Automatica 137 (2022) 110134

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Distributed networked learning with correlated data☆

Lingzhou Hong *, Alfredo Garcia, Ceyhun Eksin

Industrial and Systems Engineering Department, Texas A&M University, College Station, TX 77843, United States of America

ARTICLE INFO

Article history:
Received 4 February 2021
Received in revised form 7 October 2021
Accepted 19 November 2021
Available online 10 January 2022

Keywords:
Large scale optimization problems and methods
Network-based computing systems
Learning theory
Statistical analysis
Parameter and state estimation
Multi-agent systems

ABSTRACT

We consider a distributed estimation method in a setting with heterogeneous streams of correlated data distributed across nodes in a network. In the considered approach, linear models are estimated locally (i.e., with only local data) subject to a network regularization term that penalizes a local model that differs from neighboring models. We analyze computation dynamics (associated with stochastic gradient updates) and information exchange (associated with exchanging current models with neighboring nodes). We provide a finite-time characterization of convergence of the weighted ensemble average estimate and compare this result to *federated* learning, an alternative approach to estimation wherein a single model is updated by locally generated gradient updates. This comparison highlights the trade-off between speed *vs* precision: while model updates take place at a faster rate in federated learning, the proposed networked approach to estimation enables the identification of models with higher precision. We illustrate the method's general applicability in two examples: estimating a Markov random field using wireless sensor networks and modeling prey escape behavior of flocking birds based on a publicly available dataset.

1. Introduction

The ever-growing size and complexity of data create scalability challenges for storage and processing. In certain application domains, data cannot be stored or processed in a single location due to geographical constraints or limited bandwidth. When data coming in heterogeneous and correlated streams, a model estimated based exclusively on local data may be of arbitrarily low precision. In this paper, we consider a distributed networked estimation that successfully addresses these concerns.

In the proposed approach, locally estimated linear models are updated in response to new gradient estimates for either a local loss measure (generalized least squares) or a network regularization function. Here, we analyze computation dynamics (associated with stochastic gradient updates) and information exchange (associated with exchange of current models with neighboring nodes to compute the gradient of network regularization). To undertake the analysis, we use a continuous-time approximation of the underlying stochastic difference equations, which allows the use of Ito's calculus.

In Theorem 3, we provide a finite-time characterization of convergence of the weighted ensemble average estimate and using an upper bound on a regularity (or dispersion) measure of the local models. Such upper bound is influenced by the local model exchanging rate and the network connectivity degree. The regularity measure can be made arbitrarily small for a large enough value of a network regularization parameter. In this case, the weighted ensemble average model is also arbitrarily close to any locally estimated model.

In Theorem 4, we provide a finite-time characterization of convergence of the weighted ensemble average model error. We show the rate of convergence is determined by *smallest* strong convexity parameter across all nodes (i.e. $\kappa > 0$) and the *slowest* data rate (i.e. $\mu > 0$). The asymptotic error is increasing in the *worst case* condition number $\frac{\eta}{\kappa} > 1$ where $\eta > 0$ is the maximum value of the Lipschitz (gradient smoothness) constants associated with each node. The asymptotic error is also increasing in data rate imbalance, i.e., $\frac{\mu'}{\mu} > 1$ where $\mu' > 0$ is the *fastest* data rate across all nodes. This characterization has no dependence on the dimension of the models. Hence, the characterization remains valid for higher-dimensional models as long as the worst-case condition number is bounded (i.e., $\kappa > 0$ and $\eta < \infty$).

We compare this performance characterization with that of an alternative approach known as *federated* learning (FL) (see, e.g., Li, Sahu, Talwalkar, and Smith (2020) and Yang, Liu, Chen,

☆ L. Hong and A. Garcia were partially supported by Grant AFOSR-15RT0767 and ECCS-2136206. C. Eksin was partially supported by NSF, United States ECCS-1933878 and NSF CCF-2008855. The material in this paper was partially presented at the 59th IEEE Conference on Decision and Control, December 14-18, 2020, Jeju Island, Republic of Korea. This paper was recommended for publication in revised form by Associate Editor Luca Schenato under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: hlz@tamu.edu (L. Hong), alfredo.garcia@tamu.edu (A. Garcia), eksinc@tamu.edu (C. Eksin).

https://doi.org/10.1016/j.automatica.2021.110134
0005-1098/

and Tong (2019)). In FL, a single model stored in a shared (centralized) parameter server is updated with locally generated gradient updates. While model updates take place at a faster rate in FL, the proposed networked approach to estimation enables the identification of models with higher precision. This is formalized in two corollaries to Theorem 8.

In the first corollary, a large enough network, i.e., one with at least $N > \sqrt{\frac{\mu'\eta}{\mu\kappa}}$ nodes in a connected topology, is shown to asymptotically exhibit higher average model precision. A networked estimation approach is also more robust to heterogeneity in noise distribution. With increasing disparities in noise variance, the FL approach is more vulnerable to noise. For example, if nodes with faster data rates are also noisier, the identified model estimate will inevitably be noisy. In the second corollary we show that the networked approach is guaranteed to outperform FL estimates when a measure of heterogeneity in noise variance across nodes (i.e. $\frac{\bar{\sigma}^2}{\min \sigma_k^2}$) exceeds the threshold $\sqrt{\frac{\mu'\eta}{\mu\kappa}}$. These corollaries highlight a trade-off between speed *vs* precision: while model updates take place at a faster rate in federated learning, the proposed networked approach to estimation enables the identification of models with higher precision.[1]

This paper is related to several strands of the literature. In a "divide and conquer" approach to distributed data (see, e.g., Predd, Kulkarni, and Poor (2009) and Zhang, Duchi, and Wainwright (2013, 2015)), individual nodes implement a particular learning algorithm to fit a model for their assigned data set and *upon each machine identifying a model*, an ensemble (or global) model is obtained by averaging individual models. This is similar to *ensemble* learning (see, e.g., Mendes-Moreira, Soares, Jorge, and Freire de Sousa (2012)), which refers to methods that combine different models into a single predictive model. For example, bootstrap aggregation (also referred to as "bagging") is a popular technique for combining regression models from *homogeneously* distributed data.[2] While a "divide and conquer" approaches coupled with a model averaging step can significantly reduce computing time and lower single-machine memory requirements, it relies on a single synchronized step (i.e., computing the ensemble average), which is executed *after all* machines have identified a model. In contrast, the approach considered in this paper deals with *asynchronous* real-time estimation and regularization for *heterogeneous* and *correlated* data streams.

The considered scheme is related to the literature on consensus optimization (see, e.g., Lian, Huang, Li, and Liu (2015), Nedic and Ozdaglar (2009) and Shi, Ling, Wu, and Yin (2015)) and the recent work on finding the best common linear model in convex machine learning problems (He, Bian, & Jaggi, 2018). However, the proposed approach cannot be interpreted as being based upon averaging local models as in consensus-based optimization. The algorithms proposed in Lian et al. (2015) and Shi et al. (2015) are designed for *batch* data while our approach deals with *streaming* data. For example, in Lian et al. (2015), gradient estimation noise is assumed independent and homogeneous, while in our approach, gradient estimation noise is *correlated* and *heterogeneous*. In addition, the algorithms proposed in Lian et al. (2015) and Shi et al. (2015), every node is *equally likely* to be selected at each iteration to update its local model. In contrast, in our approach, data streams are heterogeneous so that certain nodes have faster data streams and thus are more likely to update their models at any point in time. A network regularization penalty for networked learning has been analyzed in a series of papers by Chen, Richard, and Sayed (2014), Garcia, Wang, Huang, and Hong (2020), Marelli and Fu (2015), Nassif, Richard, Ferrari, and Sayed (2016) and Nassif, Vlaski, Richard, and Sayed (2020a, 2020b). In contrast, we consider a setting with heterogeneous nodes with *correlated* data streams *asynchronously* updating their respective models at *different rates* over time.

Finally, the paper is related to the literature of distributed algorithms to solve linear algebraic equations (such as those associated with generalized least squares) over multi-agent networks (see, e.g., Liu, Mou, and Morse (2015), Mou, Liu, and Morse (2015) and Wang, Zhou, Mou, and Corless (2019)). However, unlike these papers, we examine the consequences of heterogeneous and correlated noise in distributed generalized least squares estimation in the present paper.

The contributions of this paper are as follows. We develop a distributed estimation scheme that accounts for heterogeneous and correlated distributed datasets and heterogeneity in data processing speed by the nodes. We provide a finite-time characterization of convergence of the weighted ensemble average that captures the performance gap between the centralized and weighted average ensemble models as a function of the data heterogeneity and speed imbalance (Sections 3.1–3.3). Via a similar finite-time characterization of the FL performance, we show that the distributed estimation with network regularization outperforms FL when the number of nodes or noise variance across nodes is large (Section 3.4). We demonstrate the relative poor performance of FL when some sensors have access to highly noisy data in wireless sensor network (WSN) estimation of a Gaussian Markov random field (MRF) (Section 4.1). We also show the method's performance on a real dataset using weights proportional to the inverse of the locally estimated noise for local models (Section 4.2).

2. Data and processing model

2.1. Data model

We consider a set of nodes $\mathcal{V} = \{1, \dots, N\}$ with the ability to collect and process data streams $\mathbf{y}_i = \{\mathbf{y}_{i,k} \in \mathbb{R}^m | k \in \mathbb{N}^+\}$ of the form:

$$\mathbf{y}_{i,k} = \mathbf{X}_i \boldsymbol{w}^* + \varepsilon_{i,k} + \Lambda_i \xi_k, \quad i \in \mathcal{V} \tag{1}$$

where $\mathbf{X}_i \in \mathbb{R}^{m\times p}$, $\boldsymbol{w}^* \in \mathbb{R}^p$ is the ground truth vector of coefficients, $\{\varepsilon_{i,k} \in \mathbb{R}^{m\times 1} | k \in \mathbb{N}^+\}$ are independent and identically distributed random noise variables, and $\{\xi_k \in \mathbb{R}^{m\times 1} | k \in \mathbb{N}^+\}$ are independent realizations of a *common* noise which affects nodes differently according to the matrix $\Lambda_i \in \mathbb{R}^{m\times m}$. We consider Λ_i as a diagonal matrix with diagonal entries that are possibly different.

