Distributed Networked Real-Time Learning

Alfredo Garcia, *Senior Member, IEEE*, Luochao Wang, Jeff Huang, and Lingzhou Hong

***Abstract*—Many machine learning algorithms have been developed under the assumption that datasets are already available in batch form. Yet, in many application domains, data are only available sequentially overtime via compute nodes in different geographic locations. In this article, we consider the problem of learning a model when streaming data cannot be transferred to a single location in a timely fashion. In such cases, a distributed architecture for learning which relies on a network of interconnected "local" nodes is required. We propose a distributed scheme in which every local node implements stochastic gradient updates based upon a local data stream. To ensure robust estimation, a network regularization penalty is used to maintain a measure of cohesion in the ensemble of models. We show that the ensemble average approximates a stationary point and characterizes the degree to which individual models differ from the ensemble average. We compare the results with federated learning to conclude that the proposed approach is more robust to heterogeneity in data streams (data rates and estimation quality). We illustrate the results with an application to image classification with a deep learning model based upon convolutional neural networks.**

***Index Terms*—Asynchronous computing, distributed computing, networks, nonconvex optimization, real-time machine learning.**

I. Introduction

STREAMING datasets are pervasive in certain application domains often involving a network of compute nodes located in different geographic locations. However, most machine learning algorithms have been developed under the assumption that datasets are already available in batch form. When the data are obtained through a network of heterogeneous compute nodes, assembling a diverse batch of data points in a central processing location to update a model may imply significant latency. Recently, an architecture referred to as *federated* learning (FL, see, e.g., [1], [2]) with a central server in proximity to local nodes has been proposed. In FL, each node implements updates to a machine learning model, which is kept in the central server. This allows collaborative learning while keeping all the training data on nodes rather than in the cloud. In general, schemes that avoid the need to rely on the cloud for data storage and/or computation are referred to as "edge computing."

With high data payloads, such architecture for real-time learning is subject to an *accuracy* versus *speed* tradeoff due to asymmetries in data quality versus data rates, as we explain in what follows.

Consider nodes $i \in \{1, \dots, N\}$ generating data points $(x_{i,n}, y_{i,n})$, $n \in \mathbb{N}^+$ at different rates $\mu_i > 0$, which are used for the instantaneous computation of model updates θ_k, $k \in \mathbb{N}^+$ (striving to minimize loss ℓ). This setting could correspond, for example, with supervised deep learning in real time wherein gradient estimates (with noise variance $\sigma_i^2 > 0$) are computed via backpropagation in a relatively fast fashion. Without complete information on $\sigma_i^2 > 0$, updating the model parameters based upon every incoming data point yields high speed but possibly at the expense of low accuracy. For example, if the nodes producing noisier estimates are also *faster* at producing data, it is highly unlikely that an accurate model will be identified at all.

To illustrate this scenario, in Fig. 1, we depict the performance of FL for *deep* convolutional neural networks (CNNs) with the Modified National Institute of Standards and Technology (MNIST) database. In these simulations, each one of $N = 5$ nodes sends data according to independent Poisson processes with $\mu_0 = 8$ and $\mu_i = 1$, $i \in \{1, \dots, 4\}$. The fastest node computes gradient estimates based upon a *single* image, whereas the slower nodes compute gradient estimates based upon a batch of 64 images. This tradeoff between *speed* and *precision* is mitigated in a distributed approach to real-time learning subject to a network regularization (NR) penalty. In such an approach, each one of the $N > 1$ local nodes independently produces parameter updates based upon a single (locally obtained) data point, which speeds up computation. Evidently, with increased noise, such a scheme may fail to enable the identification of a reasonably accurate model. However, by adding a NR penalty (which is computed locally), a form of coordination between multiple local nodes is induced so that the ensemble average solution is robust to noise.[1] Specifically, we show that the ensemble average solution approximates a stationary point and that the approximation quality is $\mathcal{O}(\frac{\sum_{i=1}^N \sigma_i^2}{N^2})$, which compares quite favorably with FL, which is highly sensitive to fast and inaccurate data streams. We illustrate the results with an application to deep learning with CNNs.

Manuscript received February 28, 2020; revised March 1, 2020, May 25, 2020, and September 6, 2020; accepted September 12, 2020. Date of publication October 9, 2020; date of current version February 26, 2021. This work was supported in part by the National Science Foundation under Award ECCS-1933878 and in part by the Air Force Office of Scientific Research under Grant 15RT0767. Recommended by Associate Editor A. Olshevsky. *(Corresponding author: Alfredo Garcia.)*

Alfredo Garcia and Lingzhou Hong are with the Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: alfredo.garcia@tamu.edu; hlz@tamu.edu).

Luochao Wang and Jeff Huang are with the Department of Computer Science & Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: wangluochao@tamu.edu; jeff@cse.tamu.edu).

Digital Object Identifier 10.1109/TCNS.2020.3029992

[1]Similar NR methods have been used in multitask learning to account for inherent network structure in datasets (see, e.g., [3]–[7]). See Section II.D.

The structure of this article is as follows. In Section II, we introduce the distributed scheme that combines stochastic gradient descent with NR. In Section III, we analyze the scheme and show that it converges (in a certain sense) to a stationary point, we also compare its performance with FL. Finally, in Section IV, we report the results from a testbed on deep learning application to image processing, and in Section V, we offer conclusion.

II. Network Regularized Approach to Real-Time Learning

A. Setup

We consider a setting in which data are made available sequentially overtime via nodes $i \in \{1, \dots, N\}$ in different geographic locations. We denote the ith stream by $\{(\boldsymbol{x}_{i,n}, y_{i,n}) : n \in \mathbb{N}^+\}$ and assume these data points are independent samples from a joint distribution $\mathcal{P}_i$.

We also assume that the data streams are independent but heterogeneous, i.e., $\mathcal{P}_i \neq \mathcal{P}_j, i \neq j$. Each node strives to find a parameter specification $\theta \in \Theta \subset \mathbb{R}^p$ that minimizes the performance criteria $\mathbb{E}_{\mathcal{P}_i}[\mathcal{L}(\boldsymbol{x}_i, y_i; \theta)]$, where the loss function $\mathcal{L}(\cdot) \geq 0$ is continuously differentiable with respect to θ. Though data are distributed and heterogeneous, we consider a setting in which nodes *agree* on a common learning task. This is formalized in the first standing assumption. Let $g_i(\theta) \triangleq \nabla_\theta \mathcal{L}(\boldsymbol{x}_i, y_i; \theta)$ denote the gradient evaluated at $(\boldsymbol{x}_i, y_i) \sim \mathcal{P}_i$, and assume $g_i(\theta)$ is uniformly integrable.

Assumption 0: For all $\theta \in \Theta$, and $i \in \{1, \dots, N\}$

$$\mathbb{E}_{\mathcal{P}_i}[g_i(\theta)] = \mathbb{E}_{\mathcal{P}_j}[g_j(\theta)].$$

Let $\ell(\theta)$ denote the (ensemble) average expected loss

$$\ell(\theta) \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\mathcal{P}_i}[\mathcal{L}(\boldsymbol{x}_i, y_i; \theta)].$$

By uniform integrability, $\nabla_\theta \mathbb{E}_{\mathcal{P}_i} \mathcal{L}(\boldsymbol{x}_i, y_i; \theta) = \mathbb{E}_{\mathcal{P}_i} g_i(\theta)$. Assumption 0 thus implies that $\mathbb{E}_{\mathcal{P}_i}[g_i(\theta)] = \nabla \ell(\theta)$ for all i and θ. Let $\varepsilon_i(\theta) \triangleq g_i(\theta) - \nabla \ell(\theta)$, then it holds that $\mathbb{E}[\varepsilon_i(\theta)] = 0$. We further assume the following.

Assumption 1: For all $\theta_i \in \Theta$, the random variables $\{\varepsilon_i(\theta_i) : i \in \{1, \dots, N\}\}$ are independent and

$$\mathbb{E}[\|\varepsilon_i(\theta_i)\|^2] \leq \sigma_i^2.$$

Define $\sigma^2 = \sum_{i=1}^N \sigma_i^2$. By independence of data streams

$$\mathbb{E}[\varepsilon_i(\theta)^\intercal \varepsilon_j(\theta)] = \mathbb{E}[\varepsilon_i(\theta)]^\intercal \mathbb{E}[\varepsilon_j(\theta)] = 0$$

for all $\theta \in \Theta, j \in \{1, \dots, N\}/\{i\}$. Streams generate data over time according to independent Poisson processes $D_i(t)$ with rate $\mu_i > 0$ and $D_i(0) = 1$. We assume that the time required to compute gradient estimates and/or exchange parameters locally among neighbors or with the central server are negligible compared to the time in between model updates. In what follows, we make use of a virtual clock that produces ticks according to an aggregate counting process $D(t) = \sum_{i=1}^N D_i(t)$ with rate $\mu = \sum_{i=1}^N \mu_i$. Let $k \in \mathbb{N}^+$ denote the index set of ticks associated with the aggregate process. Since we assume the parameter is updated once a data point arrives, the kth iteration is completed at the kth tick. Index k denotes the kth step in the schemes described below.

