2020 American Control Conference
Denver, CO, USA, July 1-3, 2020

Distributed Parameter Estimation in Randomized One-hidden-layer Neural Networks

Yinsong Wang and Shahin Shahrampour

Abstract— **This paper addresses distributed parameter estimation in randomized one-hidden-layer neural networks. A group of agents sequentially receive measurements of an unknown parameter that is only partially observable to them. In this paper, we present a fully distributed estimation algorithm where agents exchange local estimates with their neighbors to collectively identify the true value of the parameter. We prove that this distributed update provides an asymptotically unbiased estimator of the unknown parameter, i.e., the first moment of the expected global error converges to zero asymptotically. We further analyze the efficiency of the proposed estimation scheme by establishing an asymptotic upper bound on the variance of the global error. Applying our method to a real-world dataset related to appliances energy prediction, we observe that our empirical findings verify the theoretical results.**

I. Introduction

Supervised learning is a fundamental machine learning problem, where given input-output data samples, a learner aims to find a mapping (or function) from inputs to outputs [1]. A good mapping is one that can be used for prediction of outputs corresponding to previously unseen inputs. Recently, deep neural networks have dominated the task of supervised learning in various applications, including computer vision [2], speech recognition [3], robotics [4], and biomedical image analysis [5]. These methods, however, are data hungry and their application to domains with few/sparse labeled samples remains an active field of research [6]. An alternative effective method for supervised learning is shallow architectures with one-hidden-layer. This architecture was motivated by the classical results of Cybenko [7] and Barron [8], showing that (under some technical assumptions) one can use sigmoidal basis functions to approximate any output that is a continuous function of the input. These results later motivated researchers to develop algorithmic frameworks to leverage shallow networks for data representation. The seminal work of Rahimi and Recht is a prominent point in case [9]. In their approach, the nonlinear basis functions are selected using Monte-Carlo sampling with a theoretical guarantee that the approximated function converges asymptotically with respect to the number of data samples and basis functions.

The problem of function approximation in supervised learning (both in shallow and deep neural networks) is often formulated via *empirical risk minimization* [1], which amounts to solving an optimization problem over a high-dimensional parameter. Due to the computational challenges associated with high-dimensional optimization, an appealing solution turns out to be decentralized training of neural networks [10]. On the other hand, recent advancement in distributed computing within control and signal processing communities [11]–[16] has provided novel decentralized techniques for parameter estimation over multi-agent networks. In these scenarios, each individual agent receives partially informative measurements about the parameter and engages in local communications with other agents to collaboratively accomplish the global task. A crucial component of these methods is a *consensus* protocol [17], allowing collective information aggregation and estimation. Distributed algorithms gained popularity due to their ability to handle large data sets, low computational burden over agents, and robustness to failure of a central agent.

Motivated by the importance of distributed computing in high-dimensional parameter estimation, in this paper, we consider distributed parameter estimation in randomized one-hidden-layer neural networks. A group of agents sequentially obtain low-dimensional measurements of the parameter (in various locations at different randomized frequencies). Despite the parameter being partially observable to each individual agent, the global spread of measurements is informative enough for a collective estimation. We propose a fully distributed update where each agent engages in local interactions with its neighboring agents to construct iterative estimates of the parameter. The update is akin to *consensus+innovation* algorithms in the distributed estimation literature [11], [13], [18].

Our main *theoretical* contribution is to characterize the first and second moments of the global estimation error. In particular, we prove that the distributed update provides an asymptotically unbiased estimator of the unknown parameter when the randomness of data samples is expected out, i.e., the first moment of the global error converges to zero asymptotically. This result also allows us to characterize the convergence rate and derive a feasible range for innovation rate. We further analyze the efficiency of the proposed estimation scheme by establishing an asymptotic upper bound on the second moment of the global error. We finally simulate our method on a real-world data related to appliances energy prediction, where we observe that our empirical findings verify the theoretical results.

II. Problem Statement

Notation: We adhere to the following notation table throughout the paper:

Yinsong Wang and Shahin Shahrampour are with the Department of Industrial and Systems Engineering at Texas A&M University, College Station, TX 77843, USA. email:{gritti@tamu.edu;shahin@tamu.edu}.