We assume individual noise random variables are zero-mean $\mathbb{E}[\varepsilon_{i,k}] = \mathbf{0}_{m\times 1}$ and independent across different subsets, i.e., $\mathbb{E}[\varepsilon_{i,k}\varepsilon_{j,k}^\intercal] = \mathbf{0}_{m\times m}$ for all i and $j \neq i$, and $\mathbb{E}\left\|\varepsilon_{i,k}\varepsilon_{i,k}^\intercal\right\| = \sigma_i^2 \mathbf{I}_m$. Also, the common noise vectors are i.i.d. with $\mathbb{E}[\xi_k] = \mathbf{0}_m$ and $\mathbb{E}\|\xi_k\|^2 = \mathbf{I}_m$. It follows the covariance matrix of the error term for $\mathbf{y}_{i,k}$ as

$$\Omega_i := \mathbb{E}[(\varepsilon_{i,k} + \Lambda_i\xi_k)(\varepsilon_{i,k} + \Lambda_i\xi_k)^\intercal] = \sigma_i^2 \mathbf{I} + \Lambda_i^2 \in \mathbb{R}^{m\times m},$$

with $\mathbf{X} = [\mathbf{X}_1^\intercal, \dots, \mathbf{X}_N^\intercal]^\intercal$ and $\mathbf{y}_k = [\mathbf{y}_{1,k}^\intercal, \dots, \mathbf{y}_{N,k}^\intercal]^\intercal$. We can combine data streams as follows:

$$\mathbf{y}_k = \mathbf{X}\boldsymbol{w}^* + \varepsilon_k + \Lambda\xi_k, \tag{2}$$

where $\Lambda = [\Lambda_1, \dots, \Lambda_N]^\intercal$ and $\varepsilon_k = [\varepsilon_{1,k}^\intercal, \dots, \varepsilon_{N,k}^\intercal]^\intercal$. The covariance matrix for the error terms is:

$$\Omega := \mathbb{E}[(\varepsilon_k + \Lambda\xi_k)(\varepsilon_k + \Lambda\xi_k)^\intercal] = \Sigma + \Lambda\Lambda^\intercal \in \mathbb{R}^{m\times m},$$

[1] We note that the preliminary version of this study appeared as a conference publication (Hong, Garcia, & Eksin, 2020). The model described here incorporates heterogeneity in the speed of data processing across nodes and does not commit to a particular choice of regularization weights, unlike the preliminary model (Hong et al., 2020). We present convergence results (Theorems 3 and 4) similar to Hong et al. (2020) with these generalizations. These generalizations provide additional insights into the implications of data rate imbalance combined with the heterogeneity of data. In addition, we provide an analytical comparison of the method's performance with FL (Theorem 8, and Corollaries 9 and 10).

[2] A careful selection of weights for computing the average model ensures a reduction of estimation variance along with other desirable properties, see, e.g., Hansen (2007) and Liu, Okui, and Yoshimura (2016).

where Σ is a block-diagonal matrix with the ith block equal to $\sigma_i^2\mathbf{I}_m$ and hence the noise across subsets is correlated. A centralized formulation of the generalized least squares (GLS) consists of minimizing the following loss function over $\boldsymbol{w}$:

$$\mathcal{L}_c \triangleq \frac{1}{2}\mathbb{E}[(\mathbf{y}_k - \mathbf{X}\boldsymbol{w})^\intercal \Omega^{-1}(\mathbf{y}_k - \mathbf{X}\boldsymbol{w})]. \tag{3}$$

2.2. A network of "Local" learners

For distributed data-processing we consider an undirected network structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where an edge $(i, j) \in \mathcal{E}$ represents the ability to exchange information between nodes i and j. This is also represented by the adjacency matrix $A \in \mathbb{R}^{N\times N}$ with $a_{i,j} = 1$ if $(i, j) \in \mathcal{E}$, and $a_{i,j} = 0$ otherwise.

In a distributed and networked estimation approach, each node i solves the following "localized" convex optimization problem,

$$\min_{\boldsymbol{w}_i}\{f_i(\boldsymbol{w}_i) + \delta\rho_i(\boldsymbol{w})\}, \tag{4}$$

where

$$f_i(\boldsymbol{w}_i) = \frac{1}{2}\mathbb{E}[(\mathbf{y}_{i,k} - \mathbf{X}_i\boldsymbol{w}_i)^T \Omega_i^{-1}(\mathbf{y}_{i,k} - \mathbf{X}_i\boldsymbol{w}_i)] \tag{5}$$

is a local loss function or measure of model fit, and $\rho_i(\boldsymbol{w}) \geq 0$ is a measure of similarity of identified models in the neighborhood of node i so that the term $\delta\rho_i(\boldsymbol{w})$ in (4) can be seen as a network regularization penalty. Here we consider L_2 norm regularization:

$$\rho_i(\boldsymbol{w}) = \frac{1}{2}\sum_{j\neq i}\hat{\alpha}_j a_{i,j}\left\|\boldsymbol{w}_i - \boldsymbol{w}_j\right\|^2, \tag{6}$$

where we have $\rho_i(\boldsymbol{w}) = 0$ if and only if $\boldsymbol{w}_j = \boldsymbol{w}_i$ for all nodes $j \neq i$ with $a_{i,j} = 1$. Here $\hat{\alpha}_j > 0$ is the weight associated to the jth neighbor. We shall return to the choice of these weights in the corollaries to the main result (Theorem 4) and in our numerical illustrations. For example, the literature on the optimal combination of forecasts (see Bates and Granger (1969) and Granger and Ramanathan (1984)) suggest a choice of weights of the form $\hat{\alpha}_i = \frac{1}{\mathrm{tr}(\Omega_i)}$. We note that these weights, i.e., the local data covariance matrices, are not available in a real dataset. Thus, they need to be estimated in practice, as we show in numerical examples (Section 4.2).

The choice of network regularization parameter $\delta > 0$ also plays an important role. When $\delta = 0$, node i ignores the neighboring models and finds the model that minimizes local loss. For large values of $\delta > 0$, node i will favor a model closer to neighboring models, possibly at the expense of increased local loss. A similar network regularization term has been successfully used in a networked approach to multi-task learning (see Chen et al. (2014) and Nassif et al. (2016, 2020a, 2020b)).

2.3. Stochastic gradient with network regularization

In what follows, we introduce a real-time model of the computation and communication processes involved in a distributed and networked estimation approach based upon individual solutions of problem (4). Specifically, we assume each node implements stochastic gradient updates subject to an additive network regularization penalty (SGN) based upon a noisy gradient estimate. Namely, upon collecting data point $\mathbf{y}_{i,k}$, node i is able to compute the following gradient estimate:

$$\nabla f_{i,k} = \mathbf{X}_i^\intercal \Omega_i^{-1}(\mathbf{X}_i\mathbf{w}_{i,k} - \mathbf{y}_{i,k}) = g_{i,k} + \mathbf{X}_i^\intercal \Omega_i^{-1}(\varepsilon_{i,k} + \Lambda_i\xi_k), \tag{7}$$

where $g_{i,k} := \nabla_{\mathbf{w}_i} f_i(\mathbf{w}_{i,k}) = \mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i(\mathbf{w}_{i,k} - \mathbf{w}^*)$. Hence, the basic iteration of SGN takes the form:

$$\boldsymbol{w}_{i,k+1} = \boldsymbol{w}_{i,k} - \gamma[\nabla f_{i,k} + \delta\nabla\rho_{i,k}], \tag{8}$$

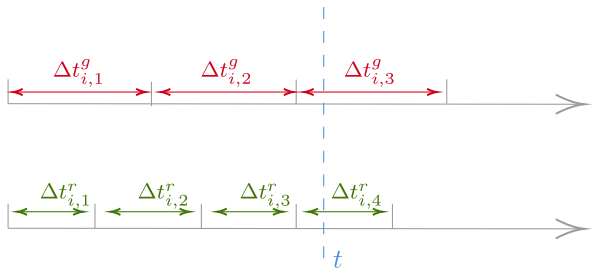

Fig. 1. By time $t > 0$, node i has executed *two* gradient updates and *three* network regularization updates ($N_{g,i}(t) = 2$ and $N_{r,i}(t) = 3$).

where $\nabla\rho_{i,k}$ is the gradient of the network regularization penalty $\rho_i(\boldsymbol{w}_k)$ and $\gamma > 0$ is the step size.

Remark 1. The basic iteration in (8) is related to the literature on consensus optimization (see e.g. Lian et al. (2015), Nedic and Ozdaglar (2009) and Shi et al. (2015)). However, the proposed approach cannot be interpreted as being based upon *averaging over local* models as in consensus-based optimization. In that literature, the basic iteration is of the form:

$$\boldsymbol{w}_{i,k+1} = \sum_j \boldsymbol{W}_{i,j}\boldsymbol{w}_{j,k} - \gamma\nabla f_{i,k}, \tag{9}$$

where $\boldsymbol{W}_k \in \mathbb{R}^{N\times N}$ is doubly stochastic. Indeed one can rewrite (8) as the form of (9) with $\boldsymbol{W}_{i,i} = 1 - \gamma\delta\sum_j \hat{\alpha}_j a_{i,j}$ and $\boldsymbol{W}_{i,j} = \gamma\delta\hat{\alpha}_j a_{i,j}$. However, the resulting matrix $\boldsymbol{W}$ is not necessarily doubly stochastic in general. Thus, the basic iteration in (8) cannot be interpreted as being based upon *averaging over local* models as in consensus-optimization.

2.4. Real-time implementation and continuous time approximation

The implementation of (8) requires that at every $k \in \mathbb{N}^+$, node i has access to the current estimates of its neighbors $\{\boldsymbol{w}_{j,k}\}_{(i,j)\in\mathcal{E}}$. Therefore to account for the real-time implementation of such updates we must model the random times required for *(i)* computing gradient estimates $\nabla f_{i,k}$, *(ii)* collect updated model parameters from neighboring nodes in order to compute the gradient of network penalty $\nabla\rho_{i,k}$. For the dynamic *(i)*, define $\Delta t^g_{i,k}$ as the random time for node i to compute $\nabla f_{i,k}$ and $t^g_{i,k} = \sum_l^k \Delta t^g_{i,l}$ as the time-point to obtain $\nabla f_{i,k}$. For the dynamic *(ii)*, $t^r_{i,k}$ marks the time at which the kth penalty gradient update is executed by node i and $\Delta t^r_{i,k} = t^r_{i,k} - t^r_{i,k-1}$ is the time required by node i to collect updated models from neighbors and compute the penalty gradient (see Fig. 1 for illustration).

We assume $\Delta t^g_{i,k}$'s are i.i.d. with $\mathbb{E}[\Delta t^g_{i,k}] = \Delta t^g_i$ and $\Delta t^r_{i,k}$'s are i.i.d with $\mathbb{E}[\Delta t^r_{i,k}] = \Delta t^r$ (i.e. all nodes execute penalty updates at the same rate). We define the counting processes $N_{g,i}(t) = \max\{k \in \mathbb{N}^+ | t^g_{i,k} < t\}$ and $N_{r,i}(t) = \max\{k \in \mathbb{N}^+ | t^r_{i,k} < t\}$, then $w.p.1$,

$$\lim_{t\to\infty}\frac{N_{r,i}(t)}{t} = \frac{1}{\Delta t^r} := \hat{\beta}, \quad \lim_{t\to\infty}\frac{N_{g,i}(t)}{t} = \frac{1}{\Delta t^g_i} := \mu_i.$$

A continuous-time embedding of $\mathbf{w}_{i,t}$ is obtained as:

$$\boldsymbol{w}_{i,t} = \boldsymbol{w}_{i,0} - \gamma\sum_{k=1}^{N_{g,i}(t)}\nabla f_{i,k} - \gamma\delta\sum_{k=1}^{N_{r,i}(t)}\nabla\rho_{i,k} \tag{10}$$

Define a re-scaled process $\mathbf{w}_{i,t} := \boldsymbol{w}_{i,t/\gamma}$, and by Donsker theorem (Donsker, 1951), the rescaled noise terms can be

approximated by a Wiener processes as $\gamma \to 0$. We show that the dynamics of $\mathbf{w}_{i,t}$ can be modeled via the stochastic differential equation (see Appendix A.1):

$$\mathrm{d}\mathbf{w}_{i,t} = -(\mu_i g_{i,t} + \delta\hat{\beta}\nabla\rho_{i,t})\mathrm{d}t + \tau_i \mathbf{X}_i^\intercal \Omega_i^{-1}\mathrm{d}B_{i,t} + \varsigma_i \mathbf{X}_i^\intercal \Omega_i^{-1}\Lambda_i \mathrm{d}B_t, \tag{11}$$

where $\tau_i = \sigma_i\sqrt{\gamma\mu_i}$, and $\varsigma_i = \sqrt{\gamma\mu_i}$.[3] Here $B_{i,t}$ and B_t are the standard m dimensional Brownian Motion approximating the individual noise associated with node i and the common noise, respectively. In the following, we characterize the convergence of the SGN scheme defined in (8) via the continuous-time approximation (11).

