Hogwild! over Distributed Local Data Sets with Linearly Increasing Mini-Batch Sizes

Nhuong V. Nguyen[†] , Toan N. Nguyen[†] , Phuong Ha Nguyen[§] , Quoc Tran-Dinh[‡] , Lam M. Nguyen[††] , Marten van Dijk[†,*]

[†] University of Connecticut, [§] eBay, [‡] University of North Carolina at Chapel Hill
[††] IBM Research, Thomas J. Watson Research Center, * CWI Amsterdam

March 20, 2021

Introduction

Solve:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \mathbb{E}[f(w; \xi)] \right\} \tag{1}$$

Here, ξ is a random variable obeying some distribution, F has a Lipschitz continuous gradient and f is bounded from below for every ξ.

Type	Data Training	Collective Learning	Synchrony	Exchanged Information
SGD	Shared	No	N/A	Gradients
Hogwild!	Shared	Yes	Asynchrony	Gradient
Local SGD	Shared	Yes	Synchrony	Models
Federated Learning	Separated	Yes	Synchrony	Models
Our Framework	Separated	Yes	Asynchrony	Models

Figure 1: Summary information of SGD-based methods.

Motivation

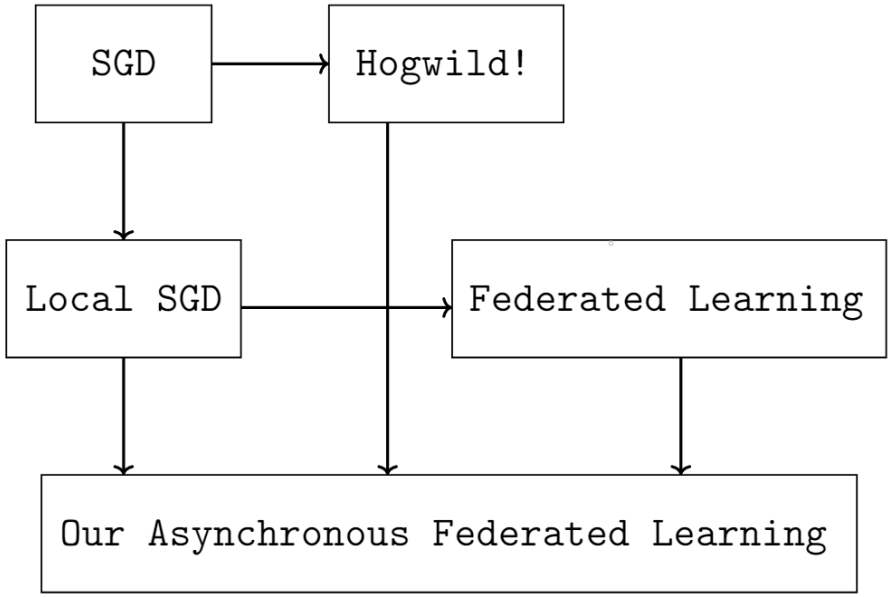


Figure 2: Related work diagram.

Motivation. Can we develop a new asynchronous Federated Learning with **rigorous convergence proof**? Yes, we can. As shown in our paper, based on the Hogwild!'s framework, we can propose a desired framework with a rigorous convergence proof for strongly convex case. Moreover, this is the first time we can show that the system with heterogeneous data sets can converge with $O(\sqrt{K})$ communications where K is the number of gradient computations.