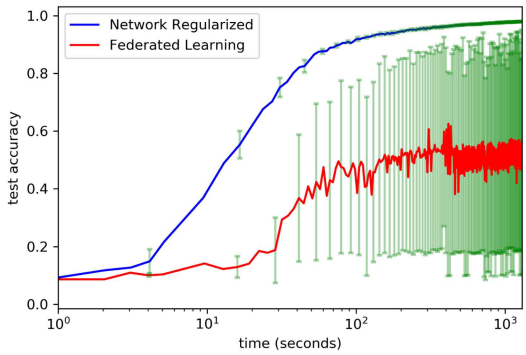


Fig. 1. Performance comparison for deep learning on MNIST with learning rate $\gamma = 0.01$. The 95% percentile is depicted with green lines.

B. Federated Real-Time Learning

In FL, gradient estimates are communicated to a central server where a model is updated as follows:

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^{N} \mathbf{1}_{i,k} g_i(\theta_k) \tag{1}$$

where γ is the learning rate, $\mathbf{1}_{i,k}$ is an indicator of whether node i performs an update: $\mathbf{1}_{i,k} = 1$ if the next gradient estimate comes from the ith stream and $\mathbf{1}_{i,k} = 0$ otherwise.

The algorithmic scheme described in (1) was first analyzed in [8] for data in batch form and has been used in the recent literature on asynchronous parallel optimization algorithms (see, for example, [9], [10], and [11]). As Fig. 1 suggests, with heterogeneous data streams, the scheme in (1) tradeoff speed in producing parameter updates at the expense of heterogeneous noise in gradient estimates. In what follows, we introduce a distributed approach that relies on an NR penalty to ensure the ensemble average approximates a stationary point (i.e., a choice of parameters with null gradient). We will show that in such a networked approach, the tradeoff between precision and speed is mitigated.

C. Distributed Approach With NR

In the NR scheme, we consider a network of local compute nodes, which we model as a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, \dots, N\}$ stands for the set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of links connecting nodes. Let $A = [\alpha_{ij}] \in \mathbb{R}^{N \times N}$ be the adjacency matrix of $\mathcal{G}$, where $\alpha_{ij} \in \{0, 1\}$ indicates whether node i communicates with node j: $\alpha_{ij} = 1$ if two nodes can exchange local information and $\alpha_{ij} = 0$ otherwise.

In this scheme, each local node i performs model updates according to a linear combination of local gradient estimate and the gradient of a consensus potential

$$\mathcal{F}(\boldsymbol{\theta}) = \frac{1}{4} \sum_i \sum_{j \neq i} \alpha_{ij} \|\theta_i - \theta_j\|^2$$

where $\boldsymbol{\theta}_t^\intercal = (\theta_{1,t}^\intercal, \ldots, \theta_{N,t}^\intercal) \in \mathbb{R}^{p\times N}$. The consensus potential is a measure of similarity across local models.[2] The update performed by node i is of the form

$$\theta_{i,k+1} = \theta_{i,k} - \gamma \mathbf{1}_{i,k}[g_i(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}] \tag{2}$$

where $a > 0$ is a regularization parameter, and

$$\nabla\mathcal{F}_{i,k} \triangleq \nabla_{\theta_i}\mathcal{F}(\boldsymbol{\theta}_k) = \sum_{j\neq i}\alpha_{ij}(\theta_{i,k} - \theta_{j,k}).$$

Note that the basic iterate (2) can be interpreted as a stochastic gradient approach to solve the local problem

$$\min_{\theta_i}[\mathbb{E}_{\mathcal{P}_i}(\theta_i) + a\mathcal{F}(\boldsymbol{\theta})]$$

in which the objective function is a linear combination of loss and consensus potential.[3] When $a = 0$, each local node ignores the neighboring models. For large values of $a > 0$, the resulting dynamics reflect the countervailing effects of seeking to minimize consensus potential and improving model fit. With highly dissimilar initial models, each local node largely ignores its own data and opts for updates that lead to a model that is similar to the local average. Once approximate consensus is achieved, local gradient estimates begin to dictate the dynamics of model updates.

In what follows, it will be convenient to rewrite (2) as follows:

$$\theta_{i,k+1} = \theta_{i,k} - \gamma \mathbf{1}_{i,k}\left[\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k} + \varepsilon_{i,k}\right]. \tag{3}$$

Given that local nodes independently update and maintain their own parameters, the network regularized scheme is not subject to the possibility of biased gradient estimates stemming from update delays in FL (see [15]).

D. Literature Review

The scheme proposed in (2) has already been considered in the machine learning literature. In a series of papers (see [4]–[7]), the authors consider an approach to *multitask* learning based upon an NR penalty, as in (2). This article focuses on distributed *single-task* learning. In contrast to the papers referred earlier, we consider a nonconvex setting with heterogeneous nodes asynchronously updating their respective models at different rates over time.

The scheme proposed in (2) is also related to the literature on consensus optimization (see, e.g., [10], [16], and [17]). 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

$$\theta_{i,k+1} = \sum_j W_{i,j,k}\theta_{j,k} - \gamma g(\theta_{i,k})$$

where $\mathbf{W}_k \in \mathbb{R}^{N\times N}$ is doubly stochastic and $g(\theta_{i,k})$ is a noisy gradient estimate. Indeed, one can rewrite (2) as

$$\theta_{i,k+1} = \sum_j W_{i,j}\theta_{j,k} - \gamma \mathbf{1}_{i,k}g_i(\theta_{i,k})$$

with $W_{i,i} = 1 - \gamma a\sum_j \alpha_{i,j}$ and $W_{i,j} = \gamma a\sum_j \alpha_{i,j}$. However, the resulting matrix $\mathbf{W}$ is *not doubly stochastic* in general since we only require $a > 0$. Thus, the approach to consensus in (2) cannot be interpreted as being based upon *averaging over local* models as in consensus optimization.

The algorithms proposed in [10] and [17] are designed for *batch* data, whereas our approach deals with *streaming* data. For example, in [10], each node uses the *same* minibatch size for estimating gradients, whereas in our approach, gradient estimation noise is *heterogeneous*. In addition, in the algorithms proposed in [10] and [17], 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 are *more likely* to update their models at any given time. Finally, in [10], the objective function (loss) is defined with respect to a distribution, which is biased toward the nodes that update more often. This is in contrast to the objective function defined in this article (i.e., $\ell(\theta)$), where every node contributes to the global distribution with the same weight regardless of their updating frequency.

III. Analysis

In this section, we show that the NR scheme converges (in a certain sense) to a stationary point. To that end, we study stochastic processes $\{\theta_{i,k} : k > 0\}$ associated with each one of the $N > 1$ nodes in the network regularized approach. The proofs are given in the Appendix. We make the following standing assumptions.

Assumption 2: The graph $\mathcal{G}$ corresponding to the network of nodes is undirected ($A = A^\intercal$) and connected, i.e., there is a path between every pair of vertices.

Assumption 3 (Lipschitz): $\|\nabla\ell(\theta) - \nabla\ell(\theta')\| \leq L\|\theta - \theta'\|$ for some $L > 0$ and for all θ, θ'.

A. Preliminaries

The ensemble average $\bar{\theta}_k \triangleq \frac{1}{N}\sum_{i=1}^N \theta_{i,k}$ plays an important role in characterizing the performance of the network regularized scheme. To this end, we analyze the process $\{\overline{V}_k : k > 0\}$ defined as

$$\overline{V}_k \triangleq \frac{1}{N}\sum_{i=1}^N \|\theta_{i,k} - \bar{\theta}_k\|^2.$$

Let $e_{i,k} \triangleq \theta_{i,k} - \bar{\theta}_k$ and $V_{i,k} \triangleq \|e_{i,k}\|^2$, then $\overline{V}_k = \frac{1}{N}\sum_{i=1}^N V_{i,k}$. We now introduce some additional notations. Let $\deg(i)$ denote the degree of vertex i in graph $\mathcal{G}$ and $\overline{d} := \max_i \deg(i)$. Let $\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k]$ denote the conditional expectation of $\overline{V}_{k+1}$ given $\boldsymbol{\theta}_k$. We define $\mu_{\max} = \max\{\mu_i : 1 \leq i \leq N\}$ and $\mu_{\min} = \min\{\mu_i : 1 \leq i \leq N\}$. We first prove two intermediate results.