978-1-5386-8266-1/$31.00 ©2020 AACC

$[n]$	set $\{1, 2, 3, ..., n\}$ for any integer n
$\mathbf{x}^\top$	transpose of vector $\mathbf{x}$
$\mathbf{I}_M$	identity matrix of size M
$\mathbb{1}_n$	vector of all ones with dimension n
$\mathbb{0}$	vector of all zeros
$\|\cdot\|_p$	$\mathcal{L}_p$-norm operator
$\lambda_i(\mathbf{P})$	i-th largest eigenvalue of matrix $\mathbf{P}$
$\mathbb{E}[\cdot]$	expectation operator
$\rho(\mathbf{Q})$	spectral radius of matrix $\mathbf{Q}$
$\mathrm{Tr}[\cdot]$	trace operator
$\mathbf{A} \preceq \mathbf{B}$	$\mathbf{B} - \mathbf{A}$ is positive semi-definite

The vectors are in column format. Boldface lowercase variables (e.g., $\mathbf{a}$) are used for vectors, and boldface uppercase variables (e.g., $\mathbf{A}$) are used for matrices.

A. One-Hidden-Layer Neural Networks: The Centralized Problem

Let us consider a regression problem of the form

$$y = f(\mathbf{x}) + v,$$

where $y \in \mathcal{Y} \subseteq \mathbb{R}$ is the output, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ is the input, and v is the noise term with zero mean and constant variance. The objective is to find the *unknown* mapping (or function) $f : \mathcal{X} \to \mathcal{Y}$ based on available input-output pairs $\{(\mathbf{x}_j, y_j)\}$. Various regression methods assume different functional forms to approximate $f(\cdot)$. For example, in linear regression, the input-output relationship is assumed to follow a linear model. In this work, we focus on one-hidden-layer neural networks [7], where the approximated function $\widehat{f}(\cdot)$ is a nonlinear function of the input, and

$$\widehat{f}(\mathbf{x}) = \sum_{l=1}^{M} \theta_l \phi(\mathbf{x}, \boldsymbol{\omega}_l), \tag{1}$$

where ϕ is called a basis function (or *feature map*) parameterized by $\boldsymbol{\omega}_l$. In the above model, the parameters $\boldsymbol{\omega}_l$ and θ_l are unknown and should be learned from data (i.e., input-output pairs). The underlying intuition behind this model is that the feature map transforms the original data from dimension d to M, where often time we have $M \gg d$. Since the new space has a higher dimension, it provides more flexibility for approximation of the unknown function (as opposed to a linear model that is restrictive). It turns out that approximations of form (1) are dense in the space of continuous functions [7], i.e., they can be used to approximate any continuous function (on the unit cube).

However, from an algorithmic perspective, learning both θ_l and $\boldsymbol{\omega}_l$ is computationally expensive. For a nonlinear feature map ϕ (e.g., cosine feature map), the problem is indeed non-convex and thus hard to solve. An alternative approach was proposed in [9] where one-hidden-layer neural networks are thought as Monte-Carlo approximations of kernel expansions. In particular, if we assume that $\boldsymbol{\omega}$ is a random variable with a support Ω and a probability distribution $\tau(\boldsymbol{\omega})$, the corresponding kernel can be obtained via [19]

$$k(\mathbf{x}, \mathbf{x}') = \int_{\Omega} \phi(\mathbf{x}, \boldsymbol{\omega}) \phi(\mathbf{x}', \boldsymbol{\omega}) d\tau(\boldsymbol{\omega}). \tag{2}$$

Hence, if $\{\boldsymbol{\omega}_l\}_{l=1}^M$ are independent samples from $\tau(\boldsymbol{\omega})$, the approximated kernel expansion corresponds to (1) and learning θ_l becomes a convex optimization problem with a modest computational cost. $\{\boldsymbol{\omega}_l\}_{l=1}^M$ are then called *random features* in this model.

One such example is using cosine feature map to approximate a Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp \frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2}$ with unit width. In this case, (1) will be as follows

$$\widehat{f}(\mathbf{x}) = \sum_{l=1}^{M} \theta_l \sqrt{2} \cos(\boldsymbol{\nu}_l^\top \mathbf{x} + b_l), \tag{3}$$

where $\{\boldsymbol{\nu}_l\}_{l=1}^M$ come from a multi-variate Gaussian distribution $\mathcal{N}(\mathbb{0}, \mathbf{I}_d)$ and $\{b_l\}_{l=1}^M$ come from a uniform distribution $\mathcal{U}(0, 2\pi)$.

B. Local Measurements in Multi-agent Networks

The proposed scenario in the previous section was centralized in the sense that the estimation task was done only by one agent that has all the data $\{(\mathbf{x}_j, y_j)\}$. In this section, we propose an iterative distributed scheme where we have a network of n agents, each of which has access to a subset of data. In particular, agent $i \in [n]$ has access to only m_i data points at each iteration.

Assumption 1: Without loss of generality, we assume each agent observes the same number of data points at each time, i.e., $m_1 = m_2 = \cdots = m_n = c$ throughout the paper.

This assumption is only for the sake of presentation clarity. Our main results can be extended to the case where different agents have various numbers of measurements.