Our Asynchronous Federated Learning

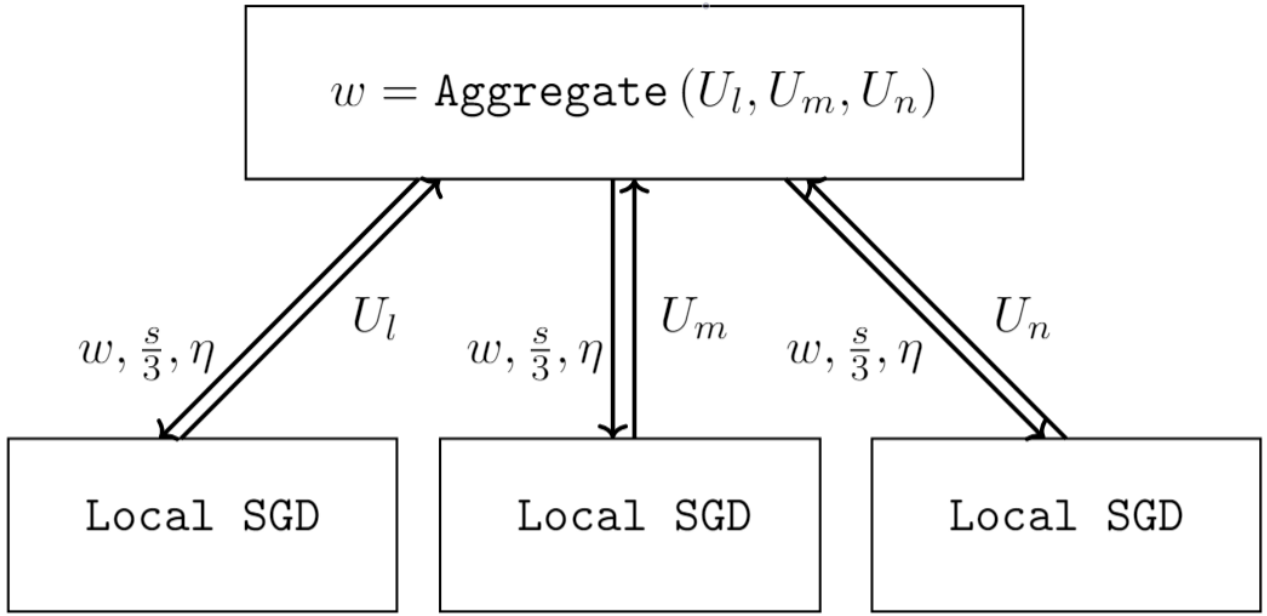


Figure 3: Example of our asynchronous Federated Learning framework.

Setting. The server and n clients are working in asynchronous fashion with i.i.d or heterogeneous data sets. At the client side, after running each s_i/n-iteration SGD with stepsize η_i, it sends the sum of computed gradients U to the server. Based on the received Us from the clients, the server issue a new global w and broadcast it to clients.

Sample Size Sequences

Lemma 1 (Sample size). Let $g > 1$. Suppose that $\tau(x) = M_1 + (x + M_0)^{1/g}$ for some $M_1 \geq d + 2$ and $M_0 \geq ((m+1)(g-1)/g)^{g/(g-1)}$, where $m \geq 0$ is an integer. Then

$$s_i = \left\lceil \frac{1}{d+1} \left(\frac{m+i+1}{d+1} \frac{g-1}{g} \right)^{1/(g-1)} \right\rceil$$

satisfies property (2).
If delay function $\tau(t) \leq \sqrt{(t/\ln t) \cdot (1 - 1/\ln t)}$ [2], then $s_i = \Theta\left(\frac{i}{\ln i}\right)$.

Delay property. We exploit the algorithm's resistance against delays $\tau(t)$ by using increasing sample size sequences $\{s_i\}$. Sample size sequences should not increase too much: We require the property that there exists a threshold d such that for all $i \geq d$,

