ON FEDERATED LEARNING WITH ENERGY HARVESTING CLIENTS

Cong Shen[*]*, Jing Yang*[†]*, and Jie Xu*[‡]

[*] University of Virginia
[†]The Pennsylvania State University
[‡]University of Miami

ABSTRACT

Catering to the proliferation of Internet of Things devices and distributed machine learning at the edge, we propose an energy harvesting federated learning (EHFL) framework in this paper. The introduction of EH implies that a client's availability to participate in any FL round cannot be guaranteed, which complicates the theoretical analysis. We derive novel convergence bounds that capture the impact of time-varying device availabilities due to the random EH characteristics of the participating clients, for both parallel and local stochastic gradient descent (SGD) with non-convex loss functions. The results suggest that having a uniform client scheduling that maximizes the minimum number of clients throughout the FL process is desirable, which is further corroborated by the numerical experiments using a real-world FL task and a state-of-the-art EH scheduler.

***Index Terms*—** Federated learning, energy harvesting, stochastic gradient descent, convergence analysis.

1. INTRODUCTION

Federated learning (FL) is a novel machine learning (ML) paradigm that builds a global ML model by training at many distributed clients. FL represents an ongoing paradigm shift towards moving the data collection and model training away from the server and to the edge [1, 2]. The proliferation of Internet of Things (IoT) devices that produce massive amount of data directly at the edge devices, the desire to reduce data transfer to the cloud, and the need to improve ML responsiveness have made FL in IoT networks an important application.

Despite its potential and impact, FL in IoT networks is a difficult task as IoT devices are highly resource constrained. In particular, this paper focuses on enabling FL with energy harvesting (EH) devices [3, 4], where the computation [5, 6, 7] and communication [8, 9, 10, 11, 12] operations of FL at an EH device depend entirely on its harvested energy. The focus of FL with EH devices is motivated by the rapid deployment of these devices in IoT networks, such as the agricultural application where devices may be exclusively powered by ambient energy sources such as wind or solar [13].

The main challenge, however, is that the introduction of EH devices complicates the already difficult FL problem. In particular, FL cannot narrowly focus on each learning round, but must consider the temporal correlation of progressive learning rounds that collectively determine the final learning outcome. With EH devices, the availability of any given client is no longer guaranteed for FL in a given round, if it does not have sufficient energy for computation and communication. Furthermore, the random evolution of the energy queue at each device also has temporal correlation that depends on both the energy arrival process and the FL client scheduling algorithm. The coupled temporal correlations of the FL process and the EH process represent a significant challenge in both theoretical analysis and algorithm design, suggesting that one cannot separately consider the EH design and FL design when optimizing the overall system performance.

In this paper, we propose an energy harvesting federated learning (EHFL) framework, where EH clients are scheduled to participate in the FL process. To address the aforementioned challenges of EHFL, we first analyze the convergence behavior of FL under an arbitrary sequence of available clients that participate in the corresponding learning rounds. This analysis is useful in that the sequence of clients can be viewed as the output of an EH client scheduler, and optimizing the resulting convergence bound sheds light on the desired behavior of the EH scheduler. A unified principle for both parallel and local stochastic gradient descent (SGD) emerges from the analysis, which suggests that a uniform client scheduling that maximizes the minimum number of clients in FL is beneficial. This theoretical result is corroborated by a numerical experiment using the standard CIFAR-10 classification task and a state-of-the-art EH scheduler.

2. THE EHFL FRAMEWORK

The proposed energy harvesting federated learning (EHFL) framework is illustrated in Fig. 1. This framework is notably

CS was supported in part by the National Science Foundation (NSF) under CNS-2002902, ECCS-2029978, and ECCS-2033671. JY was supported in part by the NSF under CNS-1956276, CNS-2003131, ECCS-2030026, and CNS-211454. JX was supported in part by the NSF under ECCS-2033681 and CNS-2044991.

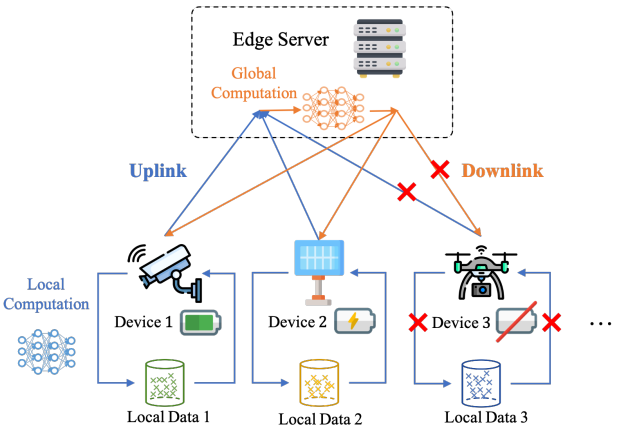

Fig. 1. Illustration of the EHFL framework.

different from standard FL, because the introduction of EH devices implies that a client's availability to participate in any round cannot be guaranteed. FL must deal with different sets of available clients that are determined *exogenously* (by the EH scheduler) in every round, which would affect the model convergence. To further complicate the analysis, such client availability is not independent over time, as clients who have participated in one round and consumed the harvested energy are less likely to have sufficient energy for the next round.

