\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^4$, which has already been done in Part I [2]. This property is needed to justify the validity of the long-term approximate model. Recall that to establish the fourth-order stability, we introduced 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 $k, \ell = 1, 2, \ldots, N$:

$$\mathbb{E}\left[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right] \le \overline{\beta}_k^4\|\boldsymbol{w}_k\|^4 + \overline{\sigma}_{s,k}^4, \quad (19)$$

for some $\overline{\beta}_k^4 \ge 0$, and $\overline{\sigma}_{s,k}^4 \ge 0$.

As explained in [9], condition (19) implies (14). To establish the mean-stability (8), we introduced a smoothness condition on the Hessian matrices of the individual costs. This smoothness condition will be adopted in the next section 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.,

$$\|\nabla_{w_k}^2 J_k(w_{k,\eta}^o + \Delta w_k) - \nabla_{w_k}^2 J_k(w_{k,\eta}^o)\| \le \kappa_d\|\Delta w_k\|, \quad (20)$$

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

III. LONG-TERM NETWORK DYNAMICS

Let $\widetilde{\mathcal{W}}_i = \mathcal{W}_\eta^o - \mathcal{W}_i$. Subtracting the vector $(I_{MN} - \mu\eta\mathcal{L})\mathcal{W}_\eta^o$ from both sides of recursion (16), and using (6), we obtain:

$$\widetilde{\mathcal{W}}_i - \mu\eta\mathcal{L}\mathcal{W}_\eta^o = (I_{MN} - \mu\eta\mathcal{L})\cdot \left(\widetilde{\mathcal{W}}_{i-1} + \mu\,\mathrm{col}\left\{\nabla_{w_k} J_k(\boldsymbol{w}_{k,i-1}) - \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\right\}_{k=1}^N\right), \tag{21}$$

From the mean-value theorem [34, pp. 24], [9, Appendix D], we have:

$$\nabla_{w_k} J_k(\boldsymbol{w}_{k,i-1}) = \nabla_{w_k} J_k(w_{k,\eta}^o) - \boldsymbol{H}_{k,i-1}(w_{k,\eta}^o - \boldsymbol{w}_{k,i-1}), \tag{22}$$

where

$$\boldsymbol{H}_{k,i-1} \triangleq \int_0^1 \nabla_{w_k}^2 J_k(w_{k,\eta}^o - t(w_{k,\eta}^o - \boldsymbol{w}_{k,i-1}))dt, \tag{23}$$

and from the optimality condition of (4), we have:

$$\mathrm{col}\left\{\nabla_{w_k} J_k(w_{k,\eta}^o)\right\}_{k=1}^N = -\eta\mathcal{L}\mathcal{W}_\eta^o. \tag{24}$$

Replacing (22) into (21) and using (24), we arrive at the following recursion for $\widetilde{\mathcal{W}}_i$:

$$\widetilde{\mathcal{W}}_i = (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\boldsymbol{\mathcal{H}}_{i-1})\widetilde{\mathcal{W}}_{i-1} - \mu(I_{MN} - \mu\eta\mathcal{L})\boldsymbol{s}_i(\mathcal{W}_{i-1}) + \mu^2\eta^2\mathcal{L}^2\mathcal{W}_\eta^o, \tag{25}$$

where

$$\boldsymbol{s}_i(\mathcal{W}_{i-1}) \triangleq \mathrm{col}\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\}_{k=1}^N, \tag{26}$$

$$\boldsymbol{\mathcal{H}}_{i-1} \triangleq \mathrm{diag}\{\boldsymbol{H}_{k,i-1}\}_{k=1}^N. \tag{27}$$

We move on to motivate a long-term model for the evolution of the network error dynamics, $\widetilde{\mathcal{W}}_i$, after sufficient iterations, i.e., for $i \gg 1$. We examine the stability property of the model, the proximity of its trajectory to that of the original network dynamics (25), and subsequently employ the model to assess network performance.

A. LONG-TERM ERROR MODEL

We introduce the error matrix $\widetilde{\boldsymbol{\mathcal{H}}}_{i-1} \triangleq \mathcal{H}_\eta - \boldsymbol{\mathcal{H}}_{i-1}$, which measures the deviation of $\boldsymbol{\mathcal{H}}_{i-1}$ from the constant matrix:

$$\mathcal{H}_\eta \triangleq \mathrm{diag}\{H_{k,\eta}\}_{k=1}^N, \tag{28}$$

with each $H_{k,\eta}$ given by the value of the Hessian matrix at the regularized solution, namely, $H_{k,\eta} \triangleq \nabla_{w_k}^2 J_k(w_{k,\eta}^o)$. Let

$$\boldsymbol{\mathcal{B}}_{i-1} \triangleq (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\boldsymbol{\mathcal{H}}_{i-1}), \tag{29}$$

$$\mathcal{B}_\eta \triangleq (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\eta). \tag{30}$$

Then, we can write:

$$\boldsymbol{\mathcal{B}}_{i-1} = \mathcal{B}_\eta + \mu(I_{MN} - \mu\eta\mathcal{L})\widetilde{\boldsymbol{\mathcal{H}}}_{i-1}. \tag{31}$$

Using (31), we can rewrite the error recursion (25) as:

$$\widetilde{\mathcal{W}}_i = \mathcal{B}_\eta\widetilde{\mathcal{W}}_{i-1} - \mu(I_{MN} - \mu\eta\mathcal{L})\boldsymbol{s}_i(\mathcal{W}_{i-1}) + \mu^2\eta^2\mathcal{L}^2\mathcal{W}_\eta^o + \mu(I_{MN} - \mu\eta\mathcal{L})\boldsymbol{c}_{i-1}, \tag{32}$$

in terms of the random perturbation sequence:

$$\boldsymbol{c}_{i-1} \triangleq \widetilde{\boldsymbol{\mathcal{H}}}_{i-1}\widetilde{\mathcal{W}}_{i-1}. \tag{33}$$

Under Assumptions 1 and 5, and for small μ, it can be shown that $\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{c}_{i-1}\| = O(\mu)$, and that $\|\boldsymbol{c}_{i-1}\| = O(\mu)$ asymptotically with *high probability* (see Appendix A in [35]). Motivated by this result, we introduce the following approximate model, where the last term involving $\boldsymbol{c}_{i-1}$ in (32), which is $O(\mu^2)$, is removed:

$$\widetilde{\mathcal{W}}_i' = \mathcal{B}_\eta\widetilde{\mathcal{W}}_{i-1}' - \mu(I_{MN} - \mu\eta\mathcal{L})\boldsymbol{s}_i(\mathcal{W}_{i-1}) + \mu^2\eta^2\mathcal{L}^2\mathcal{W}_\eta^o, \tag{34}$$

for $i \gg 1$. Obviously, the iterates that are generated by (34) are generally different from the iterates generated by the original recursion (25). To highlight this fact, we are using the prime notation for the state of the long-term model. Note that the driving process $\boldsymbol{s}_i(\mathcal{W}_{i-1})$ in (34) is the *same* gradient noise process from the original recursion (25) and is evaluated at $\mathcal{W}_{i-1}$. In the following, we show that, after sufficient iterations $i \gg 1$, the error dynamics of the network relative to the solution $\mathcal{W}_\eta^o$ is well-approximated by the model (34).

B. SIZE OF APPROXIMATION ERROR

We start by showing that the mean-square difference between the trajectories $\{\widetilde{\mathcal{W}}_i, \widetilde{\mathcal{W}}_i'\}$ is asymptotically bounded by $O(\mu^2)$ and that the mean-square error performance of the long term model (34) is within $O(\mu^{\frac{3}{2}})$ from the performance of the original recursion (25). Working with recursion (34) is much more tractable for performance analysis because its dynamics is driven by the constant matrix $\mathcal{B}_\eta$ as opposed to the random matrix $\boldsymbol{\mathcal{B}}_{i-1}$ in the original error recursion (25). Therefore, we shall work with the long-term model (34) and evaluate its performance, which will provide an accurate representation for the performance of the original distributed strategy (7) to first order in the step-size μ.

Lemma 1: (Size of approximation error) Under Assumptions 1, 2, 3, and 5, and condition (19), it holds that:

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\mathcal{W}}_i - \widetilde{\mathcal{W}}_i'\|^2 = O(\mu^2), \tag{35}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\mathcal{W}}_i\|^2 = \limsup_{i\to\infty} \mathbb{E}\|\widetilde{\mathcal{W}}_i'\|^2 + O(\mu^{\frac{3}{2}}). \tag{36}$$

Proof: See Appendix A. ■

We shall discuss now the mean and mean-square error stability of the long-term approximate model (34).

C. STABILITY OF FIRST-ORDER ERROR MOMENT

Conditioning both sides of (34), invoking the conditions on the gradient noise from Assumption 2, and computing the conditional expectations, we obtain:

$$\mathbb{E}[\widetilde{\mathcal{W}}_i'|\boldsymbol{\mathcal{F}}_{i-1}] = \mathcal{B}_\eta\widetilde{\mathcal{W}}_{i-1}' + \mu^2\eta^2\mathcal{L}^2\mathcal{W}_\eta^o. \tag{37}$$

Taking expectation again, we arrive at:

$$\mathbb{E}\widetilde{\mathcal{W}}_i' = \mathcal{B}_\eta\mathbb{E}\widetilde{\mathcal{W}}_{i-1}' + \mu^2\eta^2\mathcal{L}^2\mathcal{W}_\eta^o. \tag{38}$$

The above recursion is stable if the matrix $\mathcal{B}_\eta$ in (30) is stable. This matrix has a similar form to the matrix $(I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\infty)$ encountered in Part I [2, Section III-A2]. Similarly, it can be verified that $\mathcal{B}_\eta$ is stable when condition (17) and condition

$$0 < \mu < \min_{1\le k\le N}\left\{\frac{2}{\lambda_{k,\max}}\right\}. \tag{39}$$

are satisfied. In this case, we obtain

$$\begin{aligned}\widetilde{\mathcal{W}}'_\infty &\triangleq \lim_{i\to\infty}\mathbb{E}\widetilde{\mathcal{W}}'_i \\ &= \mu^2\eta^2(I_{MN} - (I_{MN} - \mu\eta\mathcal{L})(I_{MN} - \mu\mathcal{H}_\eta))^{-1}\mathcal{L}^2\mathcal{W}^o_\eta,\end{aligned} \tag{40}$$

where the RHS in the above expression is similar to the RHS in equation (50) [2] with $\mathcal{H}_\infty$ replaced by $\mathcal{H}_\eta$.