Now, in the distributed model, the observation matrix $\mathbf{H}_{i,t} \in \mathbb{R}^{c \times M}$ at time t will be as follows

$$\mathbf{H}_{i,t} = \begin{bmatrix} \phi(\mathbf{x}_{1,i,t}, \boldsymbol{\omega}_1) & \dots & \phi(\mathbf{x}_{1,i,t}, \boldsymbol{\omega}_M) \\ \dots & \dots & \dots \\ \phi(\mathbf{x}_{c,i,t}, \boldsymbol{\omega}_1) & \dots & \phi(\mathbf{x}_{c,i,t}, \boldsymbol{\omega}_M) \end{bmatrix}, \tag{4}$$

with any agent $i \in [n]$ having access to $\{\mathbf{x}_{j,i,t}\}_{j=1}^c$. We then have the following measurement model

$$\mathbf{y}_{i,t} = \mathbf{H}_{i,t} \boldsymbol{\theta} + \mathbf{v}_{i,t},$$

where $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]^\top \in \mathbb{R}^M$ is the *unknown* parameter that needs to be learned, and $\mathbf{v}_{i,t}$ denotes the observation noise at agent i. The above local measurement model can be interpreted as iteratively collecting low-dimensional measurements of parameter $\boldsymbol{\theta}$ at c different locations using M distinct frequencies.

We follow the general assumptions of zero mean and constant variance on the noise term, i.e., we have $\mathbb{E}[\mathbf{v}_{i,t}] = \mathbb{0}$ and $\mathbb{E}[\mathbf{v}_{i,t}\mathbf{v}_{i,t}^\top] = \sigma_v^2 \mathbf{I}_c$. We further denote by $\hat{\boldsymbol{\theta}}_{i,t}$ the estimate of $\boldsymbol{\theta}$ for agent i at time t.

Assumption 2: $\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]^\top \in \mathbb{R}^M$ is globally identifiable yet locally unobservable, i.e., the following two properties hold:

- Rank of $\mathbf{G}_i \triangleq c^{-1}\mathbb{E}[\mathbf{H}_{i,t}^\top \mathbf{H}_{i,t}]$ is strictly less than M.
- $\sum_{i=1}^n \mathbf{G}_i$ is invertible.

Note that $\mathbf{G}_i$ is also the kernel matrix formed with random features at agent i where its pq-th entry is $g_i(\boldsymbol{\omega}_p, \boldsymbol{\omega}_q) =$

$\mathbb{E}[\phi(\cdot,\boldsymbol{\omega}_p)\phi(\cdot,\boldsymbol{\omega}_q)]$. We are interchanging the role of random features $\boldsymbol{\omega}$ and data $\mathbf{x}$ here since both of them are random samples from probability measures.

Assumption 3: We assume that the feature map is bounded [9] and $\sup_{\mathbf{x},\boldsymbol{\omega}}\{|\phi(\mathbf{x},\boldsymbol{\omega})|\} \leq \sqrt{2}$. This also suggests a trivial bound where $\|\mathbf{G}_i\|_2 \leq \mathrm{Tr}\,[\mathbf{G}_i] \leq 2M$.

C. Multi-agent Network Model

The interactions of agents, which in turn defines the network, is captured with the matrix $\mathbf{P}$. Formally, we denote by $[\mathbf{P}]_{ij}$, the ij-th entry of the matrix $\mathbf{P}$. When $[\mathbf{P}]_{ij} > 0$, agent i communicates with agent j. We assume that $\mathbf{P}$ is symmetric, doubly stochastic with positive diagonal elements. The assumption simply guarantees the information flow in the network. Alternatively, from the technical point of view, we respect the following hypothesis.

Assumption 4: (connectivity) The network is connected, i.e., there is a path from any agent $i \in [n]$ to another agent $j \in [n] \setminus \{i\}$. We further assume that $\mathbf{P} = \mathbf{I}_n - \alpha\mathbf{L}$, where $\mathbf{L}$ is the Laplacian matrix and $0 < \alpha < \deg^{-1}$, where deg denotes the maximum degree of connectivity in the network.

The assumption implies that the Markov chain $\mathbf{P}$ is irreducible and aperiodic, thus having a unique stationary distribution, i.e., $\mathbb{1}^\top\mathbf{P} = \mathbb{1}^\top$ is the unique (unnormalized) left eigenvector corresponding to $\lambda_1(\mathbf{P}) = 1$. It also entails that $\lambda_1(\mathbf{P})$ is unique, and the other eigenvalues of $\mathbf{P}$ are less than unit in magnitude [20].