[2]This consensus potential has been used in the literature of opinion dynamics (see, e.g., [12]).

[3]This interpretation is not novel (see, e.g., [13] and [14] for its use in swarm (flocking) optimization and in multitask learning [4]–[7]).

Lemma 1: Suppose Assumptions 0–2 hold. It holds that

$$\overline{V}_{k+1} = \overline{V}_k - \frac{2}{N}\sum_{i=1}^{N}\gamma e_{i,k}^{\intercal}\mathbf{1}_{i,k}\left[\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}\right] - \frac{2}{N}\sum_{i=1}^{N}\gamma e_{i,k}^{\intercal}\varepsilon_{i,k}\mathbf{1}_{i,k} + \frac{1}{N}\sum_{i=1}^{N}\gamma^2\|\delta_{i,k}\|^2$$

where $\delta_{i,k} = \delta_{i,k}^f + \delta_{i,k}^g + \delta_{i,k}^n$, and

$$\delta_{i,k}^f \triangleq \nabla\ell(\theta_{i,k})\mathbf{1}_{i,k} - \nabla\bar{\ell}_k, \quad \nabla\bar{\ell}_k \triangleq \frac{1}{N}\sum_{j=1}^{N}\nabla\ell(\theta_{j,k})\mathbf{1}_{j,k}$$

$$\delta_{i,k}^g \triangleq a(\nabla\mathcal{F}_{i,k}\mathbf{1}_{i,k} - \nabla\bar{\mathcal{F}}_k),\ \delta_{i,k}^n \triangleq \varepsilon_{i,k}\mathbf{1}_{i,k} - \frac{1}{N}\sum_{j=1}^{N}\varepsilon_{j,k}\mathbf{1}_{j,k}$$

$$\nabla\bar{\mathcal{F}}_k \triangleq \frac{1}{N}\sum_{j=1}^{N}\nabla\mathcal{F}_{j,k}\mathbf{1}_{j,k}.$$

Lemma 2: Suppose Assumptions 0–3 hold. Let $\xi = \mu_{\max}/\mu_{\min}$, then

$$\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k] \leq \left(1 + \frac{\kappa\gamma}{N}\right)\overline{V}_k + \frac{4\gamma^2\xi}{N}\left\|\nabla\ell(\bar{\theta}_k)\right\|^2 + \frac{\gamma^2\xi\sigma^2}{N^2}$$

where λ_2 denotes the second-smallest of the Laplacian associated with graph $\mathcal{G}$ and

$$\kappa = 2(L\mu_{\min} - a\lambda_2\mu_{\max}) + \frac{4\gamma\xi}{N}(L^2 + 2a^2\bar{d}^2).$$

B. Convergence

We are now ready to state and prove the main theorem. As in [18], convergence is described in terms of the expected value of the average squared norm of the gradient in the first K-updates. The ensuing corollary goes into further detail by describing the same result in terms of *real-time* elapsed and not just on a total number of iterations.

Theorem 1: Suppose Assumptions 0–3 hold. Choose $\gamma < \min\{\bar{\gamma}_1, \bar{\gamma}_2\}$, where

$$\bar{\gamma}_1 = N\frac{2a\lambda_2\mu_{\max} - L(2\mu_{\min} + \xi/2)}{6\xi(L^2 + 2a^2\bar{d}^2)} \text{ and } \bar{\gamma}_2 = \frac{1}{4L(2N+1)}$$

are positive by choosing $a > \frac{4\mu_{\min}L + \xi L}{4\lambda_2\mu_{\max}}$. With scheme (2), it holds that

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\ell(\bar{\theta}_k)\|^2]\right] \leq \frac{1}{\eta K}\left[\ell(\bar{\theta}_0) + L\overline{V}_0 + \frac{KL\gamma^2\xi\sigma^2}{N^2}\left(1 + \frac{1}{2N}\right)\right]$$

where $\eta = \frac{\gamma\xi}{N}(\frac{1}{2} - 2\gamma L(2 + \frac{1}{N}))$.

The regularization penalty parameter a must be high enough to ensure cohesion between local models. This condition is weaker with a higher degree of connectivity (i.e., higher values of λ_2).

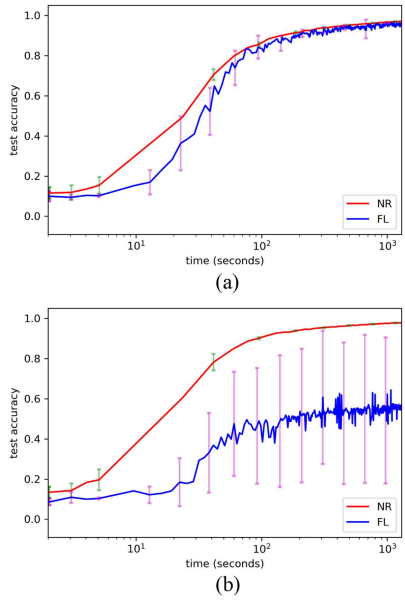


Fig. 2. Mean plot of the ensemble average computed under NR and FL schemes in heterogeneous setting. Let the network to be fully connected and set $a = 10$ and $\gamma = 0.002$ in (a) and $\gamma = 0.004$ in (b).

Note also that for fixed $N > 0$, when $a \to \infty$, then $\gamma \propto 1/a$. So convergence, as characterized by Theorem 1, may be slower. This is not necessarily the case since the conditions in Theorem 1 identify a wide range of choices for a and γ. For example, simulations indicate that for fixed γ, higher values of a may speed up convergence [see Fig. 3(c)].

C. Real-Time Performance

The analysis in Theorem 1 takes place in the time scale indexed by $k > 0$ and associated with the clicks associated with a Poisson process with rate $\mu > 0$. To embed the result in Theorem 1 in *real time*, recall that $\{D(t) : t \geq 0\}$ is the counting process governing the aggregation of all data streams. Given our assumption on computation times being negligible, the total number of updates completed in $[0, t)$ is also $D(t)$. Let us define the conditional average squared gradient norm $\|\bar{\nabla}\ell_t\|^2$ in the interval $[0, t)$ as follows:

$$\mathbb{E}[\|\bar{\nabla}\ell_t\|^2 | D(t)] \triangleq \frac{1}{D(t)}\sum_{k=1}^{D(t)}\|\nabla\ell(\bar{\theta}_k)\|^2. \tag{4}$$

Hence, the result in Theorem 1 can be reinterpreted by taking expectation of (4) over $D(t)$ as

$$\begin{aligned}\mathbb{E}[\|\bar{\nabla}\ell_t\|^2] &= \mathbb{E}\left[\mathbb{E}[\|\bar{\nabla}\ell_t\|^2|D(t)]\right] \\ &\leq \mathbb{E}\left[\frac{1}{\eta D(t)}\left(\ell(\bar{\theta}_0) + L\overline{V}_0\right) + \frac{L\gamma^2\xi\sigma^2}{\eta N^2}\left(1 + \frac{1}{2N}\right)\right] \\ &= \frac{(\ell(\bar{\theta}_0) + L\overline{V}_0)(1 - e^{-\mu t})}{\eta\mu t} + \frac{L\gamma^2\xi\sigma^2}{\eta N^2}\left(1 + \frac{1}{2N}\right).\end{aligned}$$

According to Theorem 1, and using $\gamma \sim \frac{1}{N}$, the coupling of solutions via the NR penalty implies the ensemble average

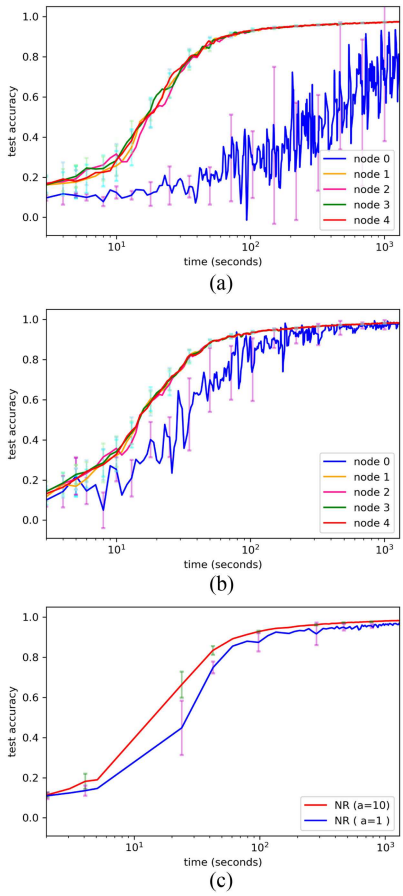


Fig. 3. Mean plot of each node computed under NR scheme in heterogeneous setting. Let the network to be fully connected and set $\gamma = 0.01$ and $a = 1$ in (a) and $a = 10$ in (b). The mean plot of the ensemble average under two choices of a is presented in (c).

approximates a stationary point in the sense that

$$\limsup_{t\to\infty} \mathbb{E}[\|\bar{\nabla}\ell_t\|^2] = \mathcal{O}\left(\frac{\sigma^2}{N^2}\right).$$

The *approximation quality* is monotonically increasing in the number of nodes. The convergence properties outlined earlier are related to the ensemble average. It is, therefore, necessary to examine the degree to which *individual* models differ from the ensemble average. This is the gist of the next result.