3. Convergence analysis

The processes $\{w_{i,t} : t > 0\}$ converge weakly (i.e. in distribution) (see Appendix A.2). To characterize convergence we will use measures of *consistency* and *regularity*. Let $\hat{\mathbf{w}}_t$ denote a weighted average solution at time t, i.e.,

$$\hat{\mathbf{w}}_t = \sum_{i=1}^N \alpha_i \mathbf{w}_{i,t}, \tag{12}$$

where $\alpha_i := \frac{\hat{\alpha}_i}{c} \in (0,1)$ are *normalized* weights with $c := \sum_{i=1}^N \hat{\alpha}_i$. Let $V_{i,t} = \left\|\mathbf{e}_{i,t}\right\|^2/2$, where $\mathbf{e}_{i,t} := \mathbf{w}_{i,t} - \hat{\mathbf{w}}_t$. To measure regularity, we will use the weighted average difference between the solutions obtained from a single node and that of the ensemble (weighted) average:

$$\bar{V}_t = \sum_{i=1}^N \frac{\alpha_i \left\|\mathbf{w}_{i,t} - \hat{\mathbf{w}}_t\right\|^2}{2} = \sum_{i=1}^N \alpha_i V_{i,t}. \tag{13}$$

To measure consistency we will examine the distance between the average and the ground truth,

$$U_t = \frac{1}{2}\left\|\hat{\mathbf{w}}_t - \mathbf{w}^*\right\|^2. \tag{14}$$

3.1. Preliminaries

We will make use of the following definitions and results in the convergence analysis. Let the Laplacian matrix of $\mathcal{G}$ be $L = D - A$, where D is the degree matrix, and A is the adjacency matrix. We define the generalized Laplacian matrix as $\hat{L} = \hat{D} - \hat{A}$, where $\hat{D}$ is a diagonal matrix whose ith diagonal entry is equal to $\sum_{j\in\mathcal{V}} \hat{\alpha}_i\hat{\alpha}_j a_{i,j}$, and $\hat{A}$ is the weighted adjacent matrix with $A_{i,j} = \hat{\alpha}_i\hat{\alpha}_j a_{i,j}$. Let λ_2 (respectively, $\hat{\lambda}_2$) denote the second smallest eigenvalue of L (respectively, $\hat{L}$).

The continuous-time gradient $g_{i,t}$ defined above is a function of $\mathbf{w}_{i,t}$, which we do not explicitly specify to simplify notation. In our analyses, we denote $g_{i,t}(\hat{\mathbf{w}}_t) = \mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i(\hat{\mathbf{w}}_t - \mathbf{w}^*)$ and $g_{i,t}(\mathbf{w}^*) = \mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i(\mathbf{w}^* - \mathbf{w}^*)$. Note that $g_{i,t}(\mathbf{w}^*) = 0$ for all $i \in \mathcal{V}$ and t, to simplify notation we will write $g(\mathbf{w}^*)$ instead. Similarly, when a property holds for all t, we drop t and write $g_{i,t}$ as g_i.

We note that g_i's are Lipschitz continuous and the corresponding loss function (noise-free version of f_i) is strongly convex with parameter κ_i. Let $\mathbf{w}_{i,1}$ and $\mathbf{w}_{i,2}$ be two input vectors taken from the function domain, then

$$\begin{aligned}\left\|g_i(\mathbf{w}_{i,1}) - g_i(\mathbf{w}_{i,2})\right\| &= \left\|\mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i(\mathbf{w}_{i,1} - \mathbf{w}_{i,2})\right\| \\ &\le \eta_i \left\|\mathbf{w}_{i,1} - \mathbf{w}_{i,2}\right\| \le \eta \left\|\mathbf{w}_{i,1} - \mathbf{w}_{i,2}\right\|,\end{aligned} \tag{15}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\eta_i := \left\|\mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i\right\|$, $\eta := \max_i \left\|\mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i\right\|$. Furthermore,

$$(g_i(\mathbf{w}_{i,1}) - g_i(\mathbf{w}_{i,2}))^\intercal(\mathbf{w}_{i,1} - \mathbf{w}_{i,2}) \ge \kappa \left\|\mathbf{w}_{i,1} - \mathbf{w}_{i,2}\right\|^2, \tag{16}$$

where $\kappa := \min \kappa_i$ with $\kappa_i := 2\lambda_{min}(\mathbf{X}_i^\intercal \Omega_i^{-1}\mathbf{X}_i)$, and $\lambda_{min}(\cdot)$ is the smallest eigenvalue.

The ratio $\frac{\eta_i}{\kappa_i} > 1$ is referred to as the condition number of the deterministic optimization problem $\min_{\mathbf{w}} f_i(\mathbf{w})$. In the analysis below, the *worst-case* condition number, i.e. $\frac{\eta}{\kappa} > 1$, plays an important role in characterizing performance.

3.2. Regularity

The following result establishes that the sum of regularity and consistency measures is equivalent to the weighted sum of the distances between individual solutions $\mathbf{w}_{i,t}$ and the ground truth $\mathbf{w}^*$.

Lemma 2. *Consider $\bar{V}_t$ in (13) and U_t in (14). We have*

$$\frac{1}{2}\sum_{i=1}^N \alpha_i \left\|\mathbf{w}_{i,t} - \mathbf{w}^*\right\|^2 = \bar{V}_t + U_t. \tag{17}$$

Proof. We expand the sum on the left-hand side of (17) as follows,

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^N \alpha_i \left\|\mathbf{w}_{i,t} - \mathbf{w}^*\right\|^2 = \frac{1}{2}\sum_{i=1}^N \alpha_i \Big[&\left\|\mathbf{w}_{i,t} - \hat{\mathbf{w}}_t\right\|^2 + \\ &\left\|\hat{\mathbf{w}}_t - \mathbf{w}^*\right\|^2 + 2\langle \mathbf{e}_{i,t}, \hat{\mathbf{w}}_t - \mathbf{w}^*\rangle \Big].\end{aligned} \tag{18}$$

Note that

$$\sum_{i=1}^N \alpha_i \mathbf{e}_{i,t} = \sum_{i=1}^N \alpha_i \mathbf{w}_{i,t} - \hat{\mathbf{w}}_t = 0. \tag{19}$$

Hence, (17) follows by the fact that the summation of the cross-product (last) term inside the brackets in (18) is zero. □

In what follows, we obtain upper bounds on the expectations of regularity $\bar{V}_t$ and consistency U_t processes in Theorems 3 and 4, respectively. Given the relation in Lemma 2, these bounds provide a bound on the average error of individual estimates generated by the SGN algorithm with respect to the ground truth. The following result provides an upper bound on the expected regularity of the estimates at a given time.

Theorem 3. *Let $\mathbf{w}_{i,t}$ evolve according to continuous time dynamics (11). Then*

$$\mathbb{E}[\bar{V}_t] \le e^{-2(\mu\kappa + \delta\beta\hat{\lambda}_2)t}\bar{V}_0 + \frac{C_1}{2(\mu\kappa + \delta\beta\hat{\lambda}_2)}(1 - e^{-2(\mu\kappa+\delta\beta\hat{\lambda}_2)t}),$$

where $\beta := \frac{\hat{\beta}}{c}$ and $C_1 > 0$ is defined as follows:

$$C_1 := \frac{1}{2}\sum_{i=1}^N \alpha_i C_{1,i}, \tag{20}$$

with $C_{1,i}$ for $i \in \mathcal{V}$ defined as,

$$\begin{aligned}C_{1,i} :=& \tau_i^2\Big(1 - 2\alpha_i\Big)\left\|\mathbf{X}_i^\intercal \Omega_i^{-1}\right\|_F^2 + \varsigma_i^2\left\|\mathbf{X}_i\Omega_i^{-1}\Lambda_i\right\|_F^2 \\ &- 2\sum_{k=1}^N \alpha_k A_{i,k} + D_1 + D_2,\end{aligned}$$

where the constants $D_1 = \sum_{k=1}^N \alpha_k^2\tau_k^2\left\|\mathbf{X}_k^\intercal \Omega_k^{-1}\right\|_F^2$ and $D_2 = \sum_{k=1}^N\sum_{k=1}^N \alpha_k\alpha_j A_{k,j}$ with $A_{i,k} := \varsigma_i\varsigma_k \mathbf{1}^\intercal(\mathbf{X}_i^\intercal \Omega_i^{-1}\Lambda_i \circ \mathbf{X}_k^\intercal \Omega_k^{-1}\Lambda_k)\mathbf{1}$ and "$\circ$" denoting the Hadamard product. In the long run,

[3] The algorithm described in this section differs from the preliminary version studied in Hong et al. (2020) in two major ways: (i) we allow heterogeneity in processing of data across nodes, and (ii) we do not consider a specific set of averaging weights.

$$\lim_{t\to\infty}\mathbb{E}[\bar{V}_t] \leq \frac{C_1}{2(\mu\kappa+\delta\beta\hat{\lambda}_2)}.$$

Proof. See Appendix A.3. □

It is not surprising that the expected difference in estimates decreases with growing δ, which penalizes disagreement with neighbors. Similarly, the larger the algebraic connectivity of the network $\hat{\lambda}_2$ or the strong convexity constant κ is, the smaller is the expected $\bar{V}_t$.

Finally, the constant term C_1 is determined by $\mathbf{X}_i, i\in\mathcal{V}$ and the covariance matrices of the common noise $\Lambda_i, i\in\mathcal{V}$. Note that the quadratic variation of $\mathbf{e}_{i,t}$ is:

$$\langle \mathbf{e}_{i,t}\rangle_t = \mathbb{E}\int_0^t \frac{1}{2}C_{1,i}\mathrm{d}s \to \mathrm{Var}(\mathbf{e}_{i,t}).$$

Hence, $\frac{t}{2}C_{i,t}$ describes the variation of $\mathbf{e}_{i,t}$, and C_1 is a weighted measure of variation. C_1 is small when we have nodes that are less affected by the common noise.

3.3. Consistency

The consistency measure $\{U_t, t\geq 0\}$ captures the performance of the average solution $\hat{\mathbf{w}}$. The following result provides a characterization of the average solution.