D. Distributed Estimation Update

To construct an iterative estimate of the parameter $\boldsymbol{\theta}$, each agent $i \in [n]$ at time t performs the following distributed update

$$\hat{\boldsymbol{\theta}}_{i,t+1} = \sum_{j=1}^{n} \mathbf{P}_{ij}\hat{\boldsymbol{\theta}}_{j,t} + \alpha\mathbf{H}_{i,t}^\top(\mathbf{y}_{i,t} - \mathbf{H}_{i,t}\hat{\boldsymbol{\theta}}_{i,t}), \quad (5)$$

where $\alpha > 0$ is the step size. The update is akin to *consensus+innovation* schemes in the distributed estimation literature [11], [13], [18], and we analyze this update in Section III in the context of one-hidden-layer neural networks. Intuitively, the first part of the update (consensus) allows agents to keep their estimates close to each other, and the second part (innovation) takes into account the new measurements.

III. Main Theoretical Results

In this section, we provide our main theoretical results. We show that the local update (5) is an asymptotically unbiased estimator of the global parameter $\boldsymbol{\theta}$. Based on this result, we derive the feasible range for step-size to guarantee convergence. We then prove that the asymptotic second moment of the collective estimation error is bounded.

A. First Moment

Let us define the *local* error for each agent $i \in [n]$ as

$$\mathbf{e}_{i,t} \triangleq \hat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}. \quad (6)$$

Subtracting $\boldsymbol{\theta}$ from both sides of the local update (5), we can write the iterative local error process as follows

$$\mathbf{e}_{i,t+1} = \sum_{j=1}^{n} \mathbf{P}_{ij}\mathbf{e}_{j,t} - \alpha\mathbf{H}_{i,t}^\top\mathbf{H}_{i,t}\mathbf{e}_{i,t} + \alpha\mathbf{H}_{i,t}^\top\mathbf{v}_{i,t}. \quad (7)$$

Stacking the local errors in a vector, we denote the *global* error by

$$\mathbf{e}_t \triangleq [\mathbf{e}_{1,t}^\top, \ldots, \mathbf{e}_{n,t}^\top]^\top. \quad (8)$$

We now characterize the global error process with the following proposition.

Proposition 1: Given Assumptions 1-4, the expected global error can be expressed as an LTI system that takes the form

$$\mathbb{E}[\mathbf{e}_t] = \mathbf{Q}\mathbb{E}[\mathbf{e}_{t-1}],$$

where

$$\mathbf{Q} \triangleq \mathbf{I}_{Mn} - \alpha\mathbf{B} \qquad \mathbf{B} \triangleq \mathbf{L} \otimes \mathbf{I}_M + c\mathbf{G}, \quad (9)$$

and $\otimes$ denotes the Kronecker product, $\mathbf{G} \triangleq \mathrm{diag}[\mathbf{G}_1, \ldots, \mathbf{G}_n]$ and $\{\mathbf{G}_i\}_{i=1}^n$ is defined in Assumption 2. The expectation is taken over the stochasticity of $\mathbf{x}$ and $\mathbf{v}$. □

The proof of proposition 1 is given in the Appendix. It shows that the agents will collectively generate estimates of the parameter $\boldsymbol{\theta}$ that are asymptotically unbiased as long as the spectral radius of $\mathbf{Q}$ is less than 1.

B. Step Size Tuning

According to Proposition 1, a sufficient condition for the convergence of the first moment is that the spectral radius of $\mathbf{Q}$ should be less than 1. The spectral radius of $\mathbf{Q}$ is decided by the following two quantities:

$$\lambda_1(\mathbf{Q}) = 1 - \alpha\lambda_{Mn}(\mathbf{B}), \quad (10)$$

and

$$\lambda_{Mn}(\mathbf{Q}) = 1 - \alpha\lambda_1(\mathbf{B}). \quad (11)$$

Now, given the condition for convergence $\rho(\mathbf{Q}) < 1$, we can derive the feasible range for step size α. According to Assumption 4, $\mathbb{1}_n$ is the (un-normalized) eigenvector of the matrix $\mathbf{L}$ associated with the unique zero eigenvalue $\lambda_n(\mathbf{L}) = 0$, because $\mathbf{L}\mathbb{1}_n = 0$. Therefore, due to Assumption 2, $\mathbf{G}\mathbb{1}_{Mn} > 0$ and $\mathbf{B}$ is always positive definite. It is then immediate that $1 - \alpha\lambda_{Mn}(\mathbf{B}) < 1$. On the other hand,

$$\alpha\lambda_1(\mathbf{B}) - 1 < 1 \Longleftrightarrow \alpha < \frac{2}{\lambda_1(\mathbf{B})}. \quad (12)$$

In conclusion, a sufficient condition for first moment convergence of global error is $\alpha < 2/\lambda_1(\mathbf{B})$.