Corollary 1: With the same assumptions and definitions in Theorem 1, it holds that

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\overline{V}_k\right] \le \frac{1}{K|\kappa|}\left[\left(\frac{N}{\gamma}+\frac{4L\gamma\xi}{\eta}\right)\overline{V}_0+\frac{4\gamma\xi}{\eta}l(\bar{\theta}_0)\right] + \frac{4L\gamma^3\xi^2\sigma^2}{\eta|\kappa|N^2}\left(1+\frac{1}{2N}\right)+\frac{\gamma\xi\sigma^2}{|\kappa|N}.$$

We embed the result in Corollary 1 in *real time*. Define the conditional average of $\bar{V}_k$ in the interval $[0, t)$ as

$$\mathbb{E}[\bar{V}_t|D(t)] \triangleq \frac{1}{D(t)}\sum_{k=1}^{D(t)}\bar{V}_k.$$

The random process $\{\bar{V}_t : t > 0\}$ tracks the average distance of individual models to the ensemble average. Similar to the discussion of Theorem 1, the *real-time* result of Corollary 1 is as follows:

$$\begin{aligned}\mathbb{E}[\bar{V}_t] &= \mathbb{E}\left[\mathbb{E}[\bar{V}_t|D(t)]\right]\\ &\le \frac{1}{D(t)|\kappa|}\left[\left(\frac{N}{\gamma}+\frac{4L\gamma\xi}{\eta}\right)\overline{V}_0+\frac{4\gamma\xi}{\eta}l(\bar{\theta}_0)\right]\\ &\quad+\frac{4L\gamma^3\xi^2\sigma^2}{\eta|\kappa|N^2}\left(1+\frac{1}{2N}\right)+\frac{\gamma\xi\sigma^2}{|\kappa|N}\\ &= \frac{1-e^{-\mu t}}{\mu t|\kappa|}\left[\left(\frac{N}{\gamma}+\frac{4L\gamma\xi}{\eta}\right)\overline{V}_0+\frac{4\gamma\xi}{\eta}l(\bar{\theta}_0)\right]\\ &\quad+\frac{4L\gamma^3\xi^2\sigma^2}{\eta|\kappa|N^2}\left(1+\frac{1}{2N}\right)+\frac{\gamma\xi\sigma^2}{|\kappa|N}.\end{aligned}$$

This implies the asymptotic difference between individual models and the ensemble average satisfies

$$\limsup_{t\to\infty}\mathbb{E}[\bar{V}_t] = \mathcal{O}\left(\frac{\sigma^2}{N}\right).$$

The NR parameter $a > 0$ plays an important role in controlling the upper bound in Corollary 1. For fixed $N > 0$, when $a \to \infty$, then $\gamma, \eta \propto 1/a$ and $|\kappa| \propto a^2$, it follows that $\mathbb{E}[\frac{1}{K}\sum_{k=0}^{K-1}\overline{V}_k] \propto 1/a$. Hence, the upper bound in Corollary 1 can be made arbitrarily small by choosing large enough a.

D. Comparison to FL

We now present the counterpart convergence result regarding to FL.

Proposition 1: Suppose Assumptions 0–3 hold. For scheme (1), with a choice $\gamma \in (0, \frac{2}{L})$, it holds that

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\ell(\theta_k)\|^2\right] \le \frac{\ell(\theta_0)}{\tilde{\eta}K}+\frac{L\gamma^2}{2\tilde{\eta}}\sum_{i=1}^{N}\frac{\mu_i}{\mu}\sigma_i^2$$

with $\tilde{\eta} = \gamma(1-\frac{L\gamma}{2})$. To embed the process in Proposition 1 in *real time*, let us define the average squared gradient norm $\|\nabla\tilde{\ell}_t\|^2$ in the interval $[0, t)$ as follows:

$$\mathbb{E}[\|\nabla\tilde{\ell}_t\|^2|D(t)] \triangleq \frac{1}{D(t)}\sum_{k=1}^{D(t)}\|\nabla\ell(\theta_k)\|^2. \tag{5}$$

Hence, the result in Proposition 1 can be reinterpreted by taking expectation of (5) over $D(t)$ as

$$\begin{aligned}\mathbb{E}[\|\nabla\tilde{\ell}_t\|^2] &= \mathbb{E}\left[\mathbb{E}[\|\nabla\tilde{\ell}_t\|^2|D(t)]\right]\\ &\le \mathbb{E}\left[\frac{\ell(\theta_0)}{\tilde{\eta}D(t)}+\frac{L\gamma^2}{2\tilde{\eta}}\sum_{i=1}^{N}\frac{\mu_i}{\mu}\sigma_i^2\right]\\ &= \frac{\ell(\theta_0)(1-e^{-\mu t})}{\tilde{\eta}\mu t}+\frac{L\gamma^2}{2\tilde{\eta}}\sum_{i=1}^{N}\frac{\mu_i}{\mu}\sigma_i^2.\end{aligned}$$

TABLE I
EXPERIMENT HYPERPARAMETERS OF THE TWO SETTINGS, INCLUDING THE DATA STREAM ID (STREAM ID), NUMBER OF NODES INVOLVED (# NODES), THE NUMBER OF IMAGES ARRIVED AS A MINIBATCH (MINIBATCH SIZE), AND THE POISSON RATE OF THE CORRESPONDING STREAM IS (μ_i)

Setting	Stream ID	# Nodes	Mini-batch Size	μ_i
Heterogeneous	D_0	1	1	8
	$D_1 - D_4$	4	64	1
Homogeneous	All streams	20	4	1

To compare FL with NR, we also make $\gamma \sim \frac{1}{N}$. The asymptotic approximation quality is given by

$$\limsup_{t\to\infty} \mathbb{E}[\|\nabla \tilde{\ell}_t\|^2] = O\left(\frac{1}{N}\sum_i \frac{\mu_i}{\mu}\sigma_i^2\right)$$

which suggests that the approximation quality is determined by the *faster* data streams. This leads to unsatisfactory performance whenever $\mu_i \propto \sigma_i^2$ (i.e., faster data streams are also less accurate). Evidently, the opposite holds true when faster nodes are also more accurate, i.e., $\mu_i \propto 1/\sigma_i^2$. However, in many real-time machine learning applications, this is not likely to be the case. Obtaining higher precision gradient estimates requires larger batches and/or increased computation. Thus, nodes with higher precision are less likely to be the faster ones.

IV. TESTBED: REAL-TIME *DEEP* LEARNING

In this section, we report the results of NR [scheme (2)] to distributed real-time learning from three aspects: the comparison with FL [scheme (1)], the effects of the regularization parameter a, and the effects of the network connectivity.

The specific learning task is to classify handwritten digits between 0 and 9 digits as given in the MNIST dataset [19]. The dataset is composed of 10 000 testing items and 60 000 training items. Each item in the dataset is a black-and-white (single-channel) image of 28×28 pixels of a handwritten digit between 0 and 9.

In the first two experiments, we implement schemes in a heterogeneous setting with 5 nodes, and the third experiment with 20 nodes in a homogeneous setting. In the testbed, MNIST streams according to independent Poisson processes. Gradient estimates are obtained with different minibatch sizes. Evidently, a smaller minibatch size implies noisier gradient estimates. The detailed experimental settings are summarized in Table I. In the heterogeneous setting, "node 0" is the fastest and noisiest in producing gradient estimates.

We use the *Ray* platform (see [20]), which is a popular library with shared memory supported, allowing information exchange between local nodes without copying as well as avoiding a central bottleneck. For low-level computation, Google TensorFlow is used. We use a CNN with two 2-D convolutions each with kernel size 5×5, stride 1 and 32, and 64 filters. Each convolution layer is followed by a max-pooling with a 2×2 filter and stride of 2. These layers are then followed by a dense layer with 256 neurons with 0.5 dropout and sigmoid activation followed by ten output neurons and softmax operation. Cross entropy is used as a performance measure (i.e., loss).[4] We present the experimental results in mean plots with stand error bar. The means are computed across ten trials under the same hyperparameters (namely, γ and a).