$$\tau\left(\sum_{j=0}^{i} s_j\right) \geq 1 + \sum_{j=i-d}^{i} s_j. \tag{2}$$

Round Step Size Sequences

Lemma 2. Let $0 \leq q \leq 1$ and $\{E_t\}$ a constant or increasing sequence with $E_t \geq 1$. For q and $\{E_t\}$ consider the set $\mathcal{Z}$ of diminishing step size sequences $\{\eta_t\}$ with $\eta_t = \alpha_t/(\mu(t+E_t)^q)$ where $\{\alpha_t\}$ is some sequence of values with $\alpha_0 \leq \alpha_t \leq 3 \cdot \alpha_0$. We assume sample size sequence $\{s_i\}$ of Lemma 1 for $g \geq 2$. For $i \geq 0$, we define $\bar{E}_i = E_{\sum_{j=0}^{i} s_j}$. We define $\bar{E}_{-1} = E_0$. If $\bar{E}_i \leq 2\bar{E}_{i-1}$ for $i \geq 0$ and if $s_0 - 1 \leq E_0$, then there exists a diminishing step size sequence $\{\eta_t\}$ in set $\mathcal{Z}$ such that

$$\eta_t = \frac{\alpha_t}{\mu(t+E_t)^q} = \frac{\alpha_0}{\mu((\sum_{j=0}^{i-1} s_j) + \bar{E}_{i-1})^q} \overset{\text{DEF}}{=} \bar{\eta}_i$$

for $t \in \left\{ \left(\sum_{j=0}^{i-1} s_j\right), \ldots, \left(\sum_{j=0}^{i-1} s_j\right) + s_i - 1 \right\}$.

Theoretical Analysis

Assumptions:

1. $F(w)$ is L-smooth: $\forall w, w' \in \mathbb{R}^d$, $\exists L > 0$:

$$\|\nabla F(w) - \nabla F(w')\| \leq L\|w - w'\|$$

2. $f(w;\xi)$ is convex for every realization of ξ, i.e, $\forall w, w' \in \mathbb{R}^d$:

$$f(w;\xi) - f(w';\xi) \geq \langle \nabla f(w';\xi), (w - w') \rangle$$

3. F is μ-strongly convex:

$$F(w) - F(w') \geq \langle \nabla F(w'), (w - w') \rangle + \frac{\mu}{2}\|w - w'\|^2$$

4. Finite variance at minimizing solutions: Let $N = 2\mathbb{E}[\|\nabla f(w_*;\xi)\|^2]$ where $w_* = \arg\min_w F(w)$, then $N < \infty$.

Theoretical Analysis

Theorem 2 (Upper Bound). For sample size sequence $\{s_i\}$ and round step size sequence $\{\bar{\eta}_i\}$ we have expected convergence rate

$$\mathbb{E}[\|w_t - w_*\|^2] \leq \frac{4 \cdot 36^2 \cdot N}{\mu^2} \frac{1}{t} + O\left(\frac{1}{t \ln t}\right). \tag{3}$$

where t represents the total number of gradient evaluations over all compute nodes performed so far.

Corollary 1 (Lower Bound). Among first order stochastic algorithms, upper bound (3) converges for increasing t to within a constant factor $8 \cdot 36^2$ of the (theoretically) best attainable expected convergence rate, which is at least $\frac{N}{2\mu^2} \frac{1}{t} \left(1 - O\left(\frac{\ln t}{t}\right)\right)$ (for each t). Notice that the factor is independent of any parameters like L, μ, sparsity, or dimension of the model.

Experimental Result

We verify our asynchronous Federated Learning framework with various objective functions under *different step sizes* and *data sets* setting. Here, diminishing$_1$ represents diminishing step size scheme over *iterations* and diminishing$_2$ represents diminishing step size scheme over *rounds*.