C. Asymptotic Second Moment

To capture the efficiency of the collective estimation, we should also study the variance of the error, which (asymptotically) amounts to the second moment in view of Proposition 1. In the next theorem, we present an asymptotic upper bound on the second moment for a feasible range of step size α.

Theorem 2: Given Assumptions 1-4, the expected second moment of the estimation error is bounded under the following condition. When $\alpha < \frac{2\lambda_{Mn}(\mathbf{B})}{(\lambda_1(\mathbf{L})+2Mc)^2}$,

$$\lim_{t\to\infty} \mathbb{E}[\mathbf{e}_t^\top \mathbf{e}_t] \leq \frac{2\alpha c M n \sigma_v^2}{2\lambda_{Mn}(\mathbf{B}) - \alpha(\lambda_1(\mathbf{L}) + 2Mc)^2}.$$

The expectation is taken over the stochasticity of data $\mathbf{x}$ and observation noise $\mathbf{v}$. □

The proof of theorem 2 is given in the Appendix. It shows that the (asymptotic) expected second moment of the estimation error is bounded by a finite value that scales linearly with respect to the number of agents n for a certain range of step size α.

IV. Numerical Experiments

We now provide empirical evidence in support of our algorithm by applying it to a regression dataset on UCI Machine Learning Repository[1]. In this dataset, the input $\mathbf{x} \in \mathbb{R}^{28}$ includes a number of attributes including temperature in kitchen area, humidity in kitchen area, temperature in living room area, humidity in laundry room area, temperature outside, pressure, etc.. The regression model aims at representing appliances energy use in terms of these features. More details about this dataset can be found in [21] as well as the UCI Machine Learning Repository. We randomly choose 16000 observations out of its 19735 observations for our simulation.

We consider observation matrices $\mathbf{H}_{i,t}$ of form (4), where the bases are cosine functions as follows

$$\phi(\mathbf{x}, \boldsymbol{\omega}) = \phi(\mathbf{x}, \boldsymbol{\nu}, b) = \sqrt{2}\cos(\mathbf{x}^\top \boldsymbol{\nu} + b), \qquad (13)$$

as described in Section II-A where $\{\boldsymbol{\nu}_l\}_{l=1}^M$ come from a multi-variate Gaussian distribution $\mathcal{N}(\mathbb{0}, \mathbf{I}_d)$ and $\{b_l\}_{l=1}^M$ come from a uniform distribution $\mathcal{U}(0, 2\pi)$. Without loss of generality, we set $M = 5$, i.e., we use five basis functions in the approximation model (3). One can consider other values for M and perform cross-validation to find the best one, but this is outside of the scope of this paper, as our focus is on estimation rather than model selection.

Network Structure: We consider a network of 40 agents. Each agent i has access to observation matrix $\mathbf{H}_{i,t}$ with $c = 40$ data points at time t. Also, each agent i is connected to 4 agents $i-2, i-1, i+1, i+2$ (with a circular shift for any number outside of the range $[1, 40]$). The matrix $\mathbf{P}$ is such that agent i is connected to itself with weight 0.84 and connected to agents $i-2, i-1, i+1, i+2$ with weight 0.04. According to estimation, the largest eigenvalue of $\mathbf{B}$ is $\lambda_1(\mathbf{B}) = 42$, so according to the step size constraint for the first moment convergence, the feasible step size range is $\alpha < 0.047$. Also, the step size recommendation in Theorem 2 is $\alpha < 0.0005$, but to achieve a faster convergence, we set $\alpha = 0.04$, which violates the latter condition in theory but works in practice.

[1]https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction

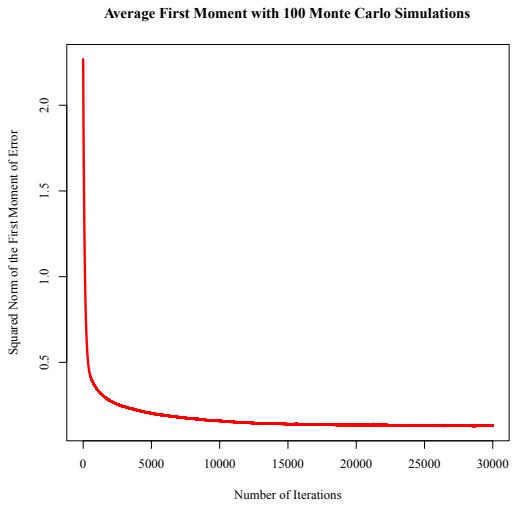

Fig. 1: The squared norm-2 of averaged (over agents) global error converges to zero over 100 Monte-Carlo simulations as the number of iterations increases.