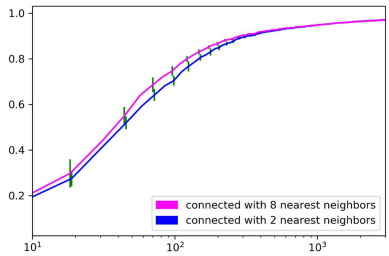


Fig. 4. Mean plot of ensemble average computed under the scheme of NR in the homogeneous setting. Set $a = 10$ and $\gamma = 0.001$.

A. Comparison to FL

In this experiment, we compare NR with FL in the heterogeneous setting. In Fig. 2, we plot the means of the ensemble average of NR and FL with different learning rates.

We can observe from Fig. 2(a) that when the learning rate is moderate, both FL and NR can converge, but the empirical standard deviation of FL is much larger than that of NR. With increased γ, FL fails to converge while NR still performs relatively well, as shown in Fig. 2(b). We can see that NR is more robust with respect to the learning rate.

B. Effects of Regularization Parameter

In this experiment, we look at the effects of changing the regularization parameter a. In Fig. 3, we present the means of each node as well as the ensemble average.

As we increase a from 1 to 10, we can observe from Fig. 3(a) and (b) that the consensus among nodes increases and the empirical mean standard deviation of the "node 0" decreases. As presented in Corollary 1, the regularization parameter a influences the degree of similarity between individual models and the ensemble average. Note that we only identify a range of values for a (lower bound) and γ (upper bound) for which convergence is guaranteed so that a higher value of a does not necessarily imply slower convergence, as shown in Fig. 3(c).

C. Effects of Network Connectivity

In the third experiment, we check the effect of increased connectivity in the homogeneous setting by using a Watts–Strogatz "small world" topology (see [21]), in which each node is connected with two (or eight) nearest neighbors.

We can see from Fig. 4 that increasing the connectivity of the topology only improves the performance slightly, meaning that only a limited connectivity is needed for the network regularized approach to enjoy a satisfactory rate of convergence.

[4]With max-pooling, the loss function is not differentiable in a set of measure zero. If in the course of execution, a nondifferentiable point is encountered, Tensorflow assumes a zero derivative. Details on the implementation are available at: https://github.com/wangluochao902/Network-Regularized-Approach

V. Conclusion

In many application domains, data streams through a network of heterogeneous nodes in different geographic locations. When there is high data payload (e.g., high-resolution video), assembling a diverse batch of data points in a central processing location in order to update a model entails significant latency. In such cases, a distributed architecture for learning, relying on a network of interconnected "local" nodes may prove advantageous. We have analyzed a distributed scheme in which every local node implements stochastic gradient updates every time a data point is obtained. To ensure robust estimation, a local regularization penalty is used to maintain a measure of cohesion in the ensemble of models. We showed that the ensemble average approximates a stationary point. The approximation quality is superior to that of FL, especially when there is heterogeneity in gradient estimation quality. We also showed that our approach is robust against changes in the learning rate and network connectivity. We illustrate the results with an application to deep learning with CNNs.

In future work, we plan to study different localized model averaging schemes. A careful selection of weights for computing local average model ensures a reduction of estimation variance. This is motivated by the literature on the optimal combination of forecasts (see [22]). For example, weights minimizing the sample mean square prediction error are of the form $\frac{\hat{\sigma}_i^{-2}}{\sum_{j=1}^N \hat{\sigma}_j^{-2}}$, where $\hat{\sigma}_i^2$ is the estimated mean squared prediction error of the ith model.

Appendix

A. Proof of Lemma 1

Note that

$$\begin{aligned}\bar{\theta}_{k+1} &= \frac{1}{N}\sum_{i=1}^N [\theta_{i,k} - \gamma \mathbf{1}_{i,k}[\nabla \ell(\theta_{i,k}) + a\nabla \mathcal{F}_{i,k} + \varepsilon_{i,k}]] \\ &= \bar{\theta}_k - \frac{\gamma}{N}\sum_{i=1}^N \nabla \ell(\theta_{i,k})\mathbf{1}_{i,k} \\ &\quad - \frac{a\gamma}{N}\sum_{i=1}^N \nabla \mathcal{F}_{i,k}\mathbf{1}_{i,k} - \frac{\gamma}{N}\sum_{i=1}^N \varepsilon_{i,k}\mathbf{1}_{i,k}.\end{aligned}$$

Hence, $e_{i,k+1} = \theta_{i,k+1} - \bar{\theta}_{k+1} = e_{i,k} - \gamma\delta_{i,k}$. Then

$$\begin{aligned}V_{i,k+1} &= (e_{i,k} - \gamma\delta_{i,k})^\intercal (e_{i,k} - \gamma\delta_{i,k}) \\ &= e_{i,k}^\intercal e_{i,k} - 2\gamma e_{i,k}^\intercal \delta_{i,k} + \gamma^2 \|\delta_{i,k}\|^2 \\ &= V_{i,k} - 2\gamma e_{i,k}^\intercal(\delta_{i,k}^f + \delta_{i,k}^g + \delta_{i,k}^n) + \gamma^2\|\delta_{i,k}\|^2\end{aligned}$$

and

$$\overline{V}_{k+1} = \overline{V}_k - \frac{2\gamma}{N}\sum_{i=1}^N e_{i,k}^\intercal(\delta_{i,k}^f + \delta_{i,k}^g + \delta_{i,k}^n) + \frac{\gamma^2}{N}\sum_{i=1}^N \|\delta_{i,k}\|^2.$$

Finally, note that

$$\begin{aligned}\sum_{i=1}^N e_{i,k}^\intercal \delta_{i,k}^f &= \sum_{i=1}^N e_{i,k}^\intercal \left[\nabla \ell(\theta_{i,k})\mathbf{1}_{i,k} - \nabla \bar{\ell}_k\right] \\ &= \sum_{i=1}^N e_{i,k}^\intercal \nabla \ell(\theta_{i,k})\mathbf{1}_{i,k} \\ \sum_{i=1}^N e_{i,k}^\intercal \delta_{i,k}^g &= a\sum_{i=1}^N e_{i,k}^\intercal(\nabla \mathcal{F}_{i,k}\mathbf{1}_{i,k} - \nabla \bar{\mathcal{F}}_k) \\ &= a\sum_{i=1}^N e_{i,k}^\intercal \nabla \mathcal{F}_{i,k}\mathbf{1}_{i,k} \\ \sum_{i=1}^N e_{i,k}^\intercal \delta_{i,k}^n &= \sum_{i=1}^N e_{i,k}^\intercal \left(\varepsilon_{i,k}\mathbf{1}_{i,k} - \frac{1}{N}\sum_{j=1}^N \varepsilon_{j,k}\mathbf{1}_{j,k}\right) \\ &= \sum_{i=1}^N e_{i,k}^\intercal \varepsilon_{i,k}\mathbf{1}_{i,k}.\end{aligned}$$

So, the result follows. ■

B. Proof of Lemma 2

In light of Lemma 1, we have

$$\begin{aligned}\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k] &= \overline{V}_k - \frac{2\gamma}{N}\sum_{i=1}^N \frac{\mu_i}{\mu} e_{i,k}^\intercal [\nabla \ell(\theta_{i,k}) + a\nabla \mathcal{F}_{i,k}] \\ &\quad + \frac{\gamma^2}{N}\sum_{i=1}^N \|\delta_{i,k}\|^2.\end{aligned}$$

Let $\mathbf{e}_k = [e_{1,k}^\intercal, e_{2,k}^\intercal, \ldots, e_{N,k}^\intercal]^\intercal$ and $\mathcal{L} = [l_{ij}]$ be the Laplacian matrix associated with the adjacency matrix A, where $l_{ii} = \sum_j a_{ij}$ and $l_{ij} = -a_{ij}$ when $i \neq j$. For an undirected graph, the Laplacian matrix is symmetric positive semidefinite. It follows that

$$\begin{aligned}\sum_{i=1}^N e_{i,k}^\intercal \nabla \mathcal{F}_{i,k} &= \sum_{i=1}^N \sum_{j=1, j\neq i}^N \alpha_{ij} e_{i,k}^\intercal (e_{i,k} - e_{j,k}) \\ &= -\sum_{i=1}^N \sum_{j\neq i}^N l_{ij} e_{i,k}^\intercal (e_{i,k} - e_{j,k}) = \sum_{i=1}^N \sum_{j\neq i}^N l_{ij} e_{i,k}^\intercal e_{j,k} \\ &= \mathbf{e}_k^\intercal (\mathcal{L}\otimes I_p)\mathbf{e}_k \geq \lambda_2 \sum_{i=1}^N \|e_{i,k}\|^2\end{aligned}$$