Federated learning model. In a typical case, the goal of FL is to solve the standard empirical risk minimization (ERM) problem:

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{D}\sum_{z\in\mathcal{D}} l(x;z),$$

in a distributed fashion, where $x \in \mathbb{R}^d$ is the machine learning model variable that we would like to optimize, $l(x;z)$ is the loss function evaluated at model x and data sample z, and $f : \mathbb{R}^d \to \mathbb{R}$ is the differentiable loss function averaged over the total dataset $\mathcal{D}$ with size D. We denote $x^* := \arg\min_{x\in\mathbb{R}^d} f(x)$, and $f^* := f(x^*)$. We denote the *maximum* number of clients in the FL system as M, and the total global dataset is the union of all local datasets at these M clients: $\mathcal{D} = \bigcup_{m=1}^{M} \mathcal{D}_m$. We assume that $\mathcal{D}_i$ has D_i data samples at client $i \in [M] := \{1, \cdots, M\}$, and all local datasets are non-overlapping, hence $\sum_i D_i = D$. Note that M is generally not the number of clients that participate in FL in any given learning round. The original ERM problem can be rewritten as

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \sum_{i=1}^{M} \frac{D_i}{D} f_i(x),$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is the local loss function for client i, averaged over its local dataset $\mathcal{D}_i$, i.e., $f_i(x) = \frac{1}{D_i}\sum_{\xi\in\mathcal{D}_i} l(x;\xi)$.

We consider that local SGD [14] is adopted to solve the FL problem. In the t-th round of local SGD, $t = 1, \cdots, T$, there are n_t clients $\mathcal{N}_t := \{m_1, \cdots, m_{n_t}\}$ who actively participate in FL. Each client independently runs K individual SGD steps before aggregating the local models at the server. Specifically, the t-th round starts with client $i \in \mathcal{N}_t$ receiving the latest global model x_t from the parameter server: $x_t^i = x_{t,0}^i = x_t$. It then runs K steps of stochastic gradient evaluation:

$$x_{t,\tau+1}^i = x_{t,\tau}^i - \eta_t \nabla \tilde{f}_i(x_{t,\tau}^i), \forall \tau = 0, \cdots, K-1. \quad (1)$$

The client's updated model after these K steps can be written as $x_{t+1}^i = x_{t,K}^i$. Notation wise, we use $\tilde{f}_i(x) := l(x;\xi_i)$ to denote the loss function of model x evaluated with a random data sample ξ_i at client i. The server collects the local models $\{x_{t+1}^i, i \in \mathcal{N}_t\}$ and computes a simple aggregation $x_{t+1} = \frac{1}{n_t}\sum_{i\in\mathcal{N}_t} x_{t+1}^i$ as the global model for the next round. Local SGD then moves on to the $(t+1)$-th round.

Energy harvesting model. In EHFL, each client $i \in [M]$ is powered by energy harvested from the ambient environment. We assume that each client has an energy queue (rechargeable batteries or capacitors) to store the harvested energy. The energy queue at each client is replenished randomly and consumed by computation and communication for FL. We assume that the energy unit is normalized so that if a device participates in one round of FL, it consumes one unit of energy. This energy unit represents the cost of both computation and communication. We assume the duration between two consecutive rounds is fixed.

Let $E_i(t)$ be the total amount of energy units available at the beginning of round t at device i, and $A_i(t)$ be the amount of energy units harvested during the t-th round. We assume $A_i(t)$ is an independent and identically distributed (IID) Bernoulli random variable with $\mathbb{E}[A_i(t)] = \lambda_i$. Different values of λ_i capture the energy heterogeneity among clients. Then, the energy level at device i evolves according to the following equation:

$$E_i(t+1) = \min\{(E_i(t) - \mathbf{1}\{i \in \mathcal{N}_t\}) + A_i(t), E_{\max}\} \quad (2)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, $E_{\max}$ is the capacity of the battery, and the energy causality condition requires that $E_i(t) \geq \mathbf{1}\{i \in \mathcal{N}_t\}$ for all i, t.