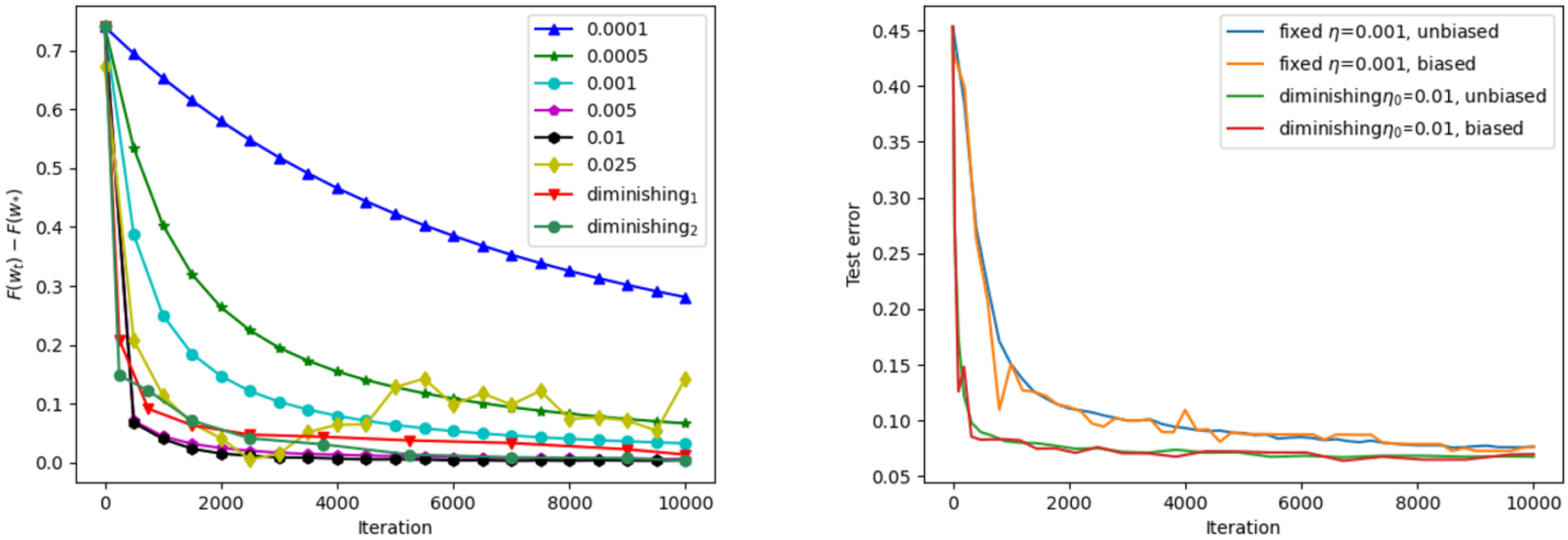
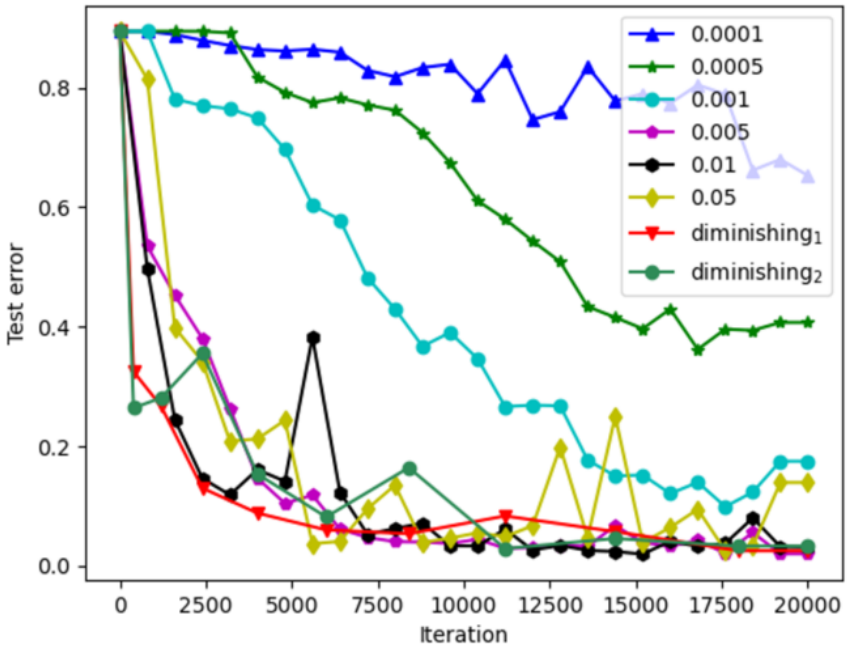


Figure 4: Our asynchronous Federated Learning for strongly convex problems: The MNIST data set - various step size sequences (left), MNIST (digit 0/1) - biased and unbiased data set (right).