where $\lambda_2 := \lambda_2(\mathcal{L})$ is the second-smallest eigenvalue of $\mathcal{L}$, also called the *algebraic connectivity* of $\mathcal{G}$ [23]. Thus

$$\begin{aligned}\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k] \leq \overline{V}_k &- \frac{2\gamma}{N}\sum_{i=1}^{N}\frac{\mu_i}{\mu}e_{i,k}^{\intercal}\nabla\ell(\theta_{i,k}) \\ &- \frac{2a\lambda_2\gamma}{N}\sum_{i=1}^{N}\frac{\mu_i}{\mu}\left\|e_{i,k}\right\|^2 + \frac{\gamma^2}{N}\sum_{i=1}^{N}\mathbb{E}[\|\delta_{i,k}\|^2|\boldsymbol{\theta}_k] \\ = \overline{V}_k &- \frac{2\gamma}{N}\sum_{i=1}^{N}\frac{\mu_i}{\mu}(\nabla\ell(\theta_{i,k}) - \nabla\ell(\bar{\theta}_k))^{\intercal}e_{i,k} \\ &- \frac{2a\lambda_2\gamma}{N}\sum_{i=1}^{N}\frac{\mu_i}{\mu}\left\|e_{i,k}\right\|^2 + \frac{\gamma^2}{N}\sum_{i=1}^{N}\mathbb{E}[\|\delta_{i,k}\|^2|\boldsymbol{\theta}_k].\end{aligned}$$

By Cauchy–Schwarz inequality and Assumption 3, we can obtain that

$$\begin{aligned}-(\nabla\ell(\theta_{i,k}) - \nabla\ell(\bar{\theta}_k))^{\intercal}e_{i,k} &\leq \left\|\nabla\ell(\theta_{i,k}) - \nabla\ell(\bar{\theta}_k)\right\|\left\|e_{i,k}\right\| \\ &\leq L\left\|e_{i,k}\right\|^2.\end{aligned}$$

Define $\bar{\mu} = \mu/N$, and by the inequalities $\frac{\mu_{\min}}{N\bar{\mu}} \leq \frac{\mu_i}{\mu} \leq \frac{\mu_{\max}}{N\bar{\mu}}$, we can obtain

$$\begin{aligned}\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k] \leq (1 + \frac{2\gamma}{N\bar{\mu}}(L\mu_{\max} - a\lambda_2\mu_{\min}))\overline{V}_k \\ + \frac{\gamma^2}{N}\sum_{i=1}^{N}\mathbb{E}[\left\|\delta_{i,k}\right\|^2|\boldsymbol{\theta}_k].\end{aligned} \tag{6}$$

We now simplify the last term in the right-hand side of (6). First, we note that

$$\mathbb{E}[\|\delta_{i,k}\|^2|\boldsymbol{\theta}_k] = \mathbb{E}[\|\delta_{i,k}^f + \delta_{i,k}^g\|^2|\boldsymbol{\theta}_k] + \mathbb{E}[\|\delta_{i,k}^n\|^2|\boldsymbol{\theta}_k]. \tag{7}$$

The first term in the right-hand side of (7) can be further described as follows:

$$\begin{aligned}&\gamma^2\mathbb{E}[\|\delta_{i,k}^f + \delta_{i,k}^g\|^2|\boldsymbol{\theta}_k] \\ &= \gamma^2\mathbb{E}\left[\left\|\left(1 - \frac{1}{N}\right)[\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}]\mathbf{1}_{i,k}\right.\right. \\ &\qquad \left.\left. + \frac{1}{N}\sum_{j\neq i}^{N}[\nabla\ell(\theta_{j,k}) + a\nabla\mathcal{F}_{j,k}]\mathbf{1}_{j,k}\right\|^2|\boldsymbol{\theta}_k\right] \\ &= \frac{\gamma^2}{N}\left[\left(1 - \frac{1}{N}\right)^2\frac{\mu_i}{\bar{\mu}}\left\|\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}\right\|^2\right. \\ &\qquad \left. + \frac{1}{N^2}\sum_{j\neq i}^{N}\frac{\mu_j}{\bar{\mu}}\left\|\nabla\ell(\theta_{j,k}) + a\nabla\mathcal{F}_{j,k}\right\|^2\right].\end{aligned}$$

This leads to

$$\begin{aligned}&\sum_{i=1}^{N}\gamma^2\mathbb{E}[\|\delta_{i,k}^f + \delta_{i,k}^g\|^2|\boldsymbol{\theta}_k] \\ &\leq \frac{\gamma^2\xi}{N}\left[\left(1 - \frac{1}{N}\right)^2\sum_{i=1}^{N}\left\|\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}\right\|^2\right. \\ &\qquad \left. + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left\|\nabla\ell(\theta_{j,k}) + a\nabla\mathcal{F}_{j,k}\right\|^2\right] \\ &\leq \frac{\gamma^2\xi}{N}\sum_{i=1}^{N}\left\|\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}\right\|^2.\end{aligned} \tag{8}$$

Finally

$$\begin{aligned}&\gamma^2\sum_{i=1}^{N}\mathbb{E}[\|\delta_{i,k}^n\|^2|\boldsymbol{\theta}_k] \\ &= \frac{\gamma^2}{N}\left[\left(1 - \frac{1}{N}\right)^2\sum_{i=1}^{N}\frac{\mu_i}{\bar{\mu}}\mathbb{E}[\|\varepsilon_{i,k}\|^2|\boldsymbol{\theta}_k]\right. \\ &\qquad \left. + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\mu_i}{\bar{\mu}}\mathbb{E}[\|\varepsilon_{j,k}\|^2|\boldsymbol{\theta}_k]\right] \\ &\leq \frac{\gamma^2}{N}\left(\frac{\mu_{\max}}{\mu_{\min}}\right)\sum_{i=1}^{N}\mathbb{E}[\|\varepsilon_{i,k}\|^2|\boldsymbol{\theta}_k] \leq \frac{\gamma^2\xi\sigma^2}{N}.\end{aligned} \tag{9}$$

We use inequalities (8) and (9) with (7) to obtain an upper bound of (6) as follows:

$$\begin{aligned}\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k] \leq &\left(1 + \frac{2\gamma}{N}\left(L\mu_{\min} - a\lambda_2\mu_{\max}\right)\right)\overline{V}_k \\ &+ \frac{\gamma^2\xi}{N^2}\sum_{i=1}^{N}\left\|\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}\right\|^2 + \frac{\gamma^2\xi\sigma^2}{N^2}.\end{aligned}$$

Finally, we analyze the third term on the right-hand side of (10). By Parallelogram law

$$\begin{aligned}&\left\|\nabla\ell(\theta_{i,k}) + a\nabla\mathcal{F}_{i,k}\right\|^2 \\ &= 2\|\nabla\ell(\theta_{i,k})\|^2 + 2\|a\nabla\mathcal{F}_{i,k}\|^2 - \|\nabla\ell(\theta_{i,k}) - a\nabla\mathcal{F}_{i,k}\|^2 \\ &\leq 2\|\nabla\ell(\theta_{i,k})\|^2 + 2\|a\nabla\mathcal{F}_{i,k}\|^2.\end{aligned}$$

In addition

$$\begin{aligned}\|\nabla\mathcal{F}_{i,k}\|^2 &= \deg(i)^2\left\|\sum_{j=1,j\neq i}^{N}\frac{\alpha_{ij}(\theta_{i,k} - \theta_{j,t})}{\deg(i)}\right\|^2 \\ &\leq \deg(i)\sum_{j=1,j\neq i}^{N}\alpha_{ij}\left\|\theta_{i,k} - \theta_{j,k}\right\|^2 \\ &\leq \bar{d}\sum_{j=1,j\neq i}^{N}\alpha_{ij}\left\|\theta_{i,k} - \theta_{j,k}\right\|^2\end{aligned}$$

which implies

$$\begin{aligned}\sum_{i=1}^{N}\|\nabla\mathcal{F}_{i,k}\|^2 &\leq \bar{d}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\alpha_{ij}\left\|\theta_{i,k} - \theta_{j,k}\right\|^2 \\ &\leq 2\bar{d}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\alpha_{ij}(\|e_{i,k}\|^2 + \|e_{j,k}\|^2)\end{aligned}$$

$$\leq 4\overline{d}^2 \sum_{i=1}^{N} \|e_{i,k}\|^2 = 4N\overline{d}^2 \overline{V}_k.$$

Thus

$$\begin{aligned}
&\sum_{i=1}^{N} \|\nabla \ell(\theta_{i,k}) + a\nabla \mathcal{F}_{i,k}\|^2 \\
&\leq 2\sum_{i=1}^{N} \|\nabla \ell(\theta_{i,k})\|^2 + 8a^2 N\overline{d}^2 \overline{V}_k \\
&\leq 4N \left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + 4N(L^2 + 2a^2\overline{d}^2)\overline{V}_k.
\end{aligned} \tag{10}$$