Theorem 4. *Let $\boldsymbol{w}_{i,t}$ evolve according to continuous time dynamics (10). Then*

$$\mathbb{E}[U_t] \leq e^{-2\mu\kappa t}U_0 + \frac{1}{2\kappa\mu}\Big(\frac{\eta\mu'-\kappa\mu}{\delta\beta\hat{\lambda}_2}C_1+C_2\Big)\Big(1-e^{-2\mu\kappa t}\Big),$$

where $\beta := \frac{\hat{\beta}}{c}$ and C_1 is defined in (20), $\mu' = \max_i \mu$, and $\mu = \min_i \mu_i$, and

$$C_2 = \frac{1}{2}(D_1+D_2). \tag{21}$$

In the long run,

$$\lim_{t\to\infty}\mathbb{E}[U_t] \leq \frac{1}{2\mu\kappa}\Big(\frac{\eta\mu'-\kappa\mu}{\delta\beta\hat{\lambda}_2}C_1+C_2\Big). \tag{22}$$

Proof. See Appendix A.4. □

Similar to the regularity measure bound, the penalty constant $\delta>0$ and the algebraic connectivity $\hat{\lambda}_2$ reduce the bound on the expected consistency. However, the long-run expected difference between the ensemble average estimate and the ground truth does not reduce to zero as $\delta\hat{\lambda}_2\to\infty$. The constant C_2, determined by the data $\mathbf{X}_i$ and matrices Λ_i, captures the performance gap in the long run due to available data. Note that the quadratic variation of $\hat{\mathbf{w}}_t-\mathbf{w}^*$ is as follows,

$$\langle \hat{\mathbf{w}}_t-\mathbf{w}^*\rangle_t = \mathbb{E}\int_0^t C_2\mathrm{d}s \to \mathrm{Var}(\hat{\mathbf{w}}_t-\mathbf{w}^*).$$

Hence tC_2 describes the variation of $\hat{\mathbf{w}}_t-\mathbf{w}^*$. In the following result, we examine asymptotic performance as the network grows in size.

Remark 5. The bounds obtained in Theorems 3 and 4 also serve to bound the performance of individual estimators $\mathbf{w}_{i,t}$. Since for node i we have: $\|\mathbf{w}_{i,t}-\mathbf{w}^*\| \leq \|\mathbf{w}_{i,t}-\hat{\mathbf{w}}_t\| + \|\hat{\mathbf{w}}_t-\mathbf{w}^*\| < \sqrt{2}(\sqrt{\frac{\bar{V}_t}{\alpha_i}}+\sqrt{U_t})$.

3.4. Asymptotic network performance

Lemma 6 (*Asymptotic Performance*)**.** *Assume the network of local learners (with associated data sets) grows as follows:*

- *Each new node (say $n>N$, $N=1,2,\dots$) is associated with a new dataset $\boldsymbol{X}_n$ and $\boldsymbol{y}_{n,k}$ given by (1) with*
$$\|X_n\|_F < L_0, \quad tr(\Omega_n)\leq M, \text{ and } \epsilon\leq\sigma_n^2,$$
where $L_0<\infty$, and $0<\epsilon<M<\infty$. We assume all entries of the weight matrix Λ_i's are bounded above by ω_2, the maximum entry of all weight matrices.
- *The network connectivity is preserved.*

If we have $\delta\hat{\lambda}_2\sim N$, and $\gamma\sim\epsilon^3$, then

$$\lim_{t\to\infty}\lim_{N\to\infty}\mathbb{E}[\|\hat{\boldsymbol{w}}_t-\boldsymbol{w}^*\|^2] < \frac{S_2\mu'\gamma m\,\omega_2}{\kappa\mu\epsilon^2} = O(\gamma^{1/3}),$$

where $S_2=\max_{k,j}\max_i \boldsymbol{x}_{i,k}^\intercal\boldsymbol{x}_{i,j}$ with $\boldsymbol{x}_{i,k}$ being the kth column of $\boldsymbol{X}_i^\intercal$.

Proof. See Hong, Garcia, and Eksin (2021) section 5.6 for details. □

The proof follows by constructing an upper bound for (22) in Theorem 4 by considering constants C_1 and C_2. We first show that all terms of C_2 can be bounded by $\frac{S_2\mu'\gamma m\omega_2}{\epsilon^2}$ by using the assumptions on each new dataset available to each new node. This bound on C_2 increases with the noise term that affects the nodes, inputs magnitude, and step size. Second, show that C_1/N goes to zero as $N\to\infty$. Combining the limiting properties of the two constants with the bound in (22) and selecting δ such that $\delta\hat{\lambda}_2\sim N$, we obtain the bound above. This bound can be controlled by the selection of the step size γ. Note that we can make $\delta\hat{\lambda}_2\sim N$ by adjusting the penalty parameter when the network retains connectivity as the number of nodes grows.

Remark 7. Lemma 6 provides an upper bound of the difference between the ensemble average estimate and the ground truth, which is $O(\gamma^{1/3})$ as $N\to\infty$. The bound is subject to the influence of the data: smaller in magnitude (smaller S_2) leads to a tighter bound. We observe that the bound is tighter when the objective function is smoother (larger κ). Since we require $\gamma\sim\epsilon^3$ (or even smaller γ), the effect of having small ϵ is offset by choosing smaller γ. Thus the bound can be controlled as small as needed by choosing the stepsize γ.

3.5. Comparison with federated learning (FL)

In this section, we firstly present the convergence property of FL, and then compare its performance with SGN.

3.5.1. Federated learning

For the federated learning, the gradient samples are sent to a central server for computing, we define the FL approximation at t as $\mathbf{w}_t=\sum_i \mathbf{w}_{i,t}$, and the continuous time embedding of FL updates is as follows:

$$\boldsymbol{w}_t = \boldsymbol{w}_0 - \gamma\sum_{i=1}^{N}\sum_{k=1}^{N_{i,g}(t)}\nabla f_{i,k}. \tag{23}$$

Define a rescaled variable $\mathbf{w}_t := \boldsymbol{w}_{t/r}$, the continuous time dynamics of $\mathbf{w}_t$ can be approximated by:

$$\mathrm{d}\mathbf{w}_t = \sum_{i=1}^{N}\Big[-\mu_i g_{i,t}\mathrm{d}t + \tau_i\mathbf{X}_i^\intercal\Omega_i^{-1}\mathrm{d}B_{i,t} + \varsigma_i\mathbf{X}_i^\intercal\Omega_i^{-1}\Lambda_i\mathrm{d}B_t\Big].$$

Now we define the measure F_t to examine the distance between the FL estimates and the ground truth and present the bounds on the expectation of $F_t=\frac{1}{2}\|\mathbf{w}_t-\mathbf{w}^*\|^2$ in the following theorem:

Theorem 8. *Let $\boldsymbol{w}_t$ evolve according to continuous time dynamics (23), then*

$$dF_t = -\sum_{i=1}^{N} \mu_i g_{i,t}^{\intercal}(\boldsymbol{w}_t - \boldsymbol{w}^*)dt + K_3 d\tilde{B}_{f,t} + C_3 dt,$$

where $K_3 d\tilde{B}_{f,t}$ is the summation of the Ito terms,

$$\begin{aligned} K_3 d\tilde{B}_{f,t} = &\sum_{i=1}^{N} \tau_i (\boldsymbol{w}_t - \boldsymbol{w}^*)^{\intercal} \boldsymbol{X}_i^{\intercal} \Omega_i^{-1} dB_{i,t} \\ &+ \sum_{i=1}^{N} \varsigma_i (\boldsymbol{w}_t - \boldsymbol{w}^*)^{\intercal} \boldsymbol{X}_i^{\intercal} \Omega_i^{-1} \Lambda_i dB_t, \end{aligned} \tag{24}$$

and C_3 is the summation of the constant terms,

$$C_3 = \frac{1}{2}\Big(\sum_{i=1}^{N} \tau_i^2 \left\|\boldsymbol{X}_i^{\intercal} \Omega_i^{-1}\right\|_F^2 + \sum_{k=1}^{N}\sum_{j=1}^{N} A_{k,j}\Big). \tag{25}$$

Moreover,

$$\mathbb{E}[F_t] \geq e^{-2\eta\mu' t} F_0 + \frac{1}{2\eta\mu'} C_3 (1 - e^{-2\eta\mu' t}), \tag{26}$$

$$\mathbb{E}[F_t] \leq e^{-2\kappa\mu t} F_0 + \frac{1}{2\kappa\mu} C_3 (1 - e^{-2\kappa\mu t}), \tag{27}$$

so that in the long run,

$$\frac{C_3}{2\eta\mu'} \leq \lim_{t\to\infty} \mathbb{E}[F_t] \leq \frac{C_3}{2\kappa\mu}.$$

Proof. See Appendix A.5. □

In this theorem, we provide both upper and lower bounds of F_t. The strong convexity parameter κ and the Lipschitz constant η influence the bounds. The constant C_3, determined by the data $\mathbf{X}_i$ and matrices Λ_i determines the quality of the asymptotic estimator derived by federated learning, and tC_3 describes the variation of $\mathbf{w}_t - \mathbf{w}^*$.

3.5.2. Comparison with federated learning

In this section, we compare the long run performance of the SGN scheme and FL scheme with streaming data. Specifically, we compare the upper bound in Theorem 4 and the lower bound in Theorem 8. This is tantamount to comparing $\frac{C_2}{\mu\kappa}$ and $\frac{C_3}{\mu'\eta}$ for large enough values of $\delta > 0$.

Corollary 9. *Assume $\hat{\alpha}_i = 1, i \in \mathcal{V}$, so that $\alpha_i = \frac{1}{N}$, (i.e. simple average) and $N > \sqrt{\frac{\mu'\eta}{\mu\kappa}}$. There exists $\bar{\delta} < \infty$ such that for all $\delta > \bar{\delta}$ it holds that:*

$$\lim_{t\to\infty} E[U_t] < \lim_{t\to\infty} E[F_t].$$

Proof. If the weights are of the form $\alpha_i = \frac{1}{N}$ (i.e. simple average) then

$$C_2 = \frac{1}{2N^2}\left(\sum_{k=1}^{N} \tau_k^2 \left\|\mathbf{X}_k^{\intercal} \boldsymbol{\Omega}_k^{-1}\right\|_F^2 + \sum_{k=1}^{N}\sum_{j=1}^{N} A_{k,j}\right) = \frac{C_3}{N^2}.$$

It follows that for $N > \sqrt{\frac{\mu'\eta}{\mu\kappa}}$, we have $\frac{C_2}{\mu\kappa} < \frac{C_3}{\mu'\eta}$, and the lower bound in Theorem 8 exceeds the upper bound in Theorem 4 whenever:

$$\frac{C_2}{\mu\kappa} + \frac{1}{\mu\kappa}\left(\frac{\mu'\eta - \mu\kappa}{\delta\beta\lambda_2}\right) < \frac{C_3}{\mu'\eta}, \tag{28}$$

Hence, for $\delta > \bar{\delta}$ with $\bar{\delta} := \frac{1}{\beta\lambda_2}\left(\frac{\mu'\eta}{\mu\kappa} - 1\right)\left(\frac{C_3}{\mu'\eta} - \frac{C_2}{\mu\kappa}\right)^{-1}$, the inequality (28) holds and the result follows. □

The above result guarantees that a large enough network ensures an ensemble averages estimate with higher quality than that obtained with federated learning. In the following corollary, we show that even for small networks but highly heterogeneous datasets (high asymmetries in individual noise), the SGN ensemble estimate is superior to the estimate obtained through federated learning.