Lemma 2: (Mean stability of long-term model) Under Assumptions 1, 2, 3, and 5, and for sufficiently small μ, the steady-state bias $\widetilde{\mathcal{W}}'_\infty = \lim_{i\to\infty}\mathbb{E}\widetilde{\mathcal{W}}'_i$ of the long-term model (34) given by (40) satisfies:

$$\mu\lim_{\mu\to 0}\left(\frac{1}{\mu}\lim_{i\to\infty}\|\mathbb{E}\widetilde{\mathcal{W}}'_i\|\right) \le \mu\frac{O(\eta^2)}{(O(1)+O(\eta))^2}. \tag{41}$$

Proof: The proof is similar to the proof of Theorem 1 in Part I [2] with $\mathcal{H}_\infty$ replaced by $\mathcal{H}_\eta$. ■

D. STABILITY OF SECOND-ORDER ERROR MOMENT

In the following, we show that the long term approximate model (34) is also mean-square stable in the sense that $\mathbb{E}\|\widetilde{\boldsymbol{w}}'_{k,i}\|^2$ tends asymptotically to a region that is bounded by $O(\mu)$. We follow the same line of reasoning as in Part I [2, Section III-A] where we studied the mean-square stability of the original model (25). Based on the inequality:

$$\begin{aligned}&\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}^o_\eta - \mathcal{W}'_i\|^2 \\ &= \limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}^o_\eta - \mathcal{W}'_\infty + \mathcal{W}'_\infty - \mathcal{W}'_i\|^2 \\ &\le 2\|\mathcal{W}^o_\eta - \mathcal{W}'_\infty\|^2 + 2\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}'_\infty - \mathcal{W}'_i\|^2,\end{aligned} \tag{42}$$

where $\mathcal{W}^o_\eta - \mathcal{W}'_\infty = \widetilde{\mathcal{W}}'_\infty$ is the steady-state bias of the long term model given by (40) and where $\mathcal{W}'_\infty - \mathcal{W}'_i$ follows the recursion:

$$\mathcal{W}'_\infty - \mathcal{W}'_i = \mathcal{B}_\eta(\mathcal{W}'_\infty - \mathcal{W}'_{i-1}) - \mu(I_{MN} - \mu\eta\mathcal{L})\boldsymbol{s}_i(\mathcal{W}_{i-1}), \tag{43}$$

and from Theorems 1 and 2 in Part I [2] and previous Lemma 2, we can establish the mean-square stability of (34). Let us introduce the mean-square perturbation vector (MSP′) at time i relative to $\mathcal{W}'_\infty$:

$$\mathrm{MSP}'_i \triangleq \mathrm{col}\left\{\mathbb{E}\|w'_{k,\infty} - \boldsymbol{w}'_{k,i}\|^2\right\}_{k=1}^N. \tag{44}$$

Lemma 3. (Mean-square stability of the long-term model): Under Assumptions 1, 2, 3, and 5, the MSP' at time i can be recursively bounded as:

$$\begin{aligned}MSP'_i &\preceq (I_N - \mu\eta L)(G'')^2 MSP'_{i-1} \\ &\quad + 3\mu^2(I_N - \mu\eta L)\mathrm{diag}\{\beta_k^2\}_{k=1}^N MSP_{i-1} + \mu^2(I_N - \mu\eta L)b\end{aligned} \tag{45}$$

where:

$$G'' \triangleq \mathrm{diag}\{\gamma_k\}_{k=1}^N, \tag{46}$$

$$b \triangleq \mathrm{col}\left\{\sigma_{s,k}^2 + 3\beta_k^2\|w^o_{k,\eta}\|^2 + 3\beta_k^2\|w^o_{k,\eta} - w_{k,\infty}\|^2\right\}_{k=1}^N, \tag{47}$$

$$\gamma_k \triangleq \max\{|1-\mu\lambda_{k,\min}|, |1-\mu\lambda_{k,\max}|\}. \tag{48}$$

and MSP_i is the mean-square perturbation vector at time i relative to the fixed point $\mathcal{W}_\infty = \mathrm{col}\{w_{k,\infty}\}_{k=1}^N$ of algorithm (7) in the absence of gradient noise (see [2, Section III-A3]). A sufficiently small μ ensures the stability of the above recursion. It follows that

$$\|\limsup_{i\to\infty} MSP'_i\|_\infty = O(\mu), \tag{49}$$

and that

$$\limsup_{i\to\infty}\mathbb{E}\|\widetilde{\boldsymbol{w}}'_i\|^2 = O(\mu) + \frac{O(\mu^2\eta^4)}{(O(1)+O(\eta))^4} = O(\mu). \tag{50}$$

Proof: See Appendix B. ■

IV. MEAN-SQUARE-ERROR PERFORMANCE

We established in Theorem 2 in Part I [2] that a network running strategy (7) is mean-square-error stable for sufficiently small μ. Specifically, we showed that $\limsup_{i\to\infty}\mathbb{E}\|\mathcal{W}^o_\eta - \mathcal{W}_i\|^2 = O(\mu)$. In the following, we assess the size of the network mean-square-deviation (MSD) using the definition [9, Chapter 11]:

$$\mathrm{MSD} \triangleq \mu\lim_{\mu\to 0}\left(\limsup_{i\to\infty}\frac{1}{\mu}\mathbb{E}\left(\frac{1}{N}\|\mathcal{W}^o_\eta - \mathcal{W}_i\|^2\right)\right). \tag{51}$$

In addition to Assumptions 1, 2, 3, 5, and condition (19) on the individual costs, $J_k(w_k)$, the gradient noise process, $\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)$, and the combination matrix, $I_N - \mu\eta L$, we introduce a smoothness condition on the noise covariance matrices.

For any $\boldsymbol{w}_k \in \boldsymbol{\mathcal{F}}_{i-1}$, we let

$$R_{s,k,i}(\boldsymbol{w}_k) \triangleq \mathbb{E}[\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)\boldsymbol{s}^\top_{k,i}(\boldsymbol{w}_k)|\boldsymbol{\mathcal{F}}_{i-1}] \tag{52}$$

denote the conditional second-order moment of the gradient noise process, which generally depends on i because the statistical distribution of $\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)$ can be iteration-dependent, and is random since it depends on the random iterate $\boldsymbol{w}_k$. We assume that, in the limit, this covariance matrix tends to a constant value when evaluated at $w^o_{k,\eta}$ and we denote the limit by:

$$R_{s,k,\eta} \triangleq \lim_{i\to\infty}\mathbb{E}[\boldsymbol{s}_{k,i}(w^o_{k,\eta})\boldsymbol{s}^\top_{k,i}(w^o_{k,\eta})|\boldsymbol{\mathcal{F}}_{i-1}]. \tag{53}$$

Assumption 6. (Smoothness condition on the noise covariance): It is assumed that the conditional second-order moment

of the noise process is locally Lipschitz continuous in a small neighborhood around $w_{k,\eta}^o$, namely,

$$\|R_{s,k,i}(w_{k,\eta}^o + \Delta w_k) - R_{s,k,i}(w_{k,\eta}^o)\| \leq \kappa_d \|\Delta w_k\|^\theta, \quad (54)$$

for small perturbations $\|\Delta w_k\| \leq \epsilon$, and for some constant $\kappa_d \geq 0$ and exponent $0 < \theta \leq 4$. □

One useful conclusion that follows from Assumption 6 is that, after sufficient iterations, we can express the covariance matrix of the gradient noise process, $\boldsymbol{s}_{k,i}(\boldsymbol{w}_k)$, in terms of the limiting matrix $R_{s,k,\eta}$ defined in (53). Specifically, following the same proof used to establish Lemma 11.1 in [9], we can show that under the smoothness condition (54) and for small step-size, the covariance matrix of the gradient noise process, $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$, at each agent k satisfies for $i \gg 1$:

$$\mathbb{E}\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\boldsymbol{s}_{k,i}^\top(\boldsymbol{w}_{k,i-1}) = R_{s,k,\eta} + O\left(\mu^{\min\left\{1,\frac{\theta}{2}\right\}}\right). \quad (55)$$

For clarity of presentation, we sketch the proof in Appendix D in [35] where we used results from Theorems 2 and 3 in Part I [2].

Before studying the steady-state network performance, we establish some properties of the matrix:

$$\mathcal{F}_\eta \triangleq \mathcal{B}_\eta^\top \otimes_b \mathcal{B}_\eta^\top, \quad (56)$$

which is defined in terms of the block Kronecker operation using blocks of size $M \times M$. In the derivation that follows, we shall use the block Kronecker product $\otimes_b$ operator [36] and the block vectorization operator bvec$(\cdot)$. As explained in [9], these operations preserve the locality of the blocks in the original matrix arguments. Since $\rho(\mathcal{F}_\eta) = (\rho(\mathcal{B}_\eta))^2$, the matrix $\mathcal{F}_\eta$ is stable under conditions (17) and (39). This matrix plays a critical role in characterizing the performance of the distributed multitask algorithm. In our derivations, the matrix $\mathcal{F}_\eta$ will also appear transformed under the orthonormal transformation:

$$\overline{\mathcal{F}}_\eta \triangleq (\mathcal{V} \otimes_b \mathcal{V})^\top \mathcal{F}_\eta (\mathcal{V} \otimes_b \mathcal{V}), \quad (57)$$

where $\mathcal{V} \triangleq V \otimes I_M$.

Lemma 4: (Coefficient matrix $\mathcal{F}_\eta$) For sufficiently small step-size, it holds that

$$(I - \mathcal{F}_\eta)^{-1} = O(\mu^{-1}), \quad (58)$$

and

$$\begin{aligned}
&(I - \overline{\mathcal{F}}_\eta)^{-1} \\
&= X^{-1} + W \quad (59)\\
&= \mu^{-1} \cdot \begin{bmatrix} O(1) & 0 \\ 0 & (O(1) + O(\eta))^{-1} \end{bmatrix} \\
&\quad + \mu^{-1} \cdot \begin{bmatrix} (O(1) + O(\eta))^{-1} & (O(1) + O(\eta))^{-1} \\ (O(1) + O(\eta))^{-1} & (O(1) + O(\eta))^{-2} \end{bmatrix} \quad (60)
\end{aligned}$$

where X is an $N^2 \times N^2$ block diagonal matrix with each block of dimension $M^2 \times M^2$:

$$\begin{aligned}
X = \mu \cdot \mathrm{diag}\Big\{ &\mathrm{diag}\Big\{ (1 - \mu\eta\lambda_m)(1 - \mu\eta\lambda_p)(H_{mm} \oplus H_{pp}) \\
&+ \eta(\lambda_m + \lambda_p - \mu\eta\lambda_m\lambda_p) I_{M^2} \Big\}_{p=1}^N \Big\}_{m=1}^N, \quad (61)
\end{aligned}$$

with $\oplus$ denoting the Kronecker sum operator [37]:

$$H_{mm} \oplus H_{pp} = H_{mm} \otimes I_M + I_M \otimes H_{pp}, \quad (62)$$

$$H_{mn} \triangleq (v_m^\top \otimes I_M)\mathcal{H}_\eta(v_n \otimes I_M). \quad (63)$$

The matrix W is an $N^2 \times N^2$ block matrix arising from the matrices $\{H_{mn} | m \neq n\}$. Moreover, we can also write:

$$\begin{aligned}
(I - \mathcal{F}_\eta)^{-1} = &(\mathcal{V} \otimes_b \mathcal{V}) X^{-1} (\mathcal{V} \otimes_b \mathcal{V})^\top \\
&+ \mu^{-1}(O(1) + O(\eta))^{-1}. \quad (64)
\end{aligned}$$

Proof: See Appendix C. ■

As we shall see in Theorem 1, it turns out that the decomposition in (59) is very useful to highlight some important facts arising in the steady-state performance of the multitask algorithm.