Benchmark: Since this dataset is real-world and the ground truth value $\boldsymbol{\theta}$ is unknown, we consider the solution of the centralized problem as the baseline. The local error at time t is then calculated as the difference between local estimates $\hat{\boldsymbol{\theta}}_{i,t}$ and the centralized estimates as given in (6). We run update (5) for 30000 iterations such that the process reaches a steady state. To verify our results, we repeat the update process using Monte-Carlo simulations for 100 times by giving the agents random data points to estimate the expectations.

Performance: We visualize the error process in Proposition 1 by presenting the plot of squared norm-2 of the expected global error (averaged over agents), i.e., the squared norm-2 of $\mathbb{E}[\mathbf{e}_t]$ (divided by 40) given in Proposition 1 against number of iterations t. The vertical axis in Fig. 1 represents the average global error obtained by repeating Monte-Carlo simulations to form an estimate of the expected global error. The horizontal axis shows the number of iterations. By setting the number of Monte-Carlo simulations as 100, we can expect the squared norm-2 of the average global error converging to the squared norm-2 of the expected global error in Proposition 1. As we can observe, the estimation of the expected global error converges to zero verifying that agents form asymptotically unbiased estimators of the parameter.

We next plot the expected squared norm-2 of global error, i.e., $\mathbb{E}[\mathbf{e}_t^\top \mathbf{e}_t]$ (divided by 40) given in Theorem 2 estimated over 100 Monte-Carlo simulations. The vertical axis in Fig. 2 represents the squared norm-2 of the global error averaged over Monte-Carlo simulations. The horizontal axis shows the number of iterations. We observe that though the step size

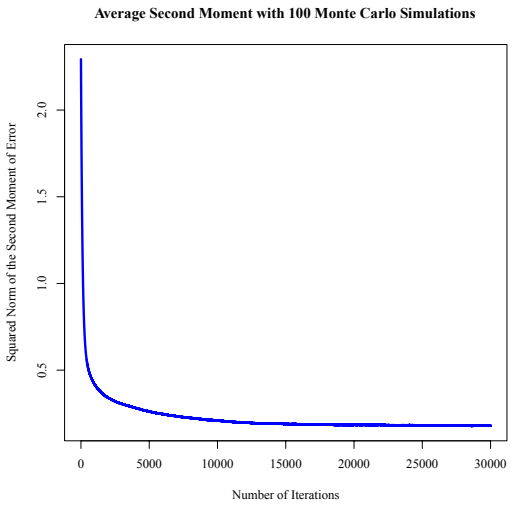

Fig. 2: The estimates across all agents have a finite variance.

does not satisfy the (sufficient) condition in Theorem 2, the second moment converges.

V. Conclusion

In this paper, we considered a distributed scheme for parameter estimation in randomized one-hidden-layer neural networks. A network of agents exchange local estimates of the parameter, formed using partial observations, to collaboratively identify the true value of the parameter. Our main contribution is to characterize the behavior of this distributed estimation scheme. We showed that the global estimation error is asymptotically unbiased and its second moment is finite under mild assumptions. Interestingly, our results shed light on the interplay of step size and network structure, which can be used for optimal design in practice. We verified this empirically by applying our method to a real-world data. Future directions include studying the estimation problem when the parameter has some dynamics [22] or the random frequencies are generated from a time-varying distribution. Due to the non-stationary nature of the problem in these two cases, the theoretical analysis becomes challenging and interesting to explore.

Appendix

For presentation clarity, we use the following definitions in the proofs:

$$\begin{aligned}\mathbf{B}_t &\triangleq \mathbf{L}\otimes\mathbf{I}_M + \text{diag}[\mathbf{H}_{1,t}^\top\mathbf{H}_{1,t},\dots,\mathbf{H}_{n,t}^\top\mathbf{H}_{n,t}]\\ \mathbf{E}_{i,t} &\triangleq \mathbf{H}_{i,t}^\top\mathbf{v}_{i,t}\\ \mathbf{E}_t &\triangleq [\mathbf{E}_{1,t}^\top,\dots,\mathbf{E}_{n,t}^\top]^\top.\end{aligned} \tag{14}$$