Experimental Result



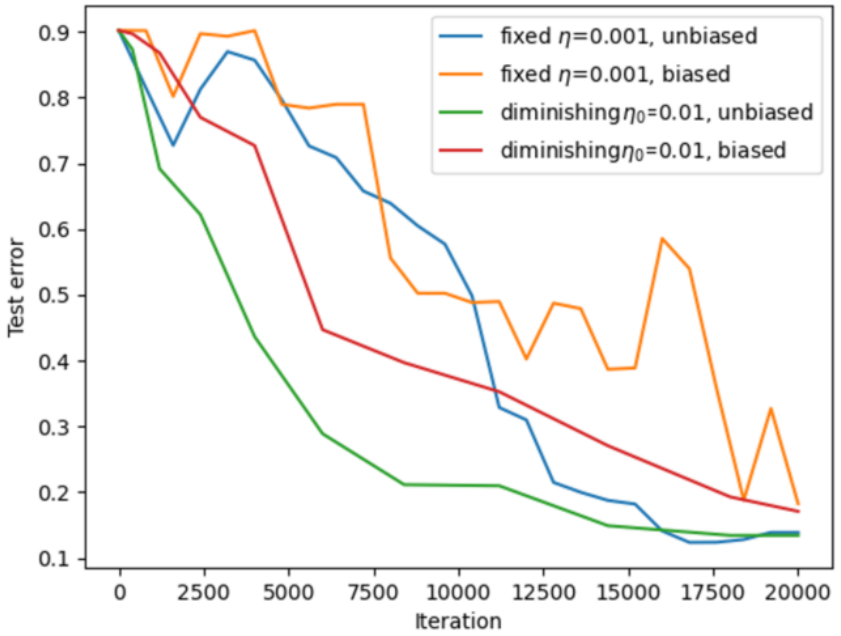


Figure 5: Our asynchronous Federated Learning for non-convex problems: The MNIST data set - various step size sequences (left), MNIST - biased and unbiased data set (right).

Contributions

1. Propose an asynchronous distributed learning framework based on Hogwild! algorithm, which can work well with iid and heterogeneous data sets.
2. Provide a general recipe for constructing increasing sample size sequences and diminishing round step size sequences so that our algorithm maintains τ as an invariant. For strongly convex problems, we prove that $\tau(t)$ can be as large as $\approx \sqrt{t/\ln t}$ for which our recipe shows a diminishing round step size sequence of $O\left(\frac{\ln i}{i^2}\right)$, where i indicates the round number, that allows a sample size sequence of $\Theta\left(\frac{i}{\ln i}\right)$; the sample size sequence can almost *linearly* increases from round to round.
3. Provide an upper bound of $O(1/t)$ on the expected convergence rate $\mathbb{E}[\|w_t - w_*\|^2]$, where w_* represents the global minimum in (1) and t is the SGD iteration number (each local node computes a subset of the w_t).
4. Let K be the total number of gradient computations (summed over all local nodes) and let T be the number of communication rounds in our algorithm. Then, T scales less than linear with K due to the increasing sample size sequence. For strongly convex problems with diminishing step sizes we show $T = O(\sqrt{K})$ for heterogeneous local data while having $O(1/K)$ convergence rate.

References

1. Niu, Feng, Benjamin Recht, Christopher R, and Stephen J. Wright:Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In NIPS 2014
2. Lam M. Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg and Martin Takac: SGD and Hogwild! Convergence without the bounded gradients assumption. In ICML 2018
3. McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas: Communication-efficient learning of deep networks from decentralized data. In AISTATS 2017

Thanks for your listening!