Lemma 5: (Steady-state network performance) Under Assumptions 1, 2, 3, 4, 5, and 6, it holds that:

$$\begin{aligned}
&\limsup_{i \to \infty} \frac{1}{N} \mathbb{E}\|\mathcal{W}_\eta^o - \mathcal{W}_i\|^2 \\
&= \frac{1}{N} \sum_{n=0}^{\infty} \mathrm{Tr}\left(\mathcal{B}_\eta^n \mathcal{Y} (\mathcal{B}_\eta^\top)^n\right) + O(\mu^{1+\theta_m}) \\
&= \frac{1}{N} (\mathrm{bvec}(\mathcal{Y}^\top))^\top (I - \mathcal{F}_\eta)^{-1} \mathrm{bvec}\,(I_{MN}) + O(\mu^{1+\theta_m}), \quad (65)
\end{aligned}$$

where $\theta_m = \frac{1}{2}\min\{1, \theta\}$, $\mathcal{B}_\eta$ and $\mathcal{F}_\eta$ are defined in (30) and (56), and

$$\mathcal{Y} \triangleq \mu^2 (I_{MN} - \mu\eta\mathcal{L}) \mathcal{S}_\eta (I_{MN} - \mu\eta\mathcal{L}), \quad (66)$$

$$\mathcal{S}_\eta \triangleq \mathrm{diag}\left\{R_{s,k,\eta}\right\}_{k=1}^N, \quad (67)$$

Proof: The proof is a direct extension of the arguments used to establish Theorem 11.2 in [9] for single-task diffusion adaptation. See Appendix F in [35]. ■

Theorem 1. (Network MSD performance): Under the same conditions of Lemma 5, it holds from Lemma 4 that the steady-state network *MSD* defined in (51) can be written as:

$$\begin{aligned}
&\mathrm{MSD} = \\
&\frac{\mu}{2N} \sum_{m=1}^N \mathrm{Tr}\left(\left(\sum_{k=1}^N [v_m]_k^2 H_{k,\eta} + \eta\lambda_m I \right)^{-1} \left(\sum_{k=1}^N [v_m]_k^2 R_{s,k,\eta} \right) \right) \\
&\quad + \frac{O(\mu)}{(O(1) + O(\eta))}. \quad (68)
\end{aligned}$$

Proof: See Appendix D. ■

As the derivation in Appendix D reveals, the second term on the RHS of (68) results from the matrix W in (59) which is zero when $\{H_{mn} = 0, m \neq n\}$. When the Hessian matrices are uniform across the agents:

$$H_{k,\eta} \equiv H_\eta, \qquad k = 1, \ldots, N, \tag{69}$$

we have $H_{mn} = 0$ for $m \neq n$. In this case, the network MSD in (68) simplifies to:

$$\text{MSD} = \frac{\mu}{2N} \sum_{m=1}^{N} \text{Tr}\left(\left(H_\eta + \eta\lambda_m I\right)^{-1} \left(\sum_{k=1}^{N} [v_m]_k^2 R_{s,k,\eta} \right) \right). \tag{70}$$

Moreover, in the non-uniform Hessian matrices scenario, by letting $\eta = \mu^{-\epsilon}$ with $\epsilon > 0$ chosen such that Assumption 3 is satisfied, we obtain:

$$\frac{O(\mu)}{(O(1) + O(\eta))} = O(\mu^{1+\epsilon}). \tag{71}$$

In this case, the first term on the RHS of (68) dominates the factor $O(\mu^{1+\epsilon})$ and when we evaluate the network MSD according to definition (51), the last term on the RHS of (68) disappears when computing the limit as $\mu \to 0$. As we will see by simulations, the first term on the RHS of (68) provides a good approximation for the network MSD for any $\eta \geq 0$.

The first term on the RHS of (68) reveals explicitly the influence of the step-size μ, regularization strength η, network topology (through the eigenvalues λ_m and eigenvectors v_m of the Laplacian), gradient noise (through the covariance matrices $R_{s,k,\eta}$), and data characteristics (through the Hessian matrices $H_{k,\eta}$) on the network MSD performance. Observe that this term consists of the sum of N individual terms, each associated with an eigenvalue λ_m of the Laplacian matrix, and given by:

$$\frac{\mu}{2N} \text{Tr}\left(\left(\overline{H}_{m,\eta} + \eta\lambda_m I\right)^{-1} \overline{R}_{m,\eta} \right), \tag{72}$$

where $\overline{H}_{m,\eta}$ and $\overline{R}_{m,\eta}$ are transformed versions of $\mathcal{H}_\eta$ in (28) and $\mathcal{S}_\eta$ in (67) at the m-th eigenvalue λ_m, respectively:

$$\overline{H}_{m,\eta} \triangleq \sum_{k=1}^{N} [v_m]_k^2 H_{k,\eta} = (v_m^\top \otimes I_M)\mathcal{H}_\eta (v_m \otimes I_M), \tag{73}$$

$$\overline{R}_{m,\eta} \triangleq \sum_{k=1}^{N} [v_m]_k^2 R_{s,k,\eta} = (v_m^\top \otimes I_M)\mathcal{S}_\eta (v_m \otimes I_M). \tag{74}$$

As shown in Section VI-A, under some assumptions on the data and noise characteristics, the individual terms in (72) are decaying functions of η or λ_m.

Before proceeding, we note that expression (68) can be written alternatively as:

$$\begin{aligned} \text{MSD} = {} & \frac{\mu}{2N} \text{Tr}\left(\left(\sum_{k=1}^{N} H_{k,\eta} \right)^{-1} \left(\sum_{k=1}^{N} R_{s,k,\eta} \right) \right) + \\ & \frac{\mu}{2N} \sum_{m=2}^{N} \text{Tr}\left(\left(\sum_{k=1}^{N} [v_m]_k^2 H_{k,\eta} + \eta\lambda_m I \right)^{-1} \left(\sum_{k=1}^{N} [v_m]_k^2 R_{s,k,\eta} \right) \right) \\ & + \frac{O(\mu)}{(O(1) + O(\eta))}, \end{aligned} \tag{75}$$

where we used the fact that $\lambda_1 = 0$ and $v_1 = \frac{1}{\sqrt{N}} \mathbb{1}_N$. Expression (75) allows us to recover the network MSD performance of the single-task diffusion adaptation employed to estimate $w^\star$ given by:

$$w^\star \triangleq \arg\min_{w} \sum_{k=1}^{N} J_k(w). \tag{76}$$

To see this, we recall the expression for the network MSD performance of single-task diffusion adaptation derived in [9, pp. 594]:

$$\text{MSD} = \frac{\mu}{2N} \text{Tr}\left(\left(\sum_{k=1}^{N} H_{k,\star} \right)^{-1} \left(\sum_{k=1}^{N} R_{s,k,\star} \right) \right), \tag{77}$$

where $H_{k,\star} \triangleq \nabla_{w_k}^2 J_k(w^\star)$ and $R_{s,k,\star}$ is the covariance of the gradient noise in (53) at $w^\star$. In order to estimate $w^\star$ using the multitask strategy (7), a very large η needs to be chosen. In this case, we have $H_{k,\eta} = H_{k,\star}$ and $R_{s,k,\eta} = R_{s,k,\star}$. Moreover, the second and third terms on the RHS of expression (75) will be $O(\mu/\eta)$ which are negligible for a very large η. Thus, we obtain (77).

V. MULTITASK LEARNING BENEFIT

Now that we have studied in detail the mean-square performance of the multitask strategy (7) relative to w_η^o, the minimizer of the regularized cost (4), we will use the results to examine the benefit of multitask learning compared to the non-cooperative solution under the smoothness assumption.

Since each cost is strongly convex, each agent k is able to estimate w_k^o on its own, if desired, in a non-cooperative manner by running strategy (7) with $\eta = 0$. We know from previous established results on single-agent adaptation [9, pp. 390] that the network MSD in that case will be given by:

$$\text{MSD}_{\text{ncop}} = \frac{\mu}{2N} \text{Tr}\left(\sum_{k=1}^{N} H_{k,o}^{-1} R_{s,k,o} \right), \tag{78}$$

where $H_{k,o} \triangleq \nabla_{w_k}^2 J_k(w_k^o)$ and $R_{s,k,o}$ is the covariance of the gradient noise in (53) at w_k^o. Note that expression (68) allows us to recover the mean-square-error performance of stand-alone adaptive agents. In particular, it can be easily verified

that (68) reduces to expression $\frac{\mu}{2}\mathrm{Tr}(H_{k,o}^{-1}R_{s,k,o})$ for the mean-square deviation of single agent learner [9, pp. 390] when the network size is set to $N = 1$ and the topology is removed.

When $\eta > 0$ is used, the graph Laplacian regularizer (4) induces a bias relative to $\mathcal{W}^o$. However, when $\mathcal{W}^o$ is smooth, we expect that promoting the smoothness of $\mathcal{W}^o$ through regularization can improve the network MSD performance despite the bias induced in the estimation.

A. INDUCED BIAS RELATIVE TO $\mathcal{W}^o$

The bias of the strategy (7) is given by:

$$\mathbb{E}\left(\mathcal{W}^o - \boldsymbol{\mathcal{W}}_i\right) = (\mathcal{W}^o - \mathcal{W}_\eta^o) + \mathbb{E}\left(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\right). \tag{79}$$

From the triangle inequality we have:

$$\begin{aligned}\limsup_{i\to\infty} \|\mathbb{E}\left(\mathcal{W}^o - \boldsymbol{\mathcal{W}}_i\right)\| \\ \leq \|\mathcal{W}^o - \mathcal{W}_\eta^o\| + \limsup_{i\to\infty} \|\mathbb{E}\left(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\right)\|,\end{aligned} \tag{80}$$

where $\limsup_{i\to\infty} \|\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)\|$ can be replaced by $\|\widetilde{\mathcal{W}}'_\infty\|$ in (40). From expression (32) in Part I [2] we know that the smoother $\mathcal{W}^o$ is, the smaller $\|\mathcal{W}^o - \mathcal{W}_\eta^o\|$ will be. Furthermore, we know from Theorem 4 in Part I [2] that $\limsup_{i\to\infty} \|\mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)\|$ is $O(\mu)$. Thus, for a smooth signal $\mathcal{W}^o$ and a small μ, the bias in (79) will be small.

B. NETWORK MSD RELATIVE TO $\mathcal{W}^o$

The mean-square-error performance of the strategy (7) relative to $\mathcal{W}^o$ is given by:

$$\begin{aligned}\mathbb{E}\left\|\mathcal{W}^o - \boldsymbol{\mathcal{W}}_i\right\|^2 = \|\mathcal{W}^o - \mathcal{W}_\eta^o\|^2 + \mathbb{E}\left\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\right\|^2 \\ + 2(\mathcal{W}^o - \mathcal{W}_\eta^o)^\top \mathbb{E}\left(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\right),\end{aligned} \tag{81}$$

and the network MSD can be expressed as:

$$\begin{aligned}\overline{\mathrm{MSD}} &= \frac{1}{N}\limsup_{i\to\infty} \mathbb{E}\left\|\mathcal{W}^o - \boldsymbol{\mathcal{W}}_i\right\|^2 \\ &= \mathrm{MSD} + \frac{1}{N}\|\mathcal{W}^o - \mathcal{W}_\eta^o\|^2 + \\ &\quad \frac{2}{N}(\mathcal{W}^o - \mathcal{W}_\eta^o)^\top \limsup_{i\to\infty} \mathbb{E}\left(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\right),\end{aligned} \tag{82}$$

where we used the bar notation to denote the network MSD relative to $\mathcal{W}^o$ and where MSD is given in (65) or (68). Note that $\limsup_{i\to\infty} \mathbb{E}(\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i)$ can be replaced by $\widetilde{\mathcal{W}}'_\infty$ in (40). In order to improve the network MSD compared to the non-cooperative case (78), the regularization strength η must be chosen such that:

$$\overline{\mathrm{MSD}} \leq \mathrm{MSD}_{\mathrm{ncop}}, \tag{83}$$

and the optimal choice of η is the one minimizing $\overline{\mathrm{MSD}}$ in (82) subject to $\eta \geq 0$.