Corollary 10. *Consider the case with $\Lambda_i = 0$ for all $i \in \mathcal{V}$, $\hat{\alpha}_i = \frac{1}{\sigma_i^2}$, and $c = \sum_{k=1}^{N} \frac{1}{\sigma_k^2}$. Let $\bar{\sigma} > 0$ be defined as:*

$$\frac{1}{\bar{\sigma}^4} := \frac{\sum_{k=1}^{N} \frac{1}{\sigma_k^4} \mu_k \left\|\mathbf{X}_k\right\|_F^2}{\sum_{k=1}^{N} \mu_k \left\|\mathbf{X}_k\right\|_F^2}.$$

If $\frac{\bar{\sigma}^2}{\min \sigma_k^2} > \sqrt{\frac{\mu'\eta}{\mu\kappa}}$, there exists $\bar{\delta} < \infty$, such that for all $\delta > \bar{\delta}$, it holds that

$$\lim_{t\to\infty} E[U_t] < \lim_{t\to\infty} E[F_t].$$

Proof. In this case, $\boldsymbol{\Omega}_k^{-1} = \frac{1}{\sigma_k^2}\boldsymbol{I}_k$, therefore $\left\|\mathbf{X}_k^{\intercal}\boldsymbol{\Omega}_k^{-1}\right\|_F^2 = \frac{1}{\sigma_k^2}\left\|\mathbf{X}_k\right\|_F^2$. Also, $\alpha_i = \frac{\hat{\alpha}_i}{c}$, and C_2 and C_3 become

$$C_2 = \frac{\gamma}{2}\sum_{k=1}^{N} \alpha_k^2 \sigma_k^2 \mu_k \left\|\mathbf{X}_k^{\intercal}\boldsymbol{\Omega}_k^{-1}\right\|_F^2 = \frac{\gamma}{2c^2}\sum_{k=1}^{N} \frac{\mu_k}{\sigma_k^4}\left\|\mathbf{X}_k\right\|_F^2,$$

$$C_3 = \frac{\gamma}{2}\sum_{k=1}^{N} \sigma_k^2 \mu_k \left\|\mathbf{X}_k^{\intercal}\boldsymbol{\Omega}_k^{-1}\right\|_F^2 = \frac{\gamma}{2}\sum_{k=1}^{N} \mu_k \left\|\mathbf{X}_k\right\|_F^2.$$

Hence, $\frac{C_3}{C_2} = c^2\bar{\sigma}^4$. Since $\frac{1}{c} \leq \min_k \sigma_k^2$, we conclude that if $\frac{\bar{\sigma}^2}{\min_k \sigma_k^2} > \sqrt{\frac{\mu\kappa}{\mu'\eta}}$, then $\frac{C_2}{\mu\kappa} < \frac{C_3}{\mu'\eta}$. The rest of the proof follows the same argument to the previous corollary. □

This second corollary indicates that with large differences in individual noise variance, the SGN approach provides a higher quality asymptotic estimate than federated learning. Corollaries 9 and 10 indirectly point to conditions under which federated learning is likely to outperform the considered networked approach. In the networked approach, the asymptotic estimation quality benefits from a noise averaging effect due to network regularization (i.e., the term $\frac{1}{N^2}$). With *homogeneous* rates (i.e., $\mu' = \mu$) and *well-conditioned* individual estimation problems (i.e. $\frac{\eta}{\kappa} \approx 1$), the sufficient condition in Corollary 9 requires a relatively *small* network size. With *heterogeneous* rates and *ill-conditioned* individual estimation problems, the requirement on network size is more stringent, and it is more likely that federated learning outperforms the proposed approach. Similarly, in Corollary 10, with *homogeneous* noise distribution (i.e. $\sigma_i^2 = \sigma_2^2 = \sigma$) it follows that $\frac{\hat{\sigma}^2}{\min_k \sigma_k^2} = 1$. Since $\eta > \kappa$, the sufficient condition in Corollary 10 *cannot be* met. To sum up, federated learning is likely to outperform the networked approach whenever the network is small relative to a measure of heterogeneity which accounts for individual estimation problems' conditioning and heterogeneity in noise distribution.

4. Numerical illustrations

We apply the proposed method to two examples to corroborate the analytical results. First, we apply the SGN algorithm to a MRF estimation problem using a WSN with synthetic data and compare our scheme with FL. Next, we look at a real-world problem: the escape behavior of European gregarious birds. In this example, we examine the robustness of the networked approach when the covariance matrices are recursively estimated.

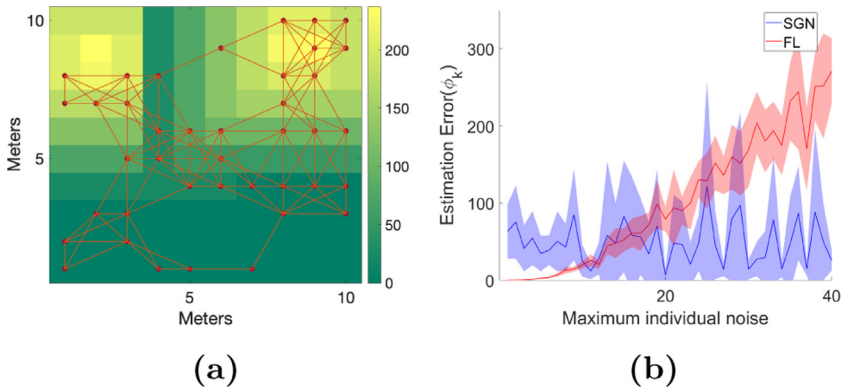

Fig. 2. (a) Network structure of sensors. The orange dots denote the nodes, and the lines represent the edges between nodes. The two heat sources are located at (2 m, 8.5 m) and (8.5 m, 9 m), marked by yellow. (b) The 95% confidence intervals of average estimation error $\bar{\phi}_k$ by SGN and FL at the final iteration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1. Temperature estimation of a field

We consider a WSN deployed to estimate the temperature on a 10 m $\times$ 10 m field, divided into 100 equal squares. We assume that the temperature within the same square is the same, and the field's true temperature is stored in the vector $\mathbf{w}^* \in \mathbb{R}^{100\times 1}$. We randomly place N sensors on the field, and each measures $\mathbf{w}^*$ using noisy local observations $\mathbf{y}_i \in \mathbb{R}^{100\times 1}$, which is corrupted by the measurement noise ε_i, a detection error that only influence sensor i; and the network disturbance ξ, a common noise that is shared by all sensors. Each sensor i only shares a portion of ξ, which is determined by a matrix Λ_i. Λ_i values reflect the relative distance between the sensors' locations and the measured squares: a sensor that is close to a square is subject to lower noise levels. We assume $\mathbf{w}^*$ is fixed but $\mathbf{y}_i$ changes at each measurement, which can be expressed as follows:

$$\mathbf{y}_i = \mathbf{w}^* + \varepsilon_i + \Lambda_i \xi, \tag{29}$$

where $\varepsilon_i \sim \mathcal{N}_m(\mathbf{0}, \sigma^2\mathbf{I})$ and $\xi \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I})$. Note that if we set $\mathbf{X}_i = \mathbf{I}$, (1) and (29) have the same form. The local lost function for each node is as (4).

We model the temperature of the field using a Gaussian MRF and let the temperature values range from 0 °F to 255 °F, as in Eksin and Ribeiro (2012). Two heat locations are located at (2, 8.5) and (8.5, 9), and temperature drops from the heat source at a rate of 25 °F/m within an area of influence of 5 m from the source. We set $N = 40$ and randomly connect the nodes to their neighbors within 2.5 m—see Fig. 2(a) for sensor locations and heat map of the field.

We would like to minimize the cost function using SGN and FL for each sensor by selecting proper $\mathbf{w}_i$. We set the stepsize $\gamma = 10^{-5}N$ for SGN and $\gamma = 10^{-5}$ for FL, the penalty parameter is set as $\delta = 100$ for SGN. We set the minimum individual noise variance as 0.01 and let the maximum change from 1 to 40. The stream parameters are set as $\mu_i = 10\sigma_i^2$, i.e., the faster nodes also generate noiser data. We define the estimation error at time point k from the lth trial as $\phi_{k,l} = \left\|\hat{\mathbf{w}}_k^{(l)} - \mathbf{w}^*\right\|$ where $\hat{\mathbf{w}}_k^{(l)}$ is the weighted average (12) from the lth trial. The average estimation error over K trials is defined as: $\bar{\phi}_k = \frac{1}{K}\sum_{l=1}^{K}\phi_{k,l}$. We run both algorithms for 10 trials and show the average estimation error with 95% confidence interval at iteration $T = 4000$ in Fig. 2(b). It shows at the final iteration, as the maximum individual noise variance increases, the average estimation error of FL increases while that of SGN is stable, which suggests that SGN is robust against noise.

4.2. Modeling flocking escape behavior

In this section, we consider a real data-set used for modeling bird escape behavior based upon the Flight Initiation Distance (FID), the distance at which animals take flight from approaching threats. In a regression model, the FID is considered as the response variable, and 7 birds' behaviors are considered as predictors. (See Morelli et al. (2019) for details.) We normalize all variables for the following analysis. At each node, the FID estimate is given by a linear model, where we denote node i's estimate with $\mathbf{w}_i = [w_0, \ldots, w_7]$ as before. The data contains 941 observations in total collected from eight European countries and 23 different bird species. We group 23 bird species into $N = 15$ nodes where each species is assigned to one node. Since all nodes contain more than 10 observations, we use the mini-batch of size 10 in the experiment to unify the data length at different nodes. In this experiment, we consider a complete network (a network with lower connectivity may only reduce the performance slightly). We use $\hat{\alpha}_i = \frac{1}{\mathrm{tr}(\Omega_i)}$ for each i, and the covariance matrix of a node is computed as the diagonals of the covariance matrix of 15 mini-batch samples. We use a fading memory update rule to compute the trace of the covariance matrix, $\mathrm{tr}(\Omega_i)$, Nocedal and Wright (2006):

$$\mathrm{tr}(\Omega_{i,k+1}) = \varphi\mathrm{tr}(\Omega_{i,k}) + (1-\varphi)\mathrm{tr}(\hat{\Omega}_{i,k+1}),$$

where $\mathrm{tr}(\hat{\Omega}_{i,k+1})$ is the ith covariance matrix trace computed at the $(k+1)$th iteration, and $\varphi \in (0, 1)$ is the fading parameter that controls the memory of the past covariance values. For SGN experiments in this section, we set parameter $\varphi = 0.9$, the step size $\gamma = 0.001$, and the regularization penalty $\delta = 100$.