The result follows by using the previous inequality to obtain an upper bound for the right-hand side of (10). ■

C. Proof of Theorem 1

By Taylor expansion and Lipschitz assumption

$$\begin{aligned}
\ell(\bar{\theta}_{k+1}) &\leq \ell(\bar{\theta}_k) + \nabla \ell(\bar{\theta}_k)^\intercal(\bar{\theta}_{k+1} - \bar{\theta}_k) + \frac{L}{2}\left\|\bar{\theta}_{k+1} - \bar{\theta}_k\right\|^2 \\
&= \ell(\bar{\theta}_k) - \frac{\gamma}{N}\sum_{i=1}^{N} \nabla \ell(\bar{\theta}_k)^\intercal \nabla \ell(\theta_{i,k}) \mathbf{1}_{i,k} \\
&\quad - \frac{a\gamma}{N}\sum_{i=1}^{N} \nabla \ell(\bar{\theta}_k)^T \nabla \mathcal{F}_{i,k} \mathbf{1}_{i,k} \\
&\quad - \frac{\gamma}{N}\sum_{i=1}^{N} \nabla \ell(\bar{\theta}_k)^T \varepsilon_{i,k} \mathbf{1}_{i,k} + \frac{L}{2}\left\|\bar{\theta}_{k+1} - \bar{\theta}_k\right\|^2.
\end{aligned}$$

Since $\sum_{i=1}^{N} \nabla \mathcal{F}_{i,k} = 0$, it follows that

$$\begin{aligned}
&E[\ell(\bar{\theta}_{k+1})\big|\, \boldsymbol{\theta}_k] \\
&\leq \ell(\bar{\theta}_k) - \frac{\gamma\xi}{N^2}\sum_{i=1}^{N} \nabla \ell(\bar{\theta}_k)^T \nabla \ell(\theta_{i,k}) + \frac{L}{2} E[\left\|\bar{\theta}_{k+1} - \bar{\theta}_k\right\|^2 |\boldsymbol{\theta}_k] \\
&\leq \ell(\bar{\theta}_k) - \frac{\gamma\xi}{N^2}\sum_{i=1}^{N} \nabla \ell(\bar{\theta}_k)^T [\nabla \ell(\theta_{i,k}) - \nabla \ell(\bar{\theta}_k)] \\
&\quad - \frac{\gamma\xi}{N}\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + \frac{L}{2} E[\left\|\bar{\theta}_{k+1} - \bar{\theta}_k\right\|^2 |\boldsymbol{\theta}_k].
\end{aligned} \tag{11}$$

Using (10) from the proof of Lemma 2, we obtain

$$\begin{aligned}
&E[\left\|\bar{\theta}_{k+1} - \bar{\theta}_k\right\|^2 \Big|\, \boldsymbol{\theta}_k] \\
&= \frac{\gamma^2}{N^3}\sum_{i=1}^{N} \frac{\mu_i}{\bar{\mu}} \|\nabla \ell(\theta_{i,k}) + a\nabla \mathcal{F}_{i,k}\|^2 \\
&\quad + \frac{\gamma^2}{N^3}\sum_{i=1}^{N} \frac{\mu_i}{\bar{\mu}} \mathbb{E}[\|\varepsilon_{i,k}\|^2 |\boldsymbol{\theta}_k] \\
&\leq \frac{4\gamma^2\xi}{N^2}\left[\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + (L^2 + 2a^2\overline{d}^2)\overline{V}_k\right] + \frac{\gamma^2\xi\sigma^2}{N^3}.
\end{aligned} \tag{12}$$

Also

$$\begin{aligned}
&- \nabla \ell(\bar{\theta}_k)^T [\nabla \ell(\theta_{i,k}) - \nabla \ell(\bar{\theta}_k)] \\
&= \frac{1}{2}\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + \frac{1}{2}\left\|\nabla \ell(\theta_{i,k}) - \nabla \ell(\bar{\theta}_k)\right\|^2 - \|\nabla \ell(\theta_{i,k})\|^2 \\
&\leq \frac{1}{2}\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + \frac{L^2}{2}\left\|\theta_{i,k} - \bar{\theta}_k\right\|^2.
\end{aligned} \tag{13}$$

Substituting (13) and (12) into (11), we obtain

$$\begin{aligned}
&E[\ell(\bar{\theta}_{k+1})\big|\, \boldsymbol{\theta}_k] \\
&\leq \ell(\bar{\theta}_k) - \frac{\gamma\xi}{2N}\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + \frac{L^2\gamma\xi}{2N}\bar{V}_k \\
&\quad + \frac{2L\gamma^2\xi}{N^2}\left[\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + (L^2 + 2a^2\overline{d}^2)\overline{V}_k\right] + \frac{L\gamma^2\xi\sigma^2}{2N^3}.
\end{aligned} \tag{14}$$

Consider the function $\ell(\bar{\theta}_k) + L\overline{V}_k$. From the inequalities in (14) and Lemma 2, we obtain

$$\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k] \leq (1 + \frac{\kappa\gamma}{N})\overline{V}_k + \frac{4\gamma^2\xi}{N}\left\|\nabla \ell(\bar{\theta}_k)\right\|^2 + \frac{\gamma^2\xi\sigma^2}{N^2}$$

$$\begin{aligned}
&\mathbb{E}[\ell(\bar{\theta}_{k+1}) + L\overline{V}_{k+1}|\boldsymbol{\theta}_k] \\
&\leq (\ell(\bar{\theta}_k) + L\overline{V}_k) - \frac{\gamma\xi}{N}\left(\frac{1}{2} - 2\gamma L(2 + \frac{1}{N})\right)\|\nabla \ell(\bar{\theta}_k)\|^2 \\
&\quad + \left[\kappa + \frac{L\xi}{2} + \frac{2\gamma\xi}{N}(L^2 + 2a^2\overline{d}^2)\right]\frac{L\gamma}{N}\overline{V}_k \\
&\quad + \frac{L\gamma^2\xi\sigma^2}{N^2}(1 + \frac{1}{2N}).
\end{aligned}$$

By choosing $a > \frac{4\mu_{\min}L + \xi L}{4\lambda_2\mu_{\max}}$, $\bar{\gamma}_1 > 0$. Given the choice $\gamma < \bar{\gamma}_1$ in the statement of Theorem 1, we have

$$\begin{aligned}
&\kappa + \frac{L\xi}{2} + \frac{2\gamma\xi}{N}(L^2 + 2a^2\overline{d}^2) \\
&= -2a\lambda_2\mu_{\max} + L\left(2\mu_{\min} + \frac{\xi}{2}\right) + \frac{6\gamma\xi}{N}\left(L^2 + 2a^2\overline{d}^2\right) \leq 0.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{\gamma\xi}{N}\left(\frac{1}{2} - 2\gamma L\left(2 + \frac{1}{N}\right)\right)\|\nabla \ell(\bar{\theta}_k)\|^2 \\
&\leq \ell(\bar{\theta}_k) + L\overline{V}_k - \mathbb{E}\left[\ell(\bar{\theta}_{k+1}) + L\overline{V}_{k+1}|\boldsymbol{\theta}_k\right] \\
&\qquad + \frac{L\gamma^2\xi\sigma^2}{N^2}\left(1 + \frac{1}{2N}\right).
\end{aligned}$$

Let $\eta = \frac{\gamma\xi}{N}(\frac{1}{2} - 2\gamma L(2 + \frac{1}{N}))$. By definition $\gamma < \bar{\gamma}_2$, we have $\eta > 0$. Since the loss function is non-negative, $l(\cdot) \geq 0$ and $\bar{V}_k \geq 0$ for all k. Taking full expectation and summing from $k = 0$ to $k = K - 1$ on both sides of the aforementioned inequality, we obtain

$$\mathbb{E}[\eta\sum_{k=0}^{K-1} \|\nabla \ell(\bar{\theta}_k)\|^2] \leq \ell(\bar{\theta}_0) + L\overline{V}_0 + \frac{KL\gamma^2\xi\sigma^2}{N^2}\left(1 + \frac{1}{2N}\right).$$

We conclude that

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla\ell(\bar{\theta}_k)\|^2]\right] \leq \frac{1}{\eta K}\left[\ell(\bar{\theta}_0)+L\overline{V}_0+\frac{KL\gamma^2\xi\sigma^2}{N^2}\left(1+\frac{1}{2N}\right)\right].$$