VI. SIMULATION RESULTS

We consider a connected network of $N = 15$ nodes and $M = 5$ with the topology shown in Fig. 1 (left). Each agent is subjected to streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ that are assumed to satisfy a linear regression model [9], [16]:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w_k^o + \boldsymbol{v}_k(i), \qquad k = 1, \ldots, N, \tag{84}$$

for some unknown $M \times 1$ vector w_k^o with $\boldsymbol{v}_k(i)$ a measurement noise. A mean-square-error cost is associated with agent k:

$$J_k(w_k) = \frac{1}{2}\,\mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w_k|^2, \qquad k = 1, \ldots, N. \tag{85}$$

The processes $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}, \boldsymbol{v}_k(i)\}$ are assumed to represent zero-mean jointly wide-sense stationary random processes satisfying: i) $\mathbb{E}\boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{\ell,j} = R_{u,k}\delta_{k,\ell}\delta_{i,j}$ where $R_{u,k} > 0$ and the Kronecker delta $\delta_{m,n} = 1$ if $m = n$ and zero otherwise; ii) $\mathbb{E}\boldsymbol{v}_k(i)\boldsymbol{v}_\ell(j) = \sigma_{v,k}^2\delta_{k,\ell}\delta_{i,j}$; iii) the regression and noise processes $\{\boldsymbol{u}_{\ell,j}, \boldsymbol{v}_k(i)\}$ are independent of each other. We set $a_{k\ell} = 0.1$ if $\ell \in \mathcal{N}_k$ and 0 otherwise. It turns out that the Laplacian matrix has 15 distinct eigenvalues. We generate $\mathcal{W}^o$ according to $\mathcal{W}^o = \mathcal{V}\overline{\mathcal{W}}^o = \mathrm{col}\{\overline{w}_m^o\}_{m=1}^N$ with:

$$\overline{w}_m^o = \frac{1}{\sqrt{M}} \cdot \mathrm{col}\left\{e^{-\tau_j \lambda_m}\right\}_{j=1}^M. \tag{86}$$

Recall from Part I [2] that $S(\mathcal{W})$ in (3) can be written as:

$$S(\mathcal{W}) = \mathcal{W}^\top \mathcal{L}\mathcal{W} = \sum_{m=2}^N \lambda_m \|\overline{w}_m\|^2, \tag{87}$$

where $\overline{w}_m = (v_m^\top \otimes I_M)\mathcal{W}$. From (87), we observe that the larger $\{\tau_j \geq 0\}$ are, the smoother the signal $\mathcal{W}^o$ is. Note that, for MSE networks, it holds that $H_k(w_k) = \nabla_{w_k}^2 J_k(w_k) = R_{u,k}\ \forall w_k$. Thus, the fixed point bias $\widetilde{\mathcal{W}}_\infty = \mathcal{W}_\eta^o - \mathcal{W}_\infty$ given by (49) in Part I [2] is equal to the long-term model bias $\widetilde{\mathcal{W}}'_\infty$ in (40). Furthermore, from the definition (6), the gradient noise process at agent k is given by:

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_k) = (\boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{k,i} - R_{u,k})(w_k^o - \boldsymbol{w}_k) + \boldsymbol{u}_{k,i}^\top \boldsymbol{v}_k(i). \tag{88}$$

Consider the covariance $R_{s,k,\eta}$ defined in (53). From (88), we have:

$$R_{s,k,\eta} = \mathbb{E}[(\boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{k,i} - R_{u,k})W_{k,\eta}(\boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{k,i} - R_{u,k})] + \sigma_{v,k}^2 R_{u,k} \tag{89}$$

where

$$W_{k,\eta} = (w_k^o - w_{k,\eta}^o)(w_k^o - w_{k,\eta}^o)^\top. \tag{90}$$

To evaluate (89), we need the fourth order moment of the regressors. Let us assume that the regressors are zero-mean real Gaussian. In this case, using the fact that $W_{k,\eta}$ is symmetric, we obtain [38]:

$$\begin{aligned}&\mathbb{E}[\boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{k,i} W_{k,\eta} \boldsymbol{u}_{k,i}^\top \boldsymbol{u}_{k,i}] \\ &= 2R_{u,k}W_{k,\eta}R_{u,k} + R_{u,k}\mathrm{Tr}(R_{u,k}W_{k,\eta}).\end{aligned} \tag{91}$$

Replacing the above expression in (89), we obtain:

$$R_{s,k,\eta} = R_{u,k}W_{k,\eta}R_{u,k} + R_{u,k}\mathrm{Tr}(R_{u,k}W_{k,\eta}) + \sigma_{v,k}^2 R_{u,k}. \tag{92}$$

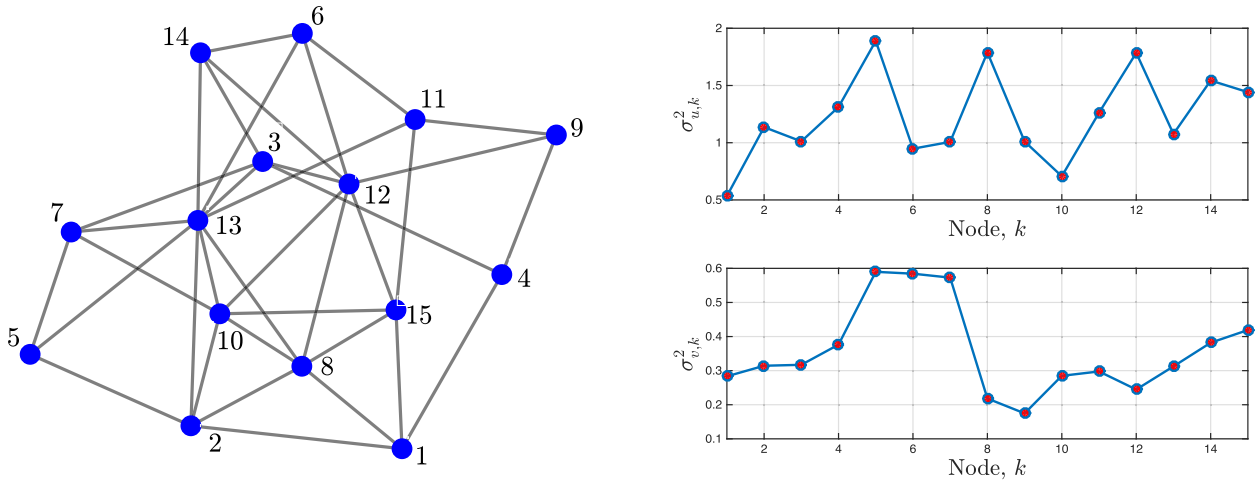

FIGURE 1. **Illustrative example.** ***(Left)*** **Network topology.** ***(Right)*** **Data and noise variances.**

A. UNIFORM DATA PROFILE

In this setting, we assume uniform covariance matrices scenario where $R_{u,k} = R_u \ \forall k$. This scenario is encountered for example in distributed denoising problems in wireless sensor networks (or image denoising) [29], [39]. In this case, R_u is equal to one. In such problems, the N sensors in the network are observing N-dimensional signal, with each entry of the signal corresponding to one sensor. Using the prior knowledge that the signal is smooth w.r.t. the underlying topology, the sensor task is to denoise the corresponding entry of the signal by performing local computations and cooperating with its neighbor in order to improve the error variance.

It turns out that in this scenario, the output of the network $\mathcal{W}^o_\eta$ can be interpreted as the output of a low-pass graph filter applied to the graph signal $\mathcal{W}^o$ [1], [29], [40]. To see this, let us first recall the notion of graph frequencies, graph Fourier transform, and graph filters [29], [39]–[41]. Consider the connected graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}, A\}$ equipped with a Laplacian matrix L, which can be decomposed as $L = V\Lambda V^\top$. A graph signal supported on the set $\mathcal{N}$ is defined as a vector $x \in \mathbb{R}^N$ whose k-th component $x_k \in \mathbb{R}$ represents the value of the signal at the k-th node. By analogy to the classical Fourier analysis, the eigenvectors of the Laplacian matrix L are used to define a graph Fourier basis V and the eigenvalues are considered as the graph frequencies. The graph Fourier transform (GFT) transforms a graph signal x into the graph frequency domain according to $\overline{x} = V^\top x$ where $\{\overline{x}_1, \ldots, \overline{x}_N\}$ are called the spectrum of x. A graph filter Φ is an operator that acts upon a graph signal x by amplifying or attenuating its spectrum as: $\Phi = \sum_{m=1}^N \Phi(\lambda_m)\overline{x}_m v_m$. The frequency response of the filter $\Phi(\lambda)$ controls how much Φ amplifies the signal spectrum. Low frequencies correspond to small eigenvalues, and low-pass or smooth filters correspond to decaying functions $\Phi(\lambda)$. Since we are dealing with vectors $w_k \in \mathbb{R}^M$ instead of scalars x_k, the graph transformation $\overline{x} = V^\top x$ becomes $\overline{\mathcal{W}} = (V^\top \otimes I_M)\mathcal{W}$. In the uniform covariance matrices scenario, we have $\mathcal{H}^o_\eta = I_N \otimes R_u$ and relation (25) in Part I [2] reduces to:

$$\overline{w}^o_{m,\eta} = \left(I_M - \eta\lambda_m \left(\eta\lambda_m I_M + R_u\right)^{-1}\right)\overline{w}^o_m, \ m = 1, \ldots, N. \tag{93}$$

since $(v_m^\top \otimes I_M)\mathcal{H}^o_\eta(v_n \otimes I_M) = R_u$ if $m = n$ and zero otherwise. By applying the matrix inversion identity [42], it holds that the m-th subvector corresponding to the m-th eigenvalue (or graph frequency) of $\overline{\mathcal{W}}^o_\eta = (V^\top \otimes I_M)\mathcal{W}^o$ satisfies:

$$\overline{w}^o_{m,\eta} = \left(\eta\lambda_m I_M + R_u\right)^{-1} R_u \overline{w}^o_m, \tag{94}$$

$$\|\overline{w}^o_{m,\eta}\| \ \leq \frac{1}{1 + \eta\frac{\lambda_m}{\lambda_{\max}(R_u)}}\|\overline{w}^o_m\|, \tag{95}$$

If $\eta = 0$, we are in the case of an all-pass graph filter since the frequency content of the output signal $\mathcal{W}^o_\eta$, is the same as the frequency content of the input signal $\mathcal{W}^o$. For $\eta > 0$, we are in the case of a low-pass graph filter since the norm of the m-th frequency content of the output signal $\mathcal{W}^o_\eta$, namely, $\|\overline{w}^o_{m,\eta}\|$, is less than or equal to the norm of the m-th frequency content of the input signal $\mathcal{W}^o$, namely, $\|\overline{w}^o_m\|$. For fixed η, as m increases, the ratio in (95) decreases. This validates the low-pass filter interpretation. The regularization parameter η controls the sharpness of the low-pass filter. We illustrate this in Fig. 2 (left). In the experiment, we set $R_u = I_M$ and $\tau_j = 0$ in (86) so that $\|\overline{w}^o_m\|^2 = 1 \ \forall m$ and we illustrate in Fig. 2 (left) the squared ℓ_2-norm of $\overline{w}^o_{m,\eta}$ for different values of η from (94). In order to visualize the frequency response of the graph filter, we also plot in dashed lines the ratio $\frac{\|\overline{w}^o_{m,\eta}\|^2}{\|\overline{w}^o_m\|^2}$ from (94) for $\lambda_m \in [0, 1.2]$ with $\|\overline{w}^o_m\|^2 = 1 \ \forall m$.