The SGN's estimation error at time point k from the lth run is given by $\phi_{k,l} = \left\|FID - \mathbf{X}\hat{\mathbf{w}}_k^{(l)}\right\|$, where $\mathbf{X}$ is the matrix containing the observations and $\hat{\mathbf{w}}_k^{(l)}$ is the SGN weighted estimation (12) from the lth trial. The average estimation error $\bar{\phi}_k$ is defined as in 4.1 with $K = 20$. When we compare the final estimator $\hat{\mathbf{w}}_T$ at $T = 3000$ of SGN with the solution of the GLS, more than half of the SGN estimators fall into the 97% confidence interval of the GLS estimators. The average estimation error of SGN at the final step ($\bar{\phi}_T = 20.45$) is close to the estimation error of GLS (19.31).

5. Conclusions

The ever-increasing dimension of data and the size of datasets have introduced new challenges to centralized estimation. For example, limited bandwidth in the current networking infrastructure may not satisfy the demands for transmitting high-volume

datasets to a central location. Hence, it is of interest to study alternatives to centralized estimation.

In this paper we consider a distributed architecture for learning a linear model via generalized least squares by relying on a network of interconnected "local" learners. In the proposed distributed scheme, each computer (or local learner) is assigned a dataset and *asynchronously* implements stochastic gradient updates based upon a sample. To ensure robust estimation, a network regularization term that penalizes models with high *local* variability is used. Unlike other model averaging schemes based upon a synchronized step, the proposed scheme implements local model averaging continuously and asynchronously. We provide finite-time performance guarantees on consistency. We illustrate the application of the proposed method for sensor estimation in a Markov Random Field, and a real dataset from ecology.

Appendix

A.1. Continuous time representation of SGN

In this section, we will derive the formula of $\mathrm{d}\mathbf{w}_{i,t}$. By Central Limit Theorem, the approximation for noise terms hold for general distributions. (The proof is omitted here, see Hong et al. (2021) 5.3.2 for detailed discussion.) For simplicity of notations, we assume both noise terms have zero-mean Gaussian distribution: $\xi_k \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$ and $\varepsilon_{i,k} \sim \mathcal{N}_m(\mathbf{0}, \sigma_i^2\mathbf{I}_m)$ for all i.

For each node $i \in \mathcal{V}$, we rewrite the scheme (10) as:

$$\begin{aligned}\mathbf{w}_{i,t} =&\mathbf{w}_{i,0} - \gamma \sum_{k=1}^{N_{g,i}(t/\gamma)} g_{i,k} - \gamma\delta \sum_{k=1}^{N_{r,i}(t/\gamma)} \nabla\rho_{i,k} + \\ &\gamma\mathbf{X}_i^\intercal\Omega_i^{-1} \sum_{k=1}^{N_{g,i}(t/\gamma)} \varepsilon_{i,k} + \gamma\mathbf{X}_i^\intercal\Omega_i^{-1}\Lambda_i \sum_{k=1}^{N_{g,i}(t/\gamma)} \xi_k.\end{aligned} \tag{30}$$

Consider the second and the third term in (30). By renewal theorem $\frac{N_g(t/\gamma)}{t/\gamma} \to \mu_i$ as $\gamma \to 0$. When $\gamma \ll \mu_i$,

$$\gamma \sum_{k=1}^{N_{g,i}(t/\gamma)} g_{i,k} = \frac{\gamma N_{g,i}(t/\gamma)}{t} \sum_{k=1}^{N_{g,i}(t/\gamma)} g_{i,k}\frac{t}{N_{g,i}(t/\gamma)} \approx \mu_i \int_0^t g_{i,s}\mathrm{d}s. \tag{31}$$

Similarly, we have

$$\gamma \sum_{k=1}^{N_{r,i}(t/\gamma)} \nabla\rho_{i,k} \approx \hat{\beta} \int_0^t \nabla\rho_{i,s}\mathrm{d}s. \tag{32}$$

Now consider the individual noise. We assume all components of $\varepsilon_{i,k}$ are independent and it is enough to illustrate one-dimension approximation. Let $\varepsilon_{i,k}^{(q)}$ be the qth dimension of $\varepsilon_{i,k}$, and for all $q \in \mathcal{D} = \{1, \ldots, m\}$,

$$\begin{aligned}\mathbb{E}\Big[\sum_{k=1}^{N_{g,i}(t/\gamma)} \gamma\varepsilon_{i,k}^{(q)}\Big] &= 0,\\ \mathrm{Var}\Big[\sum_{k=1}^{N_{g,i}(t/\gamma)} \gamma\varepsilon_{i,l}^{(q)}\Big] &= \frac{\gamma N_{g,i}(t/\gamma)}{t}\gamma\sigma_i^2 t \approx \gamma\mu_i\sigma_i^2 t.\end{aligned}$$

By Donsker theorem, we approximate $\sum_{k=1}^{N_{g,i}(t/\gamma)} \varepsilon_{i,k}$ with a standard m-dimensional Brownian Motion $B_{i,t}$. Let $\tau_i = \sigma_i\sqrt{\gamma\mu_i}$, then the individual noise term in (30) can be approximated as

$$\gamma\mathbf{X}_i^\intercal\Omega_i^{-1} \sum_{k=1}^{N_{g,i}(t/\gamma)} \varepsilon_{i,l} \approx \tau_i\mathbf{X}_i^\intercal\Omega_i^{-1}B_{i,t}. \tag{33}$$

Similar to the proof of the individual noise approximation, let $\xi_k^{(q)}$ be the qth dimension of ξ_k, for $q \in \mathcal{D}$,

$$\mathbb{E}\Big[\sum_{k=1}^{N_{g,i}(t/\gamma)} \gamma\xi_l^{(q)}\Big] = 0, \quad \mathrm{Var}\Big[\sum_{k=1}^{N_{g,i}(t/\gamma)} \gamma\xi_l^{(q)}\Big] \approx \mu_i\gamma t.$$

Let $\varsigma_i = \sqrt{\gamma\mu_i}$, we approximate the common noise term in (30) with an m-dimensional Brownian Motion B_t:

$$\gamma\mathbf{X}_i^\intercal\Omega_i^{-1}\Lambda_i \sum_{k=1}^{N_{g,i}(t/\gamma)} \xi_k \approx \varsigma_i\mathbf{X}_i\Omega_i^{-1}\Lambda_i B_t. \tag{34}$$

Substituting (31), (32) and (34) to the corresponding terms in (30), $\mathbf{w}_{i,t}$ approximately satisfies the following stochastic Ito integral:

$$\begin{aligned}\mathbf{w}_{i,t} =&\mathbf{w}_{i,0} - \mu_i\int_0^t g_{i,s}\mathrm{d}s - \delta\hat{\beta}\int_0^t \nabla\rho_{i,s}\mathrm{d}s \\ &+ \tau_i\int_0^t \mathbf{X}_i^\intercal\Omega_i^{-1}\mathrm{d}B_{i,s} + \varsigma_i\int_0^t \mathbf{X}_i^\intercal\Omega_i^{-1}\Lambda_i\mathrm{d}B_s.\end{aligned}$$

Taking the derivative of the above equation, we get (11).

A.2. Weak convergence

Consider the stacked vector $\mathbf{w}_t = [\mathbf{w}_{1,t}^\intercal, \ldots, \mathbf{w}_{N,t}^\intercal]^\intercal$ and a potential function:

$$H(\mathbf{w}) \triangleq \sum_{i=1}^N \alpha_i[\mu_i f_i(\mathbf{w}_i) + \frac{\delta\hat{\beta}}{2}\rho_i(\mathbf{w})].$$

Note that $\alpha_i\nabla_{\mathbf{w}_i}\rho_i(\mathbf{w}) = \alpha_j\nabla_{\mathbf{w}_i}\rho_j(\mathbf{w})$. Hence,

$$\nabla_{\mathbf{w}_i}H(\mathbf{w}) = \alpha_i[\mu_i\nabla_{\mathbf{w}_i}f_i(\mathbf{w}_i) + \delta\hat{\beta}\nabla_{\mathbf{w}_i}\rho_i(\mathbf{w})],$$

and the system of stochastic differential equations (11) can be written as stochastic gradient flow:

$$\mathrm{d}\mathbf{w}_t = -A^{-1}\nabla H(\mathbf{w}_t)\mathrm{d}t + \sqrt{\gamma}\mathrm{d}\tilde{B}_t,$$

where $A \succeq 0$ is the diagonal matrix with diagonal entries $\alpha_i > 0$ and $\tilde{B}_t$ is defined as:

$$\tilde{B}_{i,t} \triangleq \sqrt{\mu_i}\mathbf{X}_i^\top\Omega_i^{-1}(\sigma_i B_{i,t} + \Lambda_i B_t).$$

By theory of diffusions (see e.g. Risken (1996)), the process $\{\mathbf{w}_t : t \geq 0\}$ converges weakly (in distribution) and the limiting density is Gibbs, i.e.,

$$\lim_{t\to\infty} \Pr(\mathbf{w}_t \in \mathcal{C}) \propto \int_{\mathcal{C}} \exp(-\frac{H(\mathbf{w})}{\gamma})d\mathbf{w},$$

for $\mathcal{C} \subset \mathbb{R}^{p\times N}$. By continuous mapping theorem, the ensemble average $\{\hat{\mathbf{w}}_t : t > 0\}$ also converges in distribution. This in turn implies that both $\{\bar{V}_t : t > 0\}$ and $\{U_t : t > 0\}$ converge in distribution.

A.3. Proof of Theorem 3

The following lemma provides the differential form of the regularity measure.