■

D. Proof of Corollary 1

Since $\gamma<\bar{\gamma}_1$, it follows that

$$\begin{aligned}\kappa &< 2L\mu_{\min}-2a\lambda_2\mu_{\max}\\ &\quad+\frac{2}{3}\left(2a\lambda_2\mu_{\max}-L\left(2\mu_{\min}+\frac{\xi}{2}\right)\right)\\ &= \frac{2}{3}L\mu_{\min}-\frac{2}{3}a\lambda_2\mu_{\max}-\frac{\xi L}{3}<0\end{aligned}$$

and from Lemma 2

$$\frac{|\kappa|\,\gamma}{N}\overline{V}_k \leq \overline{V}_k-\mathbb{E}[\overline{V}_{k+1}|\boldsymbol{\theta}_k]+\frac{4\gamma^2\xi}{N}\left\|\nabla\ell(\bar{\theta}_k)\right\|^2+\frac{\gamma^2\xi\sigma^2}{N^2}.$$

Taking full expectation and summing from $k=0$ to $k=K-1$ on both sides of the aforementioned inequality

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\overline{V}_k\right] \leq \frac{N}{K|\kappa|\gamma}\overline{V}_0+\frac{4\gamma\xi}{|\kappa|}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla\ell(\bar{\theta}_k)\right\|^2+\frac{\sigma^2}{4N}\right]$$

and using Theorem 1, we obtain the result. ■

E. Proof of Proposition 1

By Assumption 3 and Taylor expansion

$$\begin{aligned}\ell(\theta_{k+1}) &\leq \ell(\theta_k)+\nabla\ell(\theta_k)^\intercal(\theta_{k+1}-\theta_k)+\frac{L}{2}\|\theta_{k+1}-\theta_k\|^2\\ &= \ell(\theta_k)-\gamma\nabla\ell(\theta_k)^\intercal\sum_{i=1}^{N}\mathbf{1}_{i,k}g_i(\theta_k)+\frac{L}{2}\|\theta_{k+1}-\theta_k\|^2.\end{aligned}$$

Taking conditional expectation on both sides

$$\begin{aligned}&\mathbb{E}[\ell(\theta_{k+1})|\,\theta_k]\\ &\leq \ell(\theta_k)-\gamma\nabla\ell(\theta_k)^\intercal\sum_{i=1}^{N}\frac{\mu_i}{\mu}\mathbb{E}[\nabla\ell(\theta_k)+\varepsilon_i]\\ &\quad+\frac{L}{2}\mathbb{E}[\|\theta_{k+1}-\theta_k\|^2|\theta_k]\\ &= \ell(\theta_k)-\gamma\|\nabla\ell(\theta_k)\|^2+\frac{L}{2}\mathbb{E}[\|\theta_{k+1}-\theta_k\|^2|\theta_k].\end{aligned}$$

Note that

$$\mathbb{E}[\|\theta_{k+1}-\theta_k\|^2|\theta_k]=\mathbb{E}[\|\gamma\sum_{i=1}^{N}\mathbf{1}_{i,k}g_i(\theta_k)\|^2|\theta_k]$$

$$\begin{aligned}&=\gamma^2\sum_{i=1}^{N}\frac{\mu_i}{\mu}\|\nabla\ell(\theta_k)+\varepsilon_i(\theta_k)\|^2\\ &\leq\gamma^2\left(\|\nabla\ell(\theta_k)\|^2+\frac{1}{\mu}\sum_{i=1}^{N}\mu_i\sigma_i^2\right)\end{aligned}$$

it follows that

$$\gamma(1-\frac{L\gamma}{2})\|\nabla\ell(\theta_k)\|^2 \leq \ell(\theta_k)-\mathbb{E}[\ell(\theta_{k+1})|\theta_k]+\frac{L\gamma^2}{2\mu}\sum_{i=1}^{N}\mu_i\sigma_i^2.$$

The results follow by taking full expectation and summing from $k=0$ to $k=K-1$ on both sides of the aforementioned inequality. ■

REFERENCES

[1] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[3] D. Hallac, J. Leskovec, and S. Boyd, "Network lasso: Clustering and optimization in large graphs," *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 387–396.

[4] J. Chen, C. Richard, and A. Sayed, "Multitask diffusion adaptation over networks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.

[5] R. Nassif, C. Richard, A. Ferrari, and A. Sayed, "Multitask diffusion adaptation over asynchronous networks," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2835–2850, Jun. 2016.

[6] R. Nassif, S. Vlaski, and A. Sayed, "Learning over multitask graphs—Part I: Stability analysis," 2019. [Online]. Available: https://arxiv.org/abs/1805.08535

[7] R. Nassif, S. Vlaski, and A. Sayed, "Learning over multitask graphs—Part II: Performance analysis," 2019. [Online]. Available: https://arxiv.org/abs/1805.08547

[8] F. Niu, B. Recht, C. Re, and S. Wright, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 693–701.

[9] J. Liu, S. Wright, C. Ré, V. Bittorf, and S. Srikrishna, "An asynchronous parallel stochastic coordinate descent algorithm," *J. Mach. Intell. Res.*, vol. 16, pp. 285–322, 2015.

[10] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2737–2745.

[11] J. Duchi, S. Chaturapruek, and C. Ré, "Asynchronous stochastic convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1531–1539.

[12] N. Friedkin and E. Johnsen, "Social influence and opinions," *J. Math. Sociol.*, vol. 15, pp. 193–205, 1990.

[13] V. Gazi and K. Passino, *Swarm Stability and Optimization*. Berlin, Germany: Springer, 2011.

[14] S. Pu and A. Garcia, "A flocking-based approach for distributed stochastic optimization," *Oper. Res.*, vol. 6, pp. 267–281, 2017.

[15] R. Leblond, F. Pedregosa, and S. Lacoste-Julien, "Improved asynchronous parallel optimization analysis for stochastic incremental methods," *J. Mach. Intell. Res.*, vol. 19, pp. 1–68, 2018.

[16] 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.

[17] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, pp. 944–966, 2015.

[18] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2341–2368, 2013.

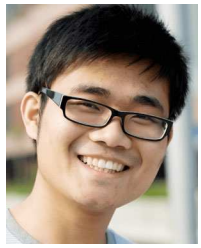
[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[20] P. Moritz *et al.*, "Ray: A distributed framework for emerging AI applications," in *Proc. 13th USENIX Symp. Oper. Syst. Des. Implementation*, 2018, pp. 561–577.
[21] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.
[22] J. M. Bates and C. M. W. Granger, "The combination of forecasts," *Oper. Res. Quart.*, vol. 20, pp. 451–468, 1969.
[23] C. Godsil and G. F. Royle, *Algebraic Graph Theory*. Berlin, Germany: Springer, 2013.

Alfredo Garcia (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Universidad de los Andes, Bogota, Colombia, in 1991, the D.E.A. degree in automatique et informatique industrielle from the Universit Paul Sabatier, Toulouse, France, in 1992, and the Ph.D. degree in industrial and operations engineering from the University of Michigan, Ann Arbor, MI, USA, in 1997.

From 1998 to 2000, he was the Commissioner with the Colombian Energy Regulatory Commission, and from 2001 to 2017, he was a Member of the Faculty with the University of Virginia and the University of Florida. He is currently a Professor with Industrial & Systems Engineering Department, Texas A&M University, College Station, TX, USA. His research interests include game theory and dynamic optimization, with applications in electricity and communication networks.

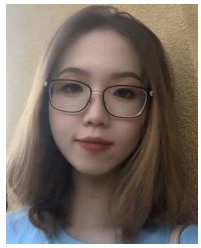
Luochao Wang is currently working toward the master's degree in computer science with Texas A&M University, College Station, TX, USA.

He was an Intern as a Machine Learning Engineer at Anadarko in 2019, working on image segmentation using CNN and oil trend forecasting with LSTM. In 2020, he was an Intern at Amazon working on internal model tracing packages. His research interests include machine learning, distributed computing, and 3-D printing.



Jeff Huang received the Ph.D. degree in computer science from the Hong Kong University of Science and Technology, Hong Kong, in 2012.

From 2013 to 2014, he was a Postdoctoral Researcher at the University of Illinois at Urbana–Champaign. He joined Texas A&M University, College Station, TX, USA, in 2014, where he is currently an Associate Professor with the Department of Computer Science & Engineering. His research focuses on developing intelligent techniques and tools for improving software performance and reliability based on fundamental program analyses and programming language theory.



Lingzhou Hong received the B.Econ. degree in statistics from the Central University of Finance and Economics, Beijing, China, in 2013, and the M.S.E. and the M.A. degrees in applied mathematics and statistics from Johns Hopkins University, Baltimore, MD, USA, in 2015 and 2017, respectively. She is currently working toward the Ph.D. degree in industrial and systems engineering from Texas A&M University, College Station, TX, USA.

Her research interests include machine learning, optimization, and statistical learning.