A. Proof of Proposition 1

Notice that $\mathbb{E}[\mathbf{H}_{i,t}^\top\mathbf{H}_{i,t}] = c\mathbf{G}_i$, entailing that

$$\mathbb{E}[\mathbf{B}_t] = \mathbf{L}\otimes\mathbf{I}_M + c\text{diag}[\mathbf{G}_1,\dots,\mathbf{G}_n] = \mathbf{B}, \tag{15}$$

in view of (14). Following the lines of the proof of Lemma 1 in [18], the error process can be expressed as the following

$$\mathbf{e}_{t+1} = \mathbf{Q}_t'\mathbf{e}_t + \alpha\mathbf{E}_t, \tag{16}$$

where

$$\mathbf{Q}_t' = \mathbf{I}_{Mn} - \alpha\mathbf{B}_t. \tag{17}$$

Taking expectation over data on both sides and noting (15), we have

$$\mathbf{Q} \triangleq \mathbb{E}[\mathbf{Q}_t'] = \mathbf{I}_{Mn} - \alpha\mathbb{E}[\mathbf{B}_t] = \mathbf{I}_{Mn} - \alpha\mathbf{B}.$$

Recalling (14), we can also immediately see from the zero-mean assumption on the noise that $\mathbb{E}[\mathbf{E}_{i,t}] = \mathbb{0}$ for every $i \in [n]$. Combining this with above and returning to (16) will finish the proof of Proposition 1.

B. Proof of Theorem 2

To prove Theorem 2, we first need to show a recursive relationship for the error process based on (16) where

$$\begin{aligned}\mathbb{E}\left[\mathbf{e}_{t+1}^\top\mathbf{e}_{t+1}\right] &= \mathbb{E}\left[(\mathbf{Q}_t'\mathbf{e}_t + \alpha\mathbf{E}_t)^\top(\mathbf{Q}_t'\mathbf{e}_t + \alpha\mathbf{E}_t)\right]\\ &= \mathbb{E}\left[\mathbf{e}_t^\top{\mathbf{Q}_t'}^\top\mathbf{Q}_t'\mathbf{e}_t\right] + \alpha^2\mathbb{E}\left[\mathbf{E}_t^\top\mathbf{E}_t\right]\\ &\le \rho\left(\mathbb{E}\left[{\mathbf{Q}_t'}^\top\mathbf{Q}_t'\right]\right)\mathbb{E}\left[\mathbf{e}_t^\top\mathbf{e}_t\right] + \alpha^2\mathbb{E}\left[\mathbf{E}_t^\top\mathbf{E}_t\right]\\ &= \lambda_1\left(\mathbb{E}\left[{\mathbf{Q}_t'}^\top\mathbf{Q}_t'\right]\right)\mathbb{E}\left[\mathbf{e}_t^\top\mathbf{e}_t\right] + \alpha^2\mathbb{E}\left[\mathbf{E}_t^\top\mathbf{E}_t\right],\end{aligned} \tag{18}$$

where we used the fact $\mathbb{E}[\mathbf{v}_{i,t}] = \mathbb{0}$, resulting in zero cross-terms in the second line. To further bound $\lambda_1(\mathbb{E}[{\mathbf{Q}_t'}^\top\mathbf{Q}_t'])$, let us recall (17), we have that

$$\begin{aligned}\mathbb{E}\left[{\mathbf{Q}_t'}^\top\mathbf{Q}_t'\right] &= \mathbb{E}\left[\mathbf{I}_{Mn} - 2\alpha\mathbf{B}_t + \alpha^2\mathbf{B}_t^2\right]\\ &= \mathbf{I}_{Mn} - 2\alpha\mathbf{B} + \alpha^2\mathbb{E}\left[\mathbf{B}_t^2\right].\end{aligned}$$

Now, observe that $\lambda_1(\mathbf{B}_t) \le \lambda_1(\mathbf{L}) + 2Mc$ due to Assumption 3. We can bound the spectral radius of the above matrix as

$$\lambda_1(\mathbb{E}[{\mathbf{Q}_t'}^\top\mathbf{Q}_t']) \le 1 - 2\alpha\lambda_{Mn}(\mathbf{B}) + \alpha^2(\lambda_1(\mathbf{L}) + 2Mc)^2. \tag{19}$$

Recalling (14), we can then bound the additive term in the recursive relation (18) as follows

$$\begin{aligned}\alpha^2\mathbb{E}\left[\mathbf{E}_t^\top\mathbf{E}_t\right] &= \alpha^2\mathbb{E}\left[\sum_{i=1}^n\mathbf{E}_{i,t}^\top\mathbf{E}_{i,t}\right]\\ &= \alpha^2\mathbb{E}\left[\sum_{i=1}^n\mathbf{v}_{i,t}^\top\mathbf{H}_{i,t}\mathbf{H}_{i,t}^\top\mathbf{v}_{i,t}\right]\\ &= \alpha^2\sum_{i=1}^n\text{Tr}\left[\mathbb{E}\left[\mathbf{H}_{i,t}\mathbf{H}_{i,t}^\top\right]\mathbb{E}\left[\mathbf{v}_{i,t}\mathbf{v}_{i,t}^\top\right]\right]\\ &= \alpha^2\sum_{i=1}^n\text{Tr}\left[\mathbb{E}\left[\mathbf{H}_{i,t}^\top\mathbf{H}_{i,t}\right]\right]\sigma_v^2\\ &= \alpha^2 c\sum_{i=1}^n\text{Tr}\left[\mathbf{G}_i\right]\sigma_v^2 \le 2\alpha^2 cMn\sigma_v^2.\end{aligned} \tag{20}$$