Lemma 11. *The regularity measure $\bar{V}_t$ satisfies*

$$d\bar{V}_t \leq -\sum_{i=1}^N \alpha_i\mu_i g_{i,t}^\intercal\boldsymbol{e}_{i,t}dt - \delta\hat{\beta}\sum_{i=1}^N \alpha_i\nabla\rho_{i,t}^\intercal\boldsymbol{e}_{i,t}dt + K_1 d\tilde{B}_t + C_1 dt, \tag{35}$$

where $K_1 d\tilde{B}_t$ is the summation of Ito terms,

$$K_1 d\tilde{B}_t = \sum_{i=1}^{N} \alpha_i K_{1,i} d\tilde{B}_t,$$

with $K_{1,i} d\tilde{B}_t$ for $i \in \mathcal{V}$ defined as,

$$K_{1,i} d\tilde{B}_t = \tau_i \boldsymbol{X}_i^\intercal \Omega_i^{-1} dB_{i,t}^\intercal \mathbf{e}_{i,t} + \varsigma_i \boldsymbol{X}_i^\intercal \Omega_i^{-1} \Lambda_i dB_t^\intercal \mathbf{e}_{i,t}.$$

Proof. By definition of $\hat{\mathbf{w}}_t$ and $\mathbf{e}_{i,t}$, it follows that

$$d\mathbf{e}_{i,t} = d\mathbf{w}_{i,t} - d\hat{\mathbf{w}}_t. \tag{36}$$

Note that for d-by-m matrices $C = [c_1, \ldots, c_d]$ and $Q = [q_1, \ldots, q_d]$, $C dB_t \cdot C dB_t = \|C\|_F^2 \, dt$, and $C dB_t \cdot Q dB_t = \mathbf{1}^\intercal (C \circ Q)\mathbf{1}$, where "$\mathbf{1}$" is a vector of all ones and "$\circ$" denotes the Hadamard product. Then by (36), the inner product of two $d\mathbf{e}_{i,t}$ is $d\mathbf{e}_{i,t} \cdot d\mathbf{e}_{i,t} = C_{1,i} dt$, and hence $C_{1,i}$ is positive. Applying Ito's lemma to $dV_{i,t}$, we obtain

$$\begin{aligned} dV_{i,t} = \mathbf{e}_{i,t} \cdot d\mathbf{e}_{i,t} + \frac{1}{2} d\mathbf{e}_{i,t} \cdot d\mathbf{e}_{i,t} \leq \Big(-\mu_i g_{i,t} + \mu' \sum_{k=1}^{N} \alpha_k g_{k,t}\Big)^\intercal \mathbf{e}_{i,t} dt \\ + \delta\beta \Big(-\nabla \rho_{i,t} + \sum_{k=1}^{N} \alpha_k \nabla \rho_{k,t}\Big)^\intercal \mathbf{e}_{i,t} dt + \frac{1}{2} C_{i,1} dt + K_{1,i} d\tilde{B}_t - K_1 d\tilde{B}_t, \end{aligned} \tag{37}$$

where $\mu' = \max_i \mu_i$. To obtain the differential form of the regularity measure, we take the weighted average of (37). Because of (19), the terms with double summation (weighting) vanish, and (35) follows. □

Now we are ready to prove Theorem 3. We first obtain an upper bound of $d\bar{V}_t$ and then integrate and take the expectation of the obtained bound to get the desired result. Consider the first term of (35), let $h_t = \min_{i \in \mathcal{V}} g_{i,t}(\hat{\mathbf{w}}_t)$. By (19), we can add a zero-valued term $h_t^\intercal \sum_{i=1}^{N} \alpha_i \mathbf{e}_{i,t}$ to the equation, and by the strong convexity of g_i (16),

$$\begin{aligned} -\sum_{i=1}^{N} \alpha_i \mu_i g_{i,t}^\intercal \mathbf{e}_{i,t} &\leq -\mu \sum_{i=1}^{N} \alpha_i (g_{i,t} - h_t)^\intercal \mathbf{e}_{i,t} \\ &\leq -\mu \sum_{i=1}^{N} \alpha_i \kappa_i \left\| \mathbf{e}_{i,t} \right\|^2 \leq -2\kappa\mu \bar{V}_t, \end{aligned} \tag{38}$$

where $\mu = \min_i \mu_i$ and $\kappa = \min_i \kappa_i$. Now we consider the second term of (35). Define the vector $\mathbf{e}_t = [\mathbf{e}_{1,t}^T, \ldots, \mathbf{e}_{N,t}^T]^T$ and the matrix $\hat{L} = \hat{L} \otimes I_m$, where $\otimes$ is the Kronecker product. Then

$$-\sum_{i=1}^{N} \sum_{j \neq i} \hat{\alpha}_i \hat{\alpha}_j a_{i,j} (\mathbf{w}_{i,t} - \mathbf{w}_{j,t})^\intercal \mathbf{e}_{i,t} = -\mathbf{e}_t^\intercal \hat{L} \mathbf{e}_t, \tag{39}$$

where $\hat{l}_{i,j}$ is the (i,j)th entry of $\hat{L}$. The second smallest eigenvalue $\hat{\lambda}_2 > 0$ satisfies (see Godsil and Royle (2001)), $\min_{x \neq 0,\, 1^\intercal x = 0} \frac{(x^\intercal \hat{L} x)}{\|x\|^2} = \hat{\lambda}_2$. Thus, we have $-\mathbf{e}_t^T \hat{L} \mathbf{e}_t \leq -\hat{\lambda}_2 \sum_{i=1}^{N} \left\| \mathbf{e}_{i,t} \right\|^2$. We assume that $\hat{\alpha}_i = c\alpha_i$ and $\beta = \hat{\beta}/c$, it follows that,

$$-\delta\hat{\beta} \sum_{i=1}^{N} \alpha_i \nabla \rho_{i,t}^\intercal \mathbf{e}_{i,t} \leq -\frac{\delta\hat{\beta}}{c} \hat{\lambda}_2 \sum_{i=1}^{N} \left\| \mathbf{e}_{i,t} \right\|^2 = -2\delta\beta\hat{\lambda}_2 \bar{V}_t.$$

An upper bound for $d\bar{V}_t$ follows

$$d\bar{V}_t \leq -2(\mu\kappa + \delta\beta\hat{\lambda}_2)\bar{V}_t dt + K_1 d\tilde{B}_t + C_1 dt. \tag{40}$$

Now consider the derivative of $e^{2(\mu\kappa + \delta\beta\hat{\lambda}_2)t} \bar{V}_t$,

$$d(e^{2(\mu\kappa + \delta\beta\hat{\lambda}_2)t} \bar{V}_t) \leq e^{2(\mu\kappa + \delta\beta\hat{\lambda}_2)t} C_1 dt + e^{2(\mu\kappa + \delta\beta\hat{\lambda}_2)t} K_1 d\tilde{B}_t, \tag{41}$$

Integrating both sides of the inequality in (41),

$$\begin{aligned} \bar{V}_t \leq & e^{-2(\mu\kappa + \delta\beta\hat{\lambda}_2)t} \bar{V}_0 + \frac{C_1}{2(\mu\kappa + \delta\beta\hat{\lambda}_2)} (1 - e^{-2(\mu\kappa + \delta\beta\hat{\lambda}_2)t}) \\ & + e^{-2(\mu\kappa + \delta\beta\hat{\lambda}_2)t} \int_0^t e^{2(\mu\kappa + \delta\beta\hat{\lambda}_2)s} K_1 d\tilde{B}_s. \end{aligned} \tag{42}$$

Since the stochastic integral is a martingale, we obtain the desired upper bound by taking the expectation on both sides of (42). In the long run, as $t \to \infty$, the exponential terms will vanish, and the upper bound of the regularity measure follows.

A.4. Proof of Theorem 4

The proof follows a similar outline as Theorem 3. We start with Ito's Lemma to get the stochastic dynamics form of dU_t and then introduce an auxiliary variable W_t that depends on both U_t and $\bar{V}_t$ to bound $E[U_t]$.

We apply Ito's lemma to dU_t, and use the identity in (19) and $\hat{\mathbf{w}}_t - \mathbf{w}^* = (\mathbf{w}_{k,t} - \mathbf{w}^*) - \mathbf{e}_{k,t}$ and the differential form of $\hat{\mathbf{w}}_t$ to get the following form,

$$dU_t = -\sum_{i=1}^{N} \alpha_i \mu_i g_{k,t}^\intercal (\mathbf{w}_{k,t} - \mathbf{w}^*) dt + \sum_{i=1}^{N} \alpha_i \mu_i g_{k,t}^\intercal e_{kt} dt + K_2 d\tilde{B}_t + C_2 dt, \tag{43}$$

where the summation term C_2 is defined in (21) and $K_2 d\tilde{B}_t$ is given as,

$$\begin{aligned} K_2 d\tilde{B}_t = & \sum_{k=1}^{N} \alpha_k \tau_k (\hat{\mathbf{w}}_t - \mathbf{w}^*)^\intercal \mathbf{X}_k^\intercal \Omega_k^{-1} dB_{k,t} \\ & + \sum_{k=1}^{N} \alpha_k \varsigma_k (\hat{\mathbf{w}}_t - \mathbf{w}^*)^\intercal \mathbf{X}_k^\intercal \Omega_k^{-1} \Lambda_k dB_t. \end{aligned}$$

We have the upper bound on the first term of (43),

$$\begin{aligned} -\sum_{i=1}^{N} \alpha_i \mu_i g_{k,t}^\intercal (\mathbf{w}_{k,t} - \mathbf{w}^*) &\leq -\mu \sum_{i=1}^{N} \alpha_i (g_{k,t} - g(\mathbf{w}^*))^\intercal (\mathbf{w}_{k,t} - \mathbf{w}^*) \\ &\leq -2\kappa\mu(\bar{V}_t + U_t). \end{aligned} \tag{44}$$

The first equality is obtained by subtracting the zero-valued term $g(\mathbf{w}^*)$ from $g_{k,t}$. The second inequality is by strong convexity of the gradients ($g_{k,t}$), and the last equality follows from (17).

Now consider the second term of (43). Let $q_t = \max_{k \in \mathcal{V}} g_k(\hat{\mathbf{w}}_t)$, then it follows that

$$\begin{aligned} \sum_{i=1}^{N} \alpha_i \mu_i g_{k,t}^\intercal e_{k,t} &\leq \mu' \sum_{i=1}^{N} \alpha_i (g_{k,t} - q_t)^\intercal e_{k,t} \\ &\leq 2\mu' \sum_{k=1}^{N} \frac{\alpha_i}{2} (g_{k,t} - g_k(\hat{\mathbf{w}}_t))^\intercal e_{k,t} \leq 2\mu' \eta \bar{V}_t. \end{aligned} \tag{45}$$

By (43)–(45), we can obtain an upper bound of dU_t,

$$dU_t \leq -2\mu\kappa U_t dt + 2(\eta\mu' - \kappa\mu)\bar{V}_t dt + C_2 dt + K_2 d\tilde{B}_t. \tag{46}$$

We construct an auxiliary variable W_t and then follow similar steps as in the proof of Theorem 3: integrate the upper bound of dW_t and then take expectation. Now, define $W_t = U_t + \frac{\eta\mu' - \kappa\mu}{\delta\beta\hat{\lambda}_2} \bar{V}_t$.

The differential of W_t can be obtained plugging in the bounds for $d\bar{V}_t$ in (40) and for dU_t in (46) into dW_t:

$$dW_t \leq -2\mu\kappa W_t dt + \Big(\frac{\eta\mu' - \kappa\mu}{\delta\beta\hat{\lambda}_2} C_1 + C_2\Big) dt + \Big(\frac{\eta\mu' - \kappa\mu}{\delta\beta\hat{\lambda}_2} K_1 + K_2\Big) d\tilde{B}_t. \tag{47}$$

Consider the derivative of $e^{2\mu\kappa t}W_t$ and plug in the upper bound of $\mathrm{d}W_t$ in (47) to obtain,

$$\mathrm{d}(e^{2\mu\kappa t}W_t) = e^{2\mu\kappa t}\mathrm{d}W_t + 2\mu\kappa e^{2\mu\kappa t}\mathrm{d}W_t \leq e^{2\mu\kappa t}\Big(\frac{\eta\mu' - \kappa\mu}{\delta\beta\hat{\lambda}_2}C_1 + C_2\Big)\mathrm{d}t + e^{2\mu\kappa t}\Big(\frac{\eta\mu' - \kappa\mu}{\delta\beta\hat{\lambda}_2}K_1 + K_2\Big)\mathrm{d}\tilde{B}_t \tag{48}$$

We assume that all nodes have the same initial estimate, i.e., $\mathbf{w}_{i,0} = \mathbf{w}_{j,0}$ for all i, j. Then, $\bar{V}_0 = 0$, which means $W_0 = U_0$. Note that the stochastic integration of the Ito term is martingale and hence the expectation is zero. Integrating both sides of (48) and taking expectation,

$$\mathbb{E}[W_t] \leq e^{-2\mu\kappa t}U_0 + \frac{1}{2\mu\kappa}\Big(\frac{\eta\mu' - \kappa\mu}{\delta\beta\hat{\lambda}_2}C_1 + C_2\Big)\Big(1 - e^{-2\mu\kappa t}\Big). \tag{49}$$

Since $\eta\mu' - \kappa\mu > 0$, we have $\mathbb{E}[U_t] \leq \mathbb{E}[W_t]$. Hence the right-hand side of (49) is also the upper bound of $\mathbb{E}[U_t]$. In the long run, as $t \to \infty$, the exponential terms vanish and the upper bound of the consistency measure follows.