3. CONVERGENCE ANALYSIS FOR EHFL

We analyze the convergence of FL with an arbitrary sequence of participating clients $\{\mathcal{N}_1, \cdots, \mathcal{N}_T\}$ as the output of the EH scheduler, with non-convex loss functions. We first focus on a special case of *parallel SGD*, which refers to distributed SGD with per-step model average, to gain some insight of the FL convergence behavior due to the random EH characteristics. We then extend the analysis to *local SGD* with periodic model average whose period is strictly larger than one. Finally we summarize the main theoretical result and discuss its implication on the EH scheduler design.

3.1. Parallel SGD: $K = 1$

3.1.1. Assumptions

We limit our attention to L-smooth (possibly non-convex) loss functions, as stated in Assumption 1. In addition, we assume that the stochastic gradients are unbiased at all clients, and the variance is (uniformly) bounded in Assumption 2.

Assumption 1 *$l(x,\xi)$ is L-smooth: $\|\nabla l(x,\xi)-\nabla l(y,\xi)\| \leq L\|x-y\|$ for any $x,y \in \mathbb{R}^d$ and any $\xi \in \mathcal{D}$.*

Assumption 2 *SGD is unbiased at all clients: $\mathbb{E}_\xi \nabla f_i(x) = \nabla f(x), \forall i$, and its variance is bounded: $\mathbb{E}_\xi \|\nabla l(x,\xi) - \nabla f(x)\|^2 \leq \sigma^2$.*

3.1.2. Main result

We note that for non-convex loss functions, it is well-known that SGD may converge to a local minimum or saddle point, and it is a common practice to evaluate the expected gradient norms as an indicator of convergence. In particular, an algorithm achieves an ϵ-suboptimal solution if $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \epsilon$, which guarantees the convergence to a stationary point [15].

We now state our main result in Theorem 1. Due to the space limitation, detailed proofs for both theorems are deferred to the online appendix [16].

Theorem 1 *Suppose Assumptions 1 and 2 hold. Consider an energy harvesting client scheduler that produces n_t clients to participate in the t-th round parallel SGD. Assume $0 < n_{\min} \leq n_t \leq n_{\max} \leq M$, and we choose a parameter η satisfying $0 < \eta \leq \frac{1}{L}\sqrt{\frac{T}{n_{\max}}}$. Then, if we set the learning rate of SGD as*

$$\eta_t = \eta\sqrt{\frac{n_t}{T}}, , \forall t = 0, \cdots, T-1,$$

the convergence of parallel SGD with non-convex loss functions and IID local datasets satisfies:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{f(x_0)-f^*}{\eta\sqrt{n_{\min}T}-\frac{L}{2}\eta^2 n_{\min}} + \frac{L\sigma^2}{2\eta\sqrt{n_{\min}T}-L\eta^2 n_{\min}} \sim \mathcal{O}\left(\frac{1}{\sqrt{n_{\min}T}}\right). \quad (3)$$

Remark 1 The key novelty in this theorem is to establish the relationship $\eta_t = \eta\sqrt{\frac{n_t}{T}}$, which is accomplished by minimizing the derived upper bound as a general function of η_t and n_t. Theorem 1 states that if we tie the choice of learning rate to the available number of clients according to $\eta_t \sim \mathcal{O}\left(\sqrt{n_t}\right)$, then we achieve the same $\mathcal{O}\left(1/\sqrt{T}\right)$ convergence rate as the constant-client parallel SGD [14].

Remark 2 It is known that within a proper range that guarantees the convergence, selecting larger stepsize has the benefit of speeding up the SGD process. In this spirit, a particular choice of η is $\eta = \frac{1}{L}\sqrt{\frac{T}{n_{\max}}}$, which leads to $\eta_{\min} := \min \eta_t = \frac{1}{L}\sqrt{\frac{n_{\min}}{n_{\max}}}$. This results in a convergence scaling of $\mathcal{O}\left(\sqrt{\frac{n_{\max}}{n_{\min}}}\frac{1}{\sqrt{T}}\right)$. Clearly, selecting a uniform client scheduling such that $n_{\max} = n_{\min}$ minimizes the coefficient of $\frac{1}{\sqrt{T}}$. This insight thus provides a theoretical guidance for the EH scheduler design.

Remark 3 Assumption 2 corresponds to the so-called IID local dataset setting for FL. How to extend the analysis to non-IID local datasets is an interesting future research direction.

3.2. Local SGD: $K > 1$

We now analyze the case of local SGD with $K > 1$. The main result is stated as follows.