We observe that a similar behavior arises when studying the network MSD for smooth signal $\mathcal{W}^o$. When $\mathcal{W}^o$ is smooth, $W_{k,\eta}$ in (90) is small and in this case, the covariance matrix $R_{s,k,\eta}$ in (92) can be approximated by $R_{s,k,\eta} \approx \sigma^2_{v,k} R_u$. In this case,

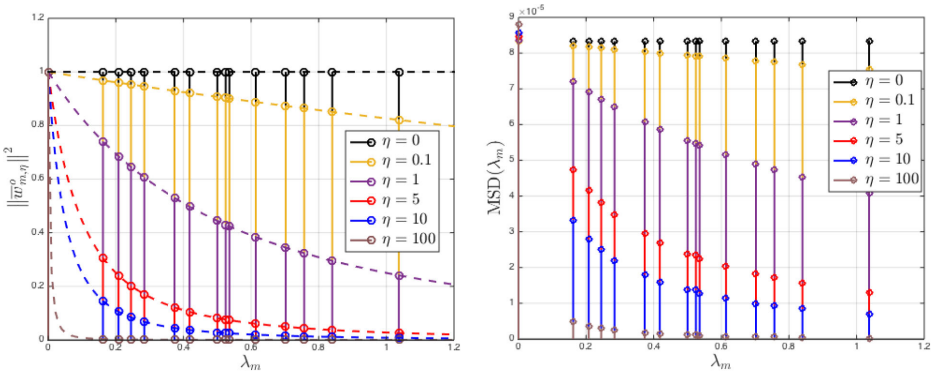

FIGURE 2. **Uniform data profile scenario.** ***(Left)*** **Spectral content of $\mathcal{W}_\eta^o$ from (94) for different η. Dashed lines correspond to the ratio $\frac{\|\overline{w}_{m,\eta}^o\|^2}{\|\overline{w}_m^o\|^2}$ for $\lambda_m \in [0, 1.2]$.** ***(Right)*** **Network MSD at λ_m from (98) (relative to $\mathcal{W}_\eta^o$ with $\mathcal{W}^o$ a smooth signal) for different η at $\mu = 0.005$.**

the network MSD expression in (70) can be approximated as:

$$\text{MSD} \approx \frac{\mu M}{2}\frac{1}{N^2}\left(\sum_{k=1}^{N}\sigma_{v,k}^2\right) + \frac{\mu}{2N}\sum_{m=2}^{N}\left(\sum_{k=1}^{N}[v_m]_k^2\sigma_{v,k}^2\right)\text{Tr}\left(\left(I_M + \eta\lambda_m R_u^{-1}\right)^{-1}\right). \tag{96}$$

Since the trace of a matrix is equal to the sum of its eigenvalues, we have:

$$\text{Tr}\left(\left(I_M + \eta\lambda_m R_u^{-1}\right)^{-1}\right) = \sum_{q=1}^{M}\frac{1}{1+\frac{\eta\lambda_m}{\lambda_q(R_u)}}. \tag{97}$$

The above expression shows that for a fixed λ_m, as η increases, the above trace decreases. We conclude that, when η increases, the sum on the RHS of (96) decreases. By further assuming uniform noise profile, i.e., $\sigma_{v,k}^2 = \sigma_v^2\ \forall k$, expression (96) reduces to:

$$\text{MSD} = \sum_{m=1}^{N}\text{MSD}(\lambda_m) \qquad \text{with}$$

$$\text{MSD}(\lambda_m) \approx \frac{\mu}{2N}\sigma_v^2\text{Tr}\left(\left(I_M + \eta\lambda_m R_u^{-1}\right)^{-1}\right). \tag{98}$$

From (97), we conclude that, for a fixed η, as λ_m increases, the corresponding trace term in (98) decreases. This case is illustrated numerically in Fig. 2 (right). In the experiment, we set $R_u = I_M$, $\sigma_v^2 = 0.1$, $\mu = 0.005$, and $\tau_j = 7 + j$ in (86) so that $\mathcal{W}^o$ is smooth and we illustrate MSD(λ_m) in (98) for different values of η.

B. VARYING DATA PROFILE

We assume that $R_{u,k} = \sigma_{u,k}^2 I_M$ for all k. In this case, expression (92) reduces to:

$$R_{s,k,\eta} = \sigma_{u,k}^4(W_{k,\eta} + \|w_k^o - w_{k,\eta}^o\|^2 I_M) + \sigma_{v,k}^2\sigma_{u,k}^2 I_M. \tag{99}$$

The variances $\sigma_{u,k}^2$ and $\sigma_{v,k}^2$ are given in Fig. 1 (right). We set $\tau_j = j$ in (86).

In order to characterize the influence of the step-size μ and the regularization parameter η on the performance of the algorithm relative to $\mathcal{W}_\eta^o$, we report in Fig. 3 (left) the squared norm of the fixed point bias $\widetilde{\mathcal{W}}_\infty = \mathcal{W}_\eta^o - \mathcal{W}_\infty$ given by (50) in Part I [2] (which is equal to the long-term model bias $\widetilde{\mathcal{W}}'_\infty$ in (40)) as a function of η where $\eta = \mu^{-\epsilon}$ with $\epsilon \geq -1$ for $\mu = \{10^{-3}, 10^{-4}, 10^{-5}\}$. We observe that for small η, the squared norm of the bias increases 40 dB per decade (when η goes from η_1 to $10\eta_1$). This means that the squared norm of the bias is on the order of η^4 for fixed μ. For large η, the bias becomes constant and we see that, when μ goes from μ_1 to $10\mu_1$, it increases 20 dB. This means that the squared norm of the bias is on the order of μ^2. Finally, we observe that for fixed η, the squared norm of the bias is on the order of μ^2. In Fig. 3 (middle), we report the network MSD learning curves relative to $\mathcal{W}_\eta^o$ obtained by running strategy (7) for $\eta = 5$ and for two different values of μ. The curves are obtained by averaging the trajectories $\{\frac{1}{N}\mathbb{E}\|\mathcal{W}_\eta^o - \boldsymbol{\mathcal{W}}_i\|^2\}$ over 200 repeated experiments. In the simulations, we use the following approximation for the network MSD expression in (68):

$$\text{MSD}_{\text{app}} = \frac{\mu}{2N}\sum_{m=1}^{N}\left(\sum_{k=1}^{N}[v_m]_k^2\sigma_{u,k}^2 + \eta\lambda_m\right)^{-1}\text{Tr}\left(\sum_{k=1}^{N}[v_m]_k^2 R_{s,k,\eta}\right) \tag{100}$$

where we use the subscript "app" to indicate that it is an approximation. Compared with (68), we see that the term $\frac{O(\mu)}{(O(1)+O(\eta))}$ has been removed in (100). It is observed that the learning curves tend to the same MSD value predicted by the theoretical expression (65) (with $O(\mu^{1+\theta_m})$ removed). Furthermore, we observe that the MSD predicted by expression (100) provides a good approximation for the performance of strategy (7). Finally, it can be observed that the MSD is on the order of μ. In Fig. 1 (right), we report the MSD learning curves relative to $\mathcal{W}_\eta^o$ obtained by running strategy (7) for $\mu = 10^{-3}$ and for six different values of η. As it can be seen, by increasing η, the network MSD decreases. Furthermore, it

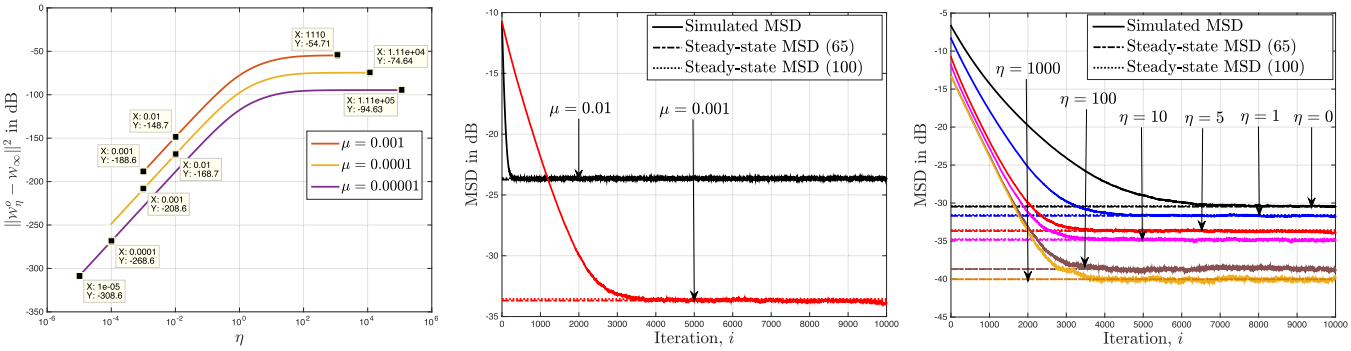

FIGURE 3. Network performance relative to w_η^o. *(Left)* Squared ℓ_2-norm of the bias (40). *(Middle)* Evolution of the learning curves for fixed regularization strength $\eta = 5$, varying step-size μ. *(Right)* Evolution of the learning curves for fixed $\mu = 0.001$, varying η.

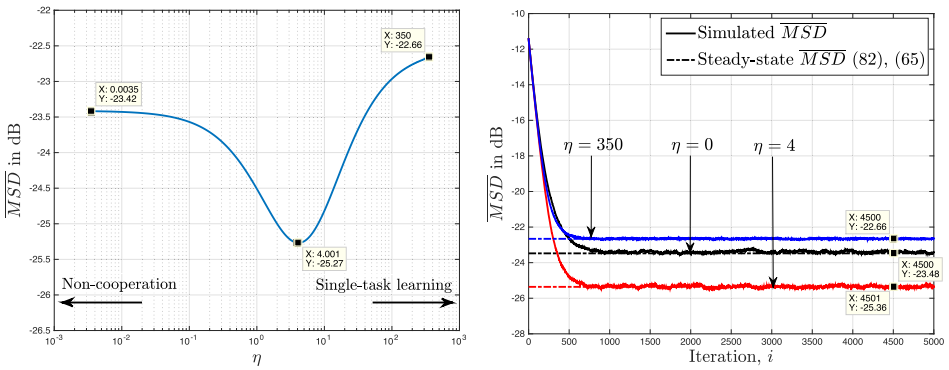

FIGURE 4. Network performance relative to w^o with a smooth signal w^o at $\mu = 0.005$. *(Left)* Network $\overline{\text{MSD}}$ as a function of the regularization strength $\eta \in [0, 350]$. *(Right)* Evolution of the network $\overline{\text{MSD}}$ learning curve for three different values of η.

is observed that expression (100) provides a good approximation for the network MSD for any $\eta \geq 0$.