A.5. Proof of Theorem 8

In this section, we first derive the formula of $\mathrm{d}\mathbf{w}_t$ in the similar way as in A.1, and then provide the convergence property of the federated learning.

We rewrite the scheme (23) as:

$$\begin{aligned}\mathbf{w}_t = \mathbf{w}_0 &- \gamma\sum_{i=1}^{N}\sum_{k=1}^{N_{g,i}(t/\gamma)}\mathbf{X}_i^{\intercal}\Omega_i^{-1}\mathbf{X}_i(\mathbf{w}_{i,l} - \mathbf{w}^*) \\ &+ \gamma\sum_{i=1}^{N}\sum_{k=1}^{N_{g,i}(t/\gamma)}\mathbf{X}_i^{\intercal}\Omega_i^{-1}\varepsilon_{i,l} + \gamma\sum_{i=1}^{N}\sum_{k=1}^{N_{g,i}(t/\gamma)}\mathbf{X}_i^{\intercal}\Omega_i^{-1}\Lambda_i\xi_l \\ &+ \gamma\sum_{i=1}^{N}\sum_{k=1}^{N_{g,i}(t/\gamma)}\epsilon_{i,l}.\end{aligned} \tag{50}$$

By (31), and (34), we can obtain the continuous approximation of $\mathrm{d}\mathbf{w}_t$. The rest of the proof follows a similar outline as Theorem 3. We apply Ito's lemma to $\mathrm{d}F_t$ to get the following form,

$$\mathrm{d}F_t = -\sum_{i=1}^{N}\mu_i g_{i,t}^{\intercal}(\mathbf{w}_t - \mathbf{w}^*)\mathrm{d}t + K_3\mathrm{d}\tilde{B}_{f,t} + C_3\mathrm{d}t,$$

where $K_3\mathrm{d}\tilde{B}_{f,t}$ and C_3 are defined in (24) and (25). Note that $(\nabla f(x_1) - \nabla f(x_2))^T(x_1 - x_2) \geq \frac{\kappa}{2}\|x_1 - x_2\|^2$, then

$$-\sum_{i=1}^{N}\mu_i g_{i,t}^{\intercal}(\mathbf{w}_t - \mathbf{w}^*)\mathrm{d}t \leq -\kappa\mu\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\mathrm{d}t.$$

It follows that

$$\mathrm{d}F_t \leq -2\mu\kappa F_t\mathrm{d}t + K_3\mathrm{d}\tilde{B}_{f,t} + C_3\mathrm{d}t.$$

Now consider $\mathrm{d}(e^{2\kappa\mu t}F_t)$,

$$\mathrm{d}(e^{2\kappa\mu t}F_t) = e^{2\kappa\mu t}\mathrm{d}F_t + 2\kappa\mu e^{2\kappa\mu t}F_t\delta\mathrm{d}t \leq e^{2\kappa\mu t}C_3\mathrm{d}t + e^{2\kappa\mu t}K_3\mathrm{d}\tilde{B}_t.$$

Integrating the left and right hand sides of the inequality and taking the expectation, by observing that the last term is a martingale we get (27). Now, we consider the lower bound of F_t. By Lipschitz continuity,

$$-\sum_{i=1}^{N}\mu_i g_{i,t}^{\intercal}(\mathbf{w}_t - \mathbf{w}^*)\mathrm{d}t \geq -\eta\mu'\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\mathrm{d}t.$$

It follows that

$$\mathrm{d}F_t \geq -2\eta\mu' F_t\mathrm{d}t + K_3\mathrm{d}\tilde{B}_{f,t} + C_3\mathrm{d}t.$$

Now consider $\mathrm{d}(e^{2\eta\mu' t}F_t)$,

$$\begin{aligned}\mathrm{d}(e^{2\eta\mu' t}F_t) &= e^{2\eta\mu' t}\mathrm{d}F_t + 2\eta\mu' e^{2\eta\mu' t}F_t\mathrm{d}t \\ &\geq e^{2\eta\mu' t}C_3\mathrm{d}t + e^{2\eta\mu' t}K_3\mathrm{d}\tilde{B}_t.\end{aligned}$$

Integrating both sides of the inequality and taking expectation, it yields (26).

References

Bates, J. M., & Granger, C. M. W. (1969). The combination of forecasts. *Operations Research Quaterly*, *20*, 451–468.

Chen, J., Richard, C., & Sayed, A. (2014). Multitask diffusion adaptation over networks. *IEEE Transactions on Signal Processing*, *62*, 4129–4144.

Donsker, M. (1951). An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*, *6*.

Eksin, C., & Ribeiro, A. (2012). Distributed network optimization with heuristic rational agents. *IEEE Transactions on Signal Processing*, *60*(10), 5396–5411.

Garcia, A., Wang, L., Huang, J., & Hong, L. (2020). Distributed networked real-time learning. *IEEE Transactions on Control of Network Systems*.

Godsil, C., & Royle, G. (2001). *Algebraic graph theory*. New York: Springer.

Granger, C. M. W., & Ramanathan, R. (1984). Improved methods of combining forecast accuracy. *Journal of Forecasting*, *19*, 197–204.

Hansen, B. (2007). Least squares model averaging. *Econometrica*, *75*(4), 1175–1189.

He, L., Bian, A., & Jaggi, M. (2018). Cola: Decentralized linear learning. In *Advances in Neural Information Processing Systems* (pp. 4536–4546).

Hong, L., Garcia, A., & Eksin, C. (2020). Distributed networked learning with correlated data. In *2020 59th IEEE Conference on Decision and Control (CDC)* (pp. 5923–5928). IEEE.

Hong, L., Garcia, A., & Eksin, C. (2021). Distributed networked learning with correlated data. https://arxiv.org/abs/1910.12783.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, *37*(3), 50–60.

Lian, X., Huang, Y., Li, Y., & Liu, J. (2015). Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems, Vol. 28 (NIPS)*.

Liu, J., Mou, S., & Morse, A. S. (2015). Asynchronous distributed algorithms for solving linear algebraic equations. *IEEE Transactions on Automatic Control*, *63*(2), 372–385.

Liu, Q., Okui, R., & Yoshimura, A. (2016). Generalized least squares model averaging. *Econometric Reviews*, *35*(8), 1692–1752.

Marelli, D. E., & Fu, M. (2015). Distributed weighted least-squares estimation with fast convergence for large-scale systems. *Automatica*, *51*, 27–39.

Mendes-Moreira, J., Soares, C., Jorge, A., & Freire de Sousa, J. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys*, *45*(1), 10–40.

Morelli, F., Benedetti, Y., Diaz, M., Grim, T., Ibanez-Alamo, J. D., Jokimaki, J., et al. (2019). Contagious fear: Escape behavior increases with flock size in European gregarious birds. *Ecology and Evolution*, *9*(10), 6096–6104.

Mou, S., Liu, J., & Morse, A. S. (2015). A distributed algorithm for solving a linear algebraic equation. *IEEE Transactions on Automatic Control*, *60*(11), 2863–2878.

Nassif, R., Richard, C., Ferrari, A., & Sayed, A. (2016). Multitask diffusion adaptation over asynchronous networks. *IEEE Transactions on Signal Processing*, *64*, 2835–2850.

Nassif, R., Vlaski, S., Richard, C., & Sayed, A. H. (2020a). Learning over multitask graphs—Part I: Stability analysis. *IEEE Open Journal of Signal Processing*, *1*, 28–45.

Nassif, R., Vlaski, S., Richard, C., & Sayed, A. H. (2020b). Learning over multitask graphs—Part II: Performance analysis. *IEEE Open Journal of Signal Processing*, *1*, 46–63.

Nedic, A., & Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, *54*(1), 48.

Nocedal, J., & Wright, S. (2006). *Numerical Optimization*. Springer Science and Business Media.

Predd, J., Kulkarni, S., & Poor, H. V. (2009). A collaborative training algorithm for distributed learning. *IEEE Transactions on Information Theory*, *55*, 1856–1869.

Risken, H. (1996). *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer.

Shi, W., Ling, Q., Wu, G., & Yin, W. (2015). Extra: An exact first- order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, *25*, 944–966.

Wang, X., Zhou, J., Mou, S., & Corless, M. (2019). A distributed algorithm for least squares solutions. *IEEE Transactions on Automatic Control*, *64*(10), 4217–4222.
Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1–19.
Zhang, Y., Duchi, J., & Wainwright, M. (2013). Communication-efficient algorithms for statistical optimization. *Journal of Machine Intelligence Research*, *14*, 3321–3363.
Zhang, Y., Duchi, J., & Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Intelligence Research*, *16*, 3299–3340.

Lingzhou Hong received BEcon. in Statistics from Central University of Finance and Economics, Beijing, China in 2013, and M.S.E and M.A. degrees in Applied Mathematics and Statistics from Johns Hopkins University in 2015 and 2017. She is currently pursuing the Ph.D degree in Industrial & Systems Engineering at Texas A&M University. Her research interests include machine learning, distributed optimization, and statistical learning.

Alfredo Garcia received B.Sc. in Electrical Engineering from the Universidad de los Andes, Colombia in 1991, D.E.A in Automatique et Informatique Industrielle from the Université Paul Sabatier in Toulouse, France in 1992, and Ph.D. in Industrial and Operations Engineering from the University of Michigan in 1997. From 1998 to 2000, he served as Commissioner in the Colombian Energy Regulatory Commission, and from 2001 to 2017 he was a member of the faculty at the University of Virginia and the University of Florida. He is currently a professor with Industrial & Systems Engineering Department, Texas A&M University. His research interests include game theory and dynamic optimization, with applications in electricity and communication networks.

Ceyhun Eksin received the B.S. degree in control engineering from Istanbul Technical University, in 2005, an M.S. degree in industrial engineering from Boğaziçi University, Istanbul, Turkey in 2008, an M.A. degree in statistics from Wharton School in 2015, and the Ph.D. degree in Electrical and Systems Engineering from the University of Pennsylvania in 2015. He was a postdoctoral researcher in Georgia Institute of Technology from 2015 to 2017. He is currently an assistant professor with the Industrial and Systems Engineering Department, Texas A&M University. His research interests focus on modeling and design of networked multi-agent systems using game theory, control theory, and statistical signal processing.