Letting

$$\Phi_a \triangleq 1 - 2\alpha\lambda_{Mn}(\mathbf{B}) + \alpha^2(\lambda_1(\mathbf{L}) + 2Mc)^2$$

$$\Phi_b \triangleq 2\alpha^2 cMn\sigma_v^2, \tag{21}$$

and using (19) and (20), we can re-write the recursive relation in (18) as

$$\mathbb{E}[\mathbf{e}_{t+1}^\top \mathbf{e}_{t+1}] \leq \Phi_a \mathbb{E}[\mathbf{e}_t^\top \mathbf{e}_t] + \Phi_b. \tag{22}$$

We can find the feasible range of α through the inequality $\Phi_a < 1$ which ensures that the recursive process (22) will converge.

$$\begin{aligned}
\Phi_a < 1 \Longleftrightarrow & 1 - 2\alpha\lambda_{Mn}(\mathbf{B}) + \alpha^2(\lambda_1(\mathbf{L}) + 2Mc)^2 < 1 \\
\Longleftrightarrow & \alpha^2(\lambda_1(\mathbf{L}) + 2Mc)^2 < 2\alpha\lambda_{Mn}(\mathbf{B}) \\
\Longleftrightarrow & \alpha < \frac{2\lambda_{Mn}(\mathbf{B})}{(\lambda_1(\mathbf{L}) + 2Mc)^2}.
\end{aligned} \tag{23}$$

Therefore, given $\alpha < \frac{2\lambda_{Mn}(\mathbf{B})}{(\lambda_1(\mathbf{L})+2Mc)^2}$, we have that

$$\begin{aligned}
\mathbb{E}[\mathbf{e}_{t+1}^\top \mathbf{e}_{t+1}] &\leq \Phi_a \mathbb{E}[\mathbf{e}_t^\top \mathbf{e}_t] + \Phi_b \\
&\leq \Phi_a^t \mathbb{E}[\mathbf{e}_1^\top \mathbf{e}_1] + \Phi_b(\Phi_a^{t-1} + ... + \Phi_a + 1) \\
&= \Phi_a^t \mathbb{E}[\mathbf{e}_1^\top \mathbf{e}_1] + \frac{\Phi_b(1 - \Phi_a^t)}{1 - \Phi_a}.
\end{aligned}$$

This upper bound will converge to $\frac{\Phi_b}{1-\Phi_a}$ as $t \to 0$, and noting definitions of Φ_a and Φ_b in (21), we derive the upper bound in the statement of Theorem 2.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NSF ECCS-1933878 Award as well as Texas A&M Triads for Transformation (T3) Program.

REFERENCES

[1] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, NY, USA:, 2001, vol. 1.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[4] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, pp. 1334–1373, 2016.

[5] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[6] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9359–9367.

[7] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 2, no. 4, pp. 303–314, 1989.

[8] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[9] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 1313–1320.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[11] U. Khan, S. Kar, A. Jadbabaie, J. M. Moura *et al.*, "On connectivity, observability, and stability in distributed estimation," in *IEEE Conference on Decision and Control (CDC)*, 2010, pp. 6639–6644.

[12] S. S. Stanković, M. S. Stanković, and D. M. Stipanović, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Transactions on Automatic Control*, vol. 56, no. 3, pp. 531–543, 2011.

[13] S. Kar, J. M. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012.

[14] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *IEEE Conference on Decision and Control*, 2013, pp. 6196–6201.

[15] N. Atanasov, R. Tron, V. M. Preciado, and G. J. Pappas, "Joint estimation and localization in sensor networks," in *IEEE Conference on Decision and Control (CDC)*, 2014, pp. 6875–6882.

[16] A. Mitra and S. Sundaram, "An approach for distributed state estimation of lti systems," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 1088–1093.

[17] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[18] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed estimation of dynamic parameters: Regret analysis," in *2016 American Control Conference (ACC)*, 2016, pp. 1066–1071.

[19] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[20] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[21] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data driven prediction models of energy use of appliances in a low-energy house," *Energy and buildings*, vol. 140, pp. 81–97, 2017.

[22] S. Shahrampour, S. Rakhlin, and A. Jadbabaie, "Online learning of dynamic parameters in social networks," in *Advances in Neural Information Processing Systems*, 2013.