In Fig. 4, we characterize the influence of η on the performance of strategy (7) relative to w^o with w^o a smooth signal generated according to (86) with $\tau_j = 7 + j$. We set $\mu = 5 \cdot 10^{-3}$. In Fig. 4 (left), we plot $\overline{\text{MSD}}$ for $\eta \in [0, 350]$. To generate $\overline{\text{MSD}}$ we use (82) with MSD replaced by MSD_{app} in (100) which has a low computational complexity. As it can be seen from this plot, $\eta = 4$ gives the best $\overline{\text{MSD}}$. In Fig. 4 (right), we report the network learning curves $\frac{1}{N}\mathbb{E}\|w^o - \mathcal{W}_i\|^2$ for $\eta = \{0, 4, 350\}$. It is observed that the learning curves tend to the same $\overline{\text{MSD}}$ value predicted by the theoretical expression (82) with MSD replaced by (65) (with $O(\mu^{1+\theta_m})$ removed) or (100).

VII. CONCLUSION

In this paper, and its accompanying Part I [2], 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 influences the limit point and the steady-state mean-square-error (MSE) performance of the algorithm. Analytical expressions illustrating these effects are derived. These expressions revealed explicitly the influence of the network topology, data settings, step-size parameter, and regularization strength on the network MSE performance and provided insights into the design of effective multitask strategies for distributed inference over networks. Illustrative examples were considered and links to spectral graph filtering were also provided.

APPENDIX A
PROOF OF LEMMA 1

To simplify the notation, we introduce the difference:

$$z_i \triangleq \widetilde{\mathcal{W}}_i - \widetilde{\mathcal{W}}'_i. \tag{101}$$

Subtracting recursions (32) and (34), we get:

$$z_i = \mathcal{B}_\eta z_{i-1} + \mu(I_{MN} - \mu\eta\mathcal{L})\mathbf{c}_{i-1}. \tag{102}$$

in terms of the random perturbation sequence $\boldsymbol{c}_{i-1}$ given in (33). For each agent k, we have from eq. (176) in Part I [2]:

$$\begin{aligned}
&\|\widetilde{\boldsymbol{H}}_{k,i-1}\| \\
&\triangleq \|H_{k,\eta} - \boldsymbol{H}_{k,i-1}\| \\
&\leq \int_0^1 \|\nabla^2_{w_k} J_k(w^o_{k,\eta}) - \nabla^2_{w_k} J_k(w^o_{k,\eta} - t\widetilde{\boldsymbol{w}}_{k,i-1})\| dt \\
&\leq \int_0^1 \kappa'_d \|t\widetilde{\boldsymbol{w}}_{k,i-1}\| dt = \frac{1}{2}\kappa'_d \|\widetilde{\boldsymbol{w}}_{k,i-1}\|, \qquad (103)
\end{aligned}$$

It follows that

$$\begin{aligned}
\|\boldsymbol{c}_{k,i-1}\| &\leq \|\widetilde{\boldsymbol{H}}_{k,i-1}\| \|w^o_{k,\eta} - \boldsymbol{w}_{k,i-1}\| \\
&\leq \frac{1}{2}\kappa'_d \|w^o_{k,\eta} - \boldsymbol{w}_{k,i-1}\|^2 \qquad (104) \\
&\leq \frac{1}{2}\kappa'_d \|\mathcal{W}^o_\eta - \boldsymbol{\mathcal{W}}_{i-1}\|^2.
\end{aligned}$$

and

$$\|\boldsymbol{c}_{k,i-1}\|^2 \leq \frac{1}{4}(\kappa'_d)^2 \|\mathcal{W}^o_\eta - \boldsymbol{\mathcal{W}}_{i-1}\|^4. \qquad (105)$$

Applying Jensen's inequality [43, pp. 77] to the convex function $\|\cdot\|^2$, we obtain from (102):

$$\mathbb{E}\|\boldsymbol{z}_{k,i}\|^2 \leq \sum_{\ell\in\mathcal{N}_k} [C]_{k,\ell} \mathbb{E}\|(I - \mu H_{\ell,\eta})\boldsymbol{z}_{\ell,i-1} + \mu \boldsymbol{c}_{\ell,i-1}\|^2. \qquad (106)$$

where C is given by:

$$C \triangleq I_N - \mu\eta L. \qquad (107)$$

Next note that

$$\begin{aligned}
&\mathbb{E}\|(I - \mu H_{k,\eta})\boldsymbol{z}_{k,i-1} + \mu \boldsymbol{c}_{k,i-1}\|^2 \\
&= \mathbb{E}\left\| t\frac{1}{t}(I_M - \mu H_{k,\eta})\boldsymbol{z}_{k,i-1} + \mu(1-t)\frac{1}{1-t}\boldsymbol{c}_{k,i-1}\right\|^2 \\
&\leq \frac{1}{t}\mathbb{E}\|(I_M - \mu H_{k,\eta})\boldsymbol{z}_{k,i-1}\|^2 + \mu^2 \frac{1}{1-t}\mathbb{E}\|\boldsymbol{c}_{k,i-1}\|^2, \qquad (108)
\end{aligned}$$

for any arbitrary positive number $t \in (0,1)$. We select $t = \gamma_k$ with γ_k defined in (48). This gives

$$\begin{aligned}
&\mathbb{E}\|(I - \mu H_{k,\eta})\boldsymbol{z}_{k,i-1} + \mu \boldsymbol{c}_{k,i-1}\|^2 \\
&\leq \gamma_k \mathbb{E}\|\boldsymbol{z}_{k,i-1}\|^2 + \mu^2 \frac{1}{1-\gamma_k}\mathbb{E}\|\boldsymbol{c}_{k,i-1}\|^2, \qquad (109)
\end{aligned}$$

Let us introduce the mean-square perturbation vector at time i:

$$\mathrm{MSP}_{z,i} \triangleq \mathrm{col}\left\{\mathbb{E}\|\boldsymbol{z}_{k,i}\|^2\right\}_{k=1}^N. \qquad (110)$$

Replacing (109) into (106), and using (105), we obtain:

$$\begin{aligned}
&\mathrm{MSP}_{z,i} \preceq CG''\mathrm{MSP}_{z,i-1} \\
&+ \mu^2 \frac{1}{4}(\kappa'_d)^2 C(I_N - G'')^{-1}(\mathbb{1}_N \otimes \mathbb{E}\|\mathcal{W}^o_\eta - \boldsymbol{\mathcal{W}}_{i-1}\|^4), \qquad (111)
\end{aligned}$$

with $G'' = \mathrm{diag}\{\gamma_k\}_{k=1}^N$. Iterating the above recursion starting from $i = 1$, we obtain:

$$\begin{aligned}
&\mathrm{MSP}_{z,i} \preceq (CG'')^i \mathrm{MSP}_{z,0} + \\
&\frac{\mu^2}{4}(\kappa'_d)^2 \sum_{j=0}^{i-1} (CG'')^j C(I_N - G'')^{-1}(\mathbb{1}_N \otimes \mathbb{E}\|\mathcal{W}^o_\eta - \boldsymbol{\mathcal{W}}_{i-1-j}\|^4) \qquad (112)
\end{aligned}$$

Under Assumption 3 and condition (39), the matrix CG'' is guaranteed to be stable. Following similar arguments to the ones used to establish eq. (70) in Part I (Appendix E) [2], and from Theorem 3 in Part I [2], we conclude that:

$$\|\limsup_{i\to\infty} \mathrm{MSP}_{z,i}\|_\infty = O(\mu^2), \qquad (113)$$

where we used the fact that $\|(I_N - G'')^{-1}\|_\infty \leq O(\mu^{-1})$. It follows that

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i - \widetilde{\boldsymbol{\mathcal{W}}}'_i\|^2 = O(\mu^2). \qquad (114)$$

Finally note that

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}'_i\|^2 \\
&= \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}'_i - \widetilde{\boldsymbol{\mathcal{W}}}_i + \widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 \\
&\leq \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}'_i - \widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 + \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 + 2|\mathbb{E}(\widetilde{\boldsymbol{\mathcal{W}}}'_i - \widetilde{\boldsymbol{\mathcal{W}}}_i)^\top \widetilde{\boldsymbol{\mathcal{W}}}_i| \\
&\leq \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}'_i - \widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 + \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 + 2\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}'_i - \widetilde{\boldsymbol{\mathcal{W}}}_i\|^2 \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2} \qquad (115)
\end{aligned}$$

where we used $|\mathbb{E}\boldsymbol{x}| \leq \mathbb{E}|\boldsymbol{x}|$ from Jensen's inequality and where we applied Hölder's inequality:

$$\mathbb{E}|\boldsymbol{x}^\top \boldsymbol{y}| \leq \left(\mathbb{E}|\boldsymbol{x}|^p\right)^{\frac{1}{p}} \left(\mathbb{E}|\boldsymbol{x}|^q\right)^{\frac{1}{q}}, \quad \text{when } 1/p + 1/q = 1.$$

Hence we get:

$$\limsup_{i\to\infty} \left(\mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}'_i\|^2 - \mathbb{E}\|\widetilde{\boldsymbol{\mathcal{W}}}_i\|^2\right) \leq O(\mu^2) + \sqrt{O(\mu^3)} = O(\mu^{\frac{3}{2}}), \qquad (116)$$

since $\mu^2 < \mu^{\frac{3}{2}}$ for small $\mu \ll 1$.

APPENDIX B
PROOF OF LEMMA 3

From (43), we have:

$$w'_{k,\infty} - \boldsymbol{w}'_{k,i} = \sum_{\ell=1}^N [C]_{k\ell} \boldsymbol{\phi}'_{\ell,i}, \qquad (117)$$

where C is defined in (107) and where $\boldsymbol{\phi}'_{k,i}$ is given by:

$$\boldsymbol{\phi}'_{k,i} = (I_M - \mu H_{k,\eta})(w'_{k,\infty} - \boldsymbol{w}'_{k,i-1}) - \mu \boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}). \qquad (118)$$

Applying Jensen's inequality [43, pp. 77] to the convex function $\|\cdot\|^2$, we obtain:

$$\mathbb{E}\|w'_{k,\infty} - \boldsymbol{w}'_{k,i}\|^2 \leq \sum_{\ell=1}^N [C]_{k\ell} \mathbb{E}\|\boldsymbol{\phi}'_{\ell,i}\|^2. \qquad (119)$$

Under Assumption 2, we have:

$$\begin{aligned}&\mathbb{E}[\|\boldsymbol{\phi}'_{k,i}\|^2|\boldsymbol{\mathcal{F}}_{i-1}]\\&=\|w'_{k,\infty}-\boldsymbol{w}'_{k,i-1}\|^2_{\Sigma_k}+\mu^2\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}],\end{aligned}\tag{120}$$

where $\Sigma_k \triangleq (I_M-\mu H_{k,\eta})^2$, which due to Assumption 1, can be bounded as follows:

$$0<\Sigma_k\leq\gamma_k^2 I_M,\tag{121}$$

where γ_k is given by (48). Taking expectation again in (120), and using the bound (138) on $\mathbb{E}[\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}]$ from Part I [2], we obtain:

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{\phi}'_{k,i}\|^2\\&=\mathbb{E}\|w'_{k,\infty}-\boldsymbol{w}'_{k,i-1}\|^2_{\Sigma_k}+\mu^2\mathbb{E}\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2\\&\leq\gamma_k^2\mathbb{E}\|w'_{k,\infty}-\boldsymbol{w}'_{k,i-1}\|^2+3\mu^2\beta_k^2\mathbb{E}\|w_{k,\infty}-\boldsymbol{w}_{k,i-1}\|^2\\&\quad+\mu^2\left(3\beta_k^2\|w^o_{k,\eta}-w_{k,\infty}\|^2+3\beta_k^2\|w^o_{k,\eta}\|^2+\sigma^2_{s,k}\right).\end{aligned}\tag{122}$$

Now, combining (122) and (119), we obtain (45).

Iterating (45) starting from $i=1$, we get:

$$\begin{aligned}\mathrm{MSP}'_i&\preceq(C(G'')^2)^i\mathrm{MSP}'_0\\&\quad+\mu^2\sum_{j=0}^{i-1}(C(G'')^2)^jC(3\mathrm{diag}\{\beta_k^2\}\mathrm{MSP}_{i-1-j}+b).\end{aligned}\tag{123}$$

Under Assumption 3 and condition (39), the matrix $C(G'')^2$ is guaranteed to be stable. Using the fact that $b=O(1)$ from Theorem 1 in Part I [2], $1-\|(G'')^2\|_\infty=O(\mu)$, and that $\|\lim_{i\to\infty}\mathrm{MSP}_i\|_\infty=O(\mu)$ from Theorem 2 in Part I [2], and following similar arguments as the one used to establish eq. (70) in Part I (Appendix E) [2], we conclude (49).

From (42), Lemma 2, and (49), we conclude (50).

APPENDIX C
PROOF OF LEMMA 4

Consider the matrix $\overline{\mathcal{F}}_\eta$ in (57). Using the block Kronecker product property:

$$(\mathcal{A}\otimes_b\mathcal{B})(\mathcal{C}\otimes_b\mathcal{D})=(\mathcal{AC}\otimes_b\mathcal{BD}),\tag{124}$$

it can be verified that:

$$\overline{\mathcal{F}}_\eta=\overline{\mathcal{B}}_\eta^{\top}\otimes_b\overline{\mathcal{B}}_\eta^{\top},\tag{125}$$

where

$$\overline{\mathcal{B}}_\eta=\mathcal{V}^{\top}\mathcal{B}_\eta\mathcal{V}=(I_{MN}-\mu\eta\mathcal{J})(I_{MN}-\mu\mathcal{V}^{\top}\mathcal{H}_\eta\mathcal{V})\tag{126}$$

so that:

$$\overline{\mathcal{B}}_\eta^{\top}=\begin{bmatrix}I-\mu H_{11} & -\mu(1-\mu\eta\lambda_2)H_{21} & \dots & -\mu(1-\mu\eta\lambda_N)H_{N1}\\ -\mu H_{12} & (1-\mu\eta\lambda_2)(I-\mu H_{22}) & \dots & -\mu(1-\mu\eta\lambda_N)H_{N2}\\ \vdots & \vdots & & \vdots\\ -\mu H_{1N} & -\mu(1-\mu\eta\lambda_2)H_{2N} & \dots & (1-\mu\eta\lambda_N)(I-\mu H_{NN})\end{bmatrix}\tag{127}$$

where H_{mn} is defined in (63). It can be verified that the matrix $Z=I-\overline{\mathcal{F}}_\eta$ is $N\times N$ blocks Z_{mn} with each block of size $M^2N\times M^2N$:

$$Z=I-\overline{\mathcal{F}}_\eta=\begin{bmatrix}Z_{11} & Z_{12} & \dots & Z_{1N}\\ Z_{21} & Z_{22} & \dots & Z_{2N}\\ \vdots & \vdots & & \vdots\\ Z_{N1} & Z_{N2} & \dots & Z_{NN}\end{bmatrix},\tag{128}$$

We denote by $[Z_{mn}]_{pq}$ the $M^2\times M^2$ (p,q)-th block of Z_{mn}. We have:

$$[Z_{mn}]_{pq}=\begin{cases}I-(1-\mu\eta\lambda_m)(1-\mu\eta\lambda_p)[(I-\mu H_{mm})\otimes(I-\mu H_{pp})],\\\qquad\text{if } m=n,\ p=q\\ \mu(1-\mu\eta\lambda_m)(1-\mu\eta\lambda_q)[(I-\mu H_{mm})\otimes H_{qp}],\\\qquad\text{if } m=n,\ p\neq q\\ \mu(1-\mu\eta\lambda_n)(1-\mu\eta\lambda_p)[H_{nm}\otimes(I-\mu H_{pp})],\\\qquad\text{if } m\neq n,\ p=q\\ -\mu^2(1-\mu\eta\lambda_n)(1-\mu\eta\lambda_q)[H_{nm}\otimes H_{qp}],\\\qquad\text{if } m\neq n,\ p\neq q.\end{cases}\tag{129}$$

We have:

$$\begin{aligned}&(I-\mu H_{mm})\otimes(I-\mu H_{pp})\\&\qquad=I-\mu H_{mm}\oplus H_{pp}+\mu^2H_{mm}\otimes H_{pp},\end{aligned}\tag{130}$$

where $H_{mm}\oplus H_{pp}$ is given by (62) and

$$1-(1-\mu\eta\lambda_m)(1-\mu\eta\lambda_p)=\mu\eta(\lambda_m+\lambda_p-\mu\eta\lambda_m\lambda_p).\tag{131}$$

Thus,

$$[Z_{mn}]_{pq} = \mu \cdot \begin{cases} (1-\mu\eta\lambda_m)(1-\mu\eta\lambda_p)(H_{mm} \oplus H_{pp}) + \\ \quad \eta(\lambda_m + \lambda_p - \mu\eta\lambda_m\lambda_p)I + O(\mu), \\ \quad \text{if } m = n,\ p = q \\ (1-\mu\eta\lambda_m)(1-\mu\eta\lambda_q)(I \otimes H_{qp}) + O(\mu), \\ \quad \text{if } m = n,\ p \neq q \\ (1-\mu\eta\lambda_n)(1-\mu\eta\lambda_p)(H_{nm} \otimes I) + O(\mu), \\ \quad \text{if } m \neq n,\ p = q \\ -\mu(1-\mu\eta\lambda_n)(1-\mu\eta\lambda_q)[H_{nm} \otimes H_{qp}] = O(\mu), \\ \quad \text{if } m \neq n,\ p \neq q \end{cases} \tag{132}$$

Before proceeding, we recall the following useful properties of the Kronecker and Kronecker sum products [37]. Let $\{\lambda_i(A), i = 1, \ldots, M\}$ and $\{\lambda_j(B), j = 1, \ldots, M\}$ denote the eigenvalues of any two $M \times M$ matrices A and B, respectively. Then,

$$\{\lambda(A \otimes B)\} = \{\lambda_i(A)\lambda_j(B)\}_{i=1,j=1}^{M,M}, \tag{133}$$

$$\{\lambda(A \oplus B)\} = \{\lambda_i(A) + \lambda_j(B)\}_{i=1,j=1}^{M,M}. \tag{134}$$

From (128), (129), and (132), it can be verified that the matrix Z can be written as:

$$Z = X + Y. \tag{135}$$

The matrix X is $N^2 \times N^2$ block diagonal defined as:

$$X \triangleq \mu \cdot \text{diag}\left\{\text{diag}\left\{[Z_{mm}]_{pp}\right\}_{p=1}^{N}\right\}_{m=1}^{N} = \mu \cdot \begin{bmatrix} O(1) & 0 \\ 0 & O(1) + O(\eta) \end{bmatrix}, \tag{136}$$

where we used the fact that for $m = 1$ and $p = 1$, we have $[Z_{11}]_{11} = \mu \cdot H_{11} \oplus H_{11} + O(\mu^2)$ which is $\mu \cdot O(1)$. This is due to property (134) and the fact that $H_{11} = \frac{1}{N}\sum_{k=1}^{N} H_{k,\eta} > 0$. For the remaining blocks of X, from (131), property (134), and the fact that $H_{mm} = \sum_{k=1}^{N}[v_m]_k^2 H_{k,\eta} > 0$, it can be verified that the matrix:

$$[Z_{mm}]_{pp} = \mu(1-\mu\eta\lambda_m)(1-\mu\eta\lambda_p)(H_{mm} \oplus H_{pp}) + \mu\eta(\lambda_m + \lambda_p - \mu\eta\lambda_m\lambda_p)I \tag{137}$$

is also positive definite when:

$$0 < (1-\mu\eta\lambda_m)(1-\mu\eta\lambda_p) \leq 1. \tag{138}$$

Furthermore, in this case, we have $[Z_{mm}]_{pp} = \mu \cdot (O(1) + O(\eta))$. The matrix $Y = Z - X$ in (135) is an $N^2 \times N^2$ block matrix where each block is $M^2 \times M^2$ given by:

$$[Y_{mn}]_{pq} = \mu \cdot \begin{cases} 0, & \text{if } m = n,\ p = q \\ (1-\mu\eta\lambda_m)(1-\mu\eta\lambda_q)(I \otimes H_{qp}) + O(\mu) \leq O(1) \\ \quad \text{if } m = n,\ p \neq q \\ (1-\mu\eta\lambda_n)(1-\mu\eta\lambda_p)(H_{nm} \otimes I) + O(\mu) \leq O(1) \\ \quad \text{if } m \neq n,\ p = q \\ -\mu(1-\mu\eta\lambda_n)(1-\mu\eta\lambda_q)[H_{nm} \otimes H_{qp}] \leq O(\mu) \\ \quad \text{if } m \neq n,\ p \neq q \end{cases} \tag{139}$$

Applying the matrix inversion identity [42], we obtain:

$$(X+Y)^{-1} = X^{-1} - X^{-1}Y(I + X^{-1}Y)^{-1}X^{-1}, \tag{140}$$

From (136), we have:

$$X^{-1} = \mu^{-1} \cdot \text{diag}\left\{\text{diag}\left\{([Z_{mm}]_{pp})^{-1}\right\}_{p=1}^{N}\right\}_{m=1}^{N} = \mu^{-1} \cdot \begin{bmatrix} O(1) & 0 \\ 0 & (O(1) + O(\eta))^{-1} \end{bmatrix}. \tag{141}$$

From (141) and (139), we have:

$$X^{-1}Y = \begin{bmatrix} 0 & O(1) \\ (O(1) + O(\eta))^{-1} & (O(1) + O(\eta))^{-1} \end{bmatrix}, \tag{142}$$

and

$$I + X^{-1}Y = \begin{bmatrix} O(1) & O(1) \\ (O(1) + O(\eta))^{-1} & O(1) \end{bmatrix}. \tag{143}$$

Applying the block inversion formula to $I + X^{-1}Y$, we obtain:

$$(I + X^{-1}Y)^{-1} = \begin{bmatrix} O(1) & O(1) \\ (O(1) + O(\eta))^{-1} & O(1) \end{bmatrix}. \tag{144}$$

Finally, from (141), (142), and (144), we conclude that:

$$X^{-1}Y(I + X^{-1}Y)^{-1}X^{-1} = \mu^{-1} \cdot \begin{bmatrix} (O(1) + O(\eta))^{-1} & (O(1) + O(\eta))^{-1} \\ (O(1) + O(\eta))^{-1} & (O(1) + O(\eta))^{-2} \end{bmatrix}. \tag{145}$$

Consider now the matrix $(I - \mathcal{F}_\eta)^{-1} = (\mathcal{V} \otimes_b \mathcal{V})(I - \overline{\mathcal{F}}_\eta)^{-1}(\mathcal{V} \otimes_b \mathcal{V})^\top$. It can be verified that:

$$\begin{aligned} &(I - \mathcal{F}_\eta)^{-1} \\ &= (\mathcal{V} \otimes_b \mathcal{V})X^{-1}(\mathcal{V} \otimes_b \mathcal{V})^\top \\ &\quad + (\mathcal{V} \otimes_b \mathcal{V})X^{-1}Y(I + X^{-1}Y)^{-1}X^{-1}(\mathcal{V} \otimes_b \mathcal{V})^\top \\ &= \sum_{m=1}^{N}\sum_{p=1}^{N}[(v_m \otimes I_M) \otimes_b (v_p \otimes I_M)]\left([Z_{mm}]_{pp}\right)^{-1} \cdot \\ &\quad [(v_m^\top \otimes I_M) \otimes_b (v_p^\top \otimes I_M)] + \mu^{-1}(O(1) + O(\eta))^{-1} \end{aligned} \tag{146}$$

APPENDIX D
PROOF OF THEOREM 1

From (65) and (146), we have:

$$(\text{bvec}(\mathcal{Y}^\top))^\top (I-\mathcal{F}_\eta)^{-1}\text{bvec}(I_{MN}) = O(\mu)(O(1)+O(\eta))^{-1} + \sum_{m=1}^{N}\sum_{p=1}^{N}(\text{bvec}(\mathcal{Y}^\top))^\top[(v_m\otimes I_M)\otimes_b(v_p\otimes I_M)] \cdot ([Z_{mm}]_{pp})^{-1}[(v_m^\top\otimes I_M)\otimes_b(v_p^\top\otimes I_M)]\text{bvec}(I_{MN}) \quad (147)$$

where the bvec operation is relative to blocks of size $M\times M$. Using the property $\text{bvec}(\mathcal{ACB}) = (\mathcal{B}^\top\otimes_b\mathcal{A})\text{bvec}(\mathcal{C})$, we obtain:

$$[(v_m^\top\otimes I_M)\otimes_b(v_p^\top\otimes I_M)]\text{bvec}(I_{MN}) = \begin{cases}\text{bvec}(I_M)=\text{vec}(I_M), & \text{if } m=p\\ 0, & \text{if } m\neq p\end{cases} \quad (148)$$

and we conclude that:

$$(\text{bvec}(\mathcal{Y}^\top))^\top (I-\mathcal{F}_\eta)^{-1}\text{bvec}(I_{MN}) = O(\mu)(O(1)+O(\eta))^{-1} + \sum_{m=1}^{N}(\text{bvec}(\mathcal{Y}^\top))^\top[(u_m\otimes I_M)\otimes_b(u_m\otimes I_M)]x_m \quad (149)$$

where

$$x_m \triangleq ([Z_{mm}]_{mm})^{-1}\text{vec}(I_M). \quad (150)$$

This vector is the unique solution to the linear system of equations:

$$[Z_{mm}]_{mm}x_m = \text{vec}(I_M), \quad (151)$$

or, equivalently, by using (137):

$$\mu\left[(1-\mu\eta\lambda_m)^2(H_{mm}\otimes I)+\frac{\eta}{2}\lambda_m(2-\mu\eta\lambda_m)\right]x_m + \mu\left[(1-\mu\eta\lambda_m)^2(I\otimes H_{mm})+\frac{\eta}{2}\lambda_m(2-\mu\eta\lambda_m)I\right]x_m = \text{vec}(I_M), \quad (152)$$

Let $X_m = \text{unvec}(x_m)$. Applying the property $\text{vec}(ACB) = (B^\top\otimes A)\text{vec}(C)$, we obtain:

$$\text{vec}(X_mT_m)+\text{vec}(T_mX_m) = \text{vec}(I_M) \quad (153)$$

where

$$T_m \triangleq \mu(1-\mu\eta\lambda_m)^2H_{mm}+\frac{\mu\eta}{2}\lambda_m(2-\mu\eta\lambda_m)I. \quad (154)$$

We conclude from the above equation that X_m is the unique solution to the continuous time Lyapunov equation:

$$X_mT_m+T_mX_m = I_M, \quad (155)$$

whose solution is given by:

$$X_m = \frac{1}{2}T_m^{-1} = \frac{1}{2\mu}\left((1-\mu\eta\lambda_m)^2H_{mm}+\frac{\eta}{2}\lambda_m(2-\mu\eta\lambda_m)I\right)^{-1}. \quad (156)$$

Using the definitions (66), (67), and applying properties

$$\text{bvec}(\mathcal{ACB}) = (\mathcal{B}^\top\otimes_b\mathcal{A})\text{bvec}(\mathcal{C}), \quad \text{and} \quad \text{Tr}(\mathcal{AB}) = (\text{bvec}(\mathcal{B}^\top))^\top\text{bvec}(\mathcal{A}), \quad (157)$$

we get:

$$\begin{aligned}
&\sum_{m=1}^{N}[\text{bvec}(\mathcal{Y}^\top)]^\top[(v_m\otimes I_M)\otimes_b(v_m\otimes I_M)]\text{vec}(X_m)\\
&=\sum_{m=1}^{N}\text{Tr}\left(\text{unbvec}\{(v_m\otimes I_M)\otimes_b(v_m\otimes I_M)\text{bvec}(X_m)\}\mathcal{Y}\right)\\
&=\sum_{m=1}^{N}\text{Tr}\left((v_m\otimes I_M)X_m(v_m^\top\otimes I_M)\mathcal{Y}\right)\\
&=\mu^2\sum_{m=1}^{N}(1-\mu\eta\lambda_m)^2\text{Tr}\left((v_m^\top\otimes I_M)\mathcal{S}_\eta(v_m\otimes I_M)X_m\right)\\
&=\frac{\mu}{2}\sum_{m=1}^{N}(1-\mu\eta\lambda_m)^2\text{Tr}\left(\left(\sum_{k=1}^{N}[v_m]_k^2R_{s,k,\eta}\right)\cdot\left((1-\mu\eta\lambda_m)^2\left(\sum_{k=1}^{N}[v_m]_k^2H_{k,\eta}\right)+\frac{\eta}{2}\lambda_m(2-\mu\eta\lambda_m)I\right)^{-1}\right)\\
&=\frac{\mu}{2}\sum_{m=1}^{N}\text{Tr}\left(\left(\sum_{k=1}^{N}[v_m]_k^2H_{k,\eta}+\frac{\eta\lambda_m(2-\mu\eta\lambda_m)}{2(1-\mu\eta\lambda_m)^2}I\right)^{-1}\cdot\left(\sum_{k=1}^{N}[v_m]_k^2R_{s,k,\eta}\right)\right)
\end{aligned} \quad (158)$$

Substituting into (149) and (65), we conclude:

$$\limsup_{i\to\infty}\frac{1}{N}\mathbb{E}\|\mathcal{W}_\eta^o-\boldsymbol{\mathcal{W}}_i\|^2 = \frac{\mu}{2N}\sum_{m=1}^{N}\text{Tr}\left(\left(\sum_{k=1}^{N}[v_m]_k^2H_{k,\eta}+\frac{\eta\lambda_m(2-\mu\eta\lambda_m)}{2(1-\mu\eta\lambda_m)^2}I\right)^{-1}\cdot\left(\sum_{k=1}^{N}[v_m]_k^2R_{s,k,\eta}\right)\right)+\frac{O(\mu)}{(O(1)+O(\eta))}+O(\mu^{1+\theta_m}) \quad (159)$$

Now, according to definition (51), dividing (159) by μ and computing the limit as $\mu\to 0$, we arrive at expression (68) for the network MSD.

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, pp. 1–5.

[2] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "Learning over multitask graphs—Part I: Stability analysis," *IEEE Open J. Signal Process.*, Apr. 2020, *arXiv:1805.08535*.

[3] 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.

[4] 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.

[5] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[6] 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.
[7] 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.
[8] 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.
[9] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations Trends Mach. Learn.*, vol. 7, no. 4-5, pp. 311–801, 2014.
[10] 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.
[11] 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.
[12] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.
[13] 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.
[14] 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.
[15] A. Hassani, J. Plata-Chaves, M. H. Bahari, M. Moonen, and A. Bertrand, "Multi-task wireless sensor network for joint distributed node-specific signal enhancement, LCMV beamforming and DOA estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 518–533, Apr. 2017.
[16] 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.
[17] 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.
[18] 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.
[19] D. Hallac, J. Leskovec, and S. Boyd, "Network Lasso: Clustering and optimization in large graphs," in *Proc. ACM SIGKDD*, Aug. 2015, pp. 387–396.
[20] J. Szurley, A. Bertrand, and M. Moonen, "Distributed adaptive node-specific signal estimation in heterogeneous and mixed-topology wireless sensor networks," *Signal Process.*, vol. 117, pp. 44–60, 2015.
[21] 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.
[22] 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.
[23] 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.
[24] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.
[25] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks with common latent representations," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 563–579, Apr. 2017.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI, USA: American Mathematical Society, 1997.
[31] R. K. Ando and T. Zhang, "Learning on graph with Laplacian regularization," in *Proc. Advances Neural Inf. Process. Syst.*, Dec. 2006, pp. 25–32.
[32] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160–6173, Dec. 2016.
[33] P.-Y. Chen and S. Liu, "Bias-variance tradeoff of graph Laplacian regularizer," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1118–1122, Aug. 2017.
[34] B. T. Polyak, *Introduction to Optimization*. New York, NY, USA: Optimization Software, 1987.
[35] R. Nassif, S. Vlaski, C. Richard, and A. H. Sayed, "Learning over multitask graphs–Part II: Performance analysis," *IEEE Open J. Signal Process.*, Apr. 2020, *arXiv:1805.08547v2*.
[36] R. H. Koning, H. Neudecker, and T. Wansbeek, "Block Kronecker products and the vecb operator," *Linear Algebra Appl.*, vol. 149, pp. 165–184, Apr. 1991.
[37] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. Princeton, NJ, USA: Princeton Univ. Press, 2005.
[38] L. Isserlis, "On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables," *Biometrika*, vol. 12, no. 1/2, pp. 134–139, Nov. 1918.
[39] D. I. Shuman, P. Vandergheynst, D. Kressner, and P. Frossard, "Distributed signal processing via Chebyshev polynomial approximation," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 4, pp. 736–751, Dec. 2018.
[40] 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.
[41] 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.
[42] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1980.
[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

ROULA NASSIF (Member, IEEE) received the Dipl.-Ing. degree from the Lebanese University, Lebanon, the M.S. degree from 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 Beirut (AUB), Lebanon. Prior to joining AUB, she was a Postdoctoral Scholar with the 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. degrees in electrical and computer engineering from Compiègne University of Technology, France. He is a Full Professor with Université Côte d'Azur, France. In 2010–2015, he was distinguished as a Member of the Institut Universitaire de France (IUF).

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 Director of the French Federal CNRS Research Association ISIS (Information, Signal, Image, Vision). 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).

He 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 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. 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. He is a member of the US National Academy of Engineering and recognized as a Highly-Cited author. 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.