Theorem 2 *Suppose Assumptions 1 and 2 hold. Consider an energy harvesting client scheduler that produces n_t clients to participate in the t-th round local SGD. Assume $0 < n_{\min} \leq n_t \leq n_{\max} \leq M$, and we choose a parameter η satisfying $0 < \eta \leq \frac{1}{2KL}\sqrt{\frac{1}{30n_{\max}}}$. Then, if we set the stepsize of SGD at the t-th round as*

$$\eta_t = \eta\sqrt{\frac{n_t}{T}}, \forall t = 0, \cdots, T-1,$$

then we achieve the following convergence of local SGD with non-convex loss functions:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{\frac{2}{K}(f(x_0)-f^*)+L\sigma^2\eta^2}{\eta\sqrt{n_{\min}T}-\sqrt{30}KL\eta^2 n_{\min}} + \frac{5KL^2\sigma^2\eta^3 n_{\max}^{\frac{3}{2}}}{\eta\sqrt{n_{\min}T}-\sqrt{30}KL\eta^2 n_{\min}\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{n_{\min}T}}\right). \quad (4)$$

Remark 4 The key challenge for analyzing local SGD is that the gradient estimation after the first step becomes *biased*, i.e., they do not represent the true gradients in expectation. Having a varying n_t means that different rounds are "heterogeneous" in terms of averaging the biased SGDs with varying variances, which cannot be easily handled when bounding the convergence rate. The proof relies on enhancing the *perturbed iterate framework* [17] to decouple the impact of each additional SGD step by a careful construction of the virtual model sequence. This allows us to derive an η_t-dependent upper bound for the average (over n_t clients) gradient for each SGD step $\tau = 0, \cdots, K-1$. This bound is then utilized in the enhanced perturbed iterate framework to derive a non-trivial (n_t, η_t)-dependent convergence rate upper bound.

Then, similar to Theorem 1, we can minimize this bound over the choice of η_t as a function of n_t.

Remark 5 Theorem 2 unifies the selection of learning rate as a function of the EH device availability for both parallel and local SGDs (at least with respect to the scaling), which suggests that the EH scheduler design can be agnostic to the SGD steps chosen by the FL task. This is an important feature that improves the generalization of the proposed EHFL framework in terms of the performance guarantees.

3.3. EH scheduler design

The convergence analysis for both parallel and local SGD indicates that maintaining a balanced number of clients participating in each round throughout the learning horizon is desirable. However, strictly maintaining a constant number of clients in the face of stochastic energy arrival and energy causality constraint is a very challenging task, not to mention the inhomogeneous EH processes at clients.

In order to gain some intuition of the desired EH scheduler design, we first ignore the stochasticity of the EH process and focus on the long-term average EH rate instead. Given the total EH rate $\Lambda := \sum_{i=1}^{M} \lambda_i$ and the energy flow conservation condition (i.e., energy consumption rate must be upper bounded by the energy arrival rate), the average number of active clients in each round must be upper bounded by Λ as well. For a clear exposition of our rationale, we assume Λ is an integer. Thus, if we are able to obtain a subset of clients $\mathcal{N}_t$ in round t such that $|\mathcal{N}_t| = \Lambda$ with high probability, then we can expect that the $n_{\min}$ throughout the learning process is maximized, and the convergence rate can thus be optimized with high probability based on our theoretical results. The problem then boils down to ensuring such a selection of $\mathcal{N}_t$ is feasible in each round, in the presence of stochastic energy arrivals and heterogeneous EH rates across the clients.

In our previous work [9], we have developed an energy queue length based *myopic* scheduling policy when $E_{\max} = \infty$. At the beginning of round t, the scheduler first selects Λ clients with the longest energy queues and forms a candidate set of active clients, denoted as $\mathcal{N}'_t$. Then, it determines $\mathcal{N}_t = \{i : i \in \mathcal{N}'_t, E_i(t) \geq 1\}$. The myopic scheduling policy has a queue-length balancing nature, i.e., it tries to equalize the battery levels of all clients by prioritizing clients with longer energy queues. As a result, it ensures that $|\mathcal{N}_t| = \Lambda$ in almost every round t. We will evaluate the performance of this myopic EH scheduling policy in the experiment.

4. SIMULATION RESULTS

Experiment setup. We have carried out an experiment on the standard real-world CIFAR-10 classification task [18] under the proposed EHFL framework. We set $M = 10$, $K = 5$, and mini-batch size of 50. The nominal learning rate initially sets to 0.15 and decays every 10 rounds with rate 0.99. On top of that, we apply a $c\sqrt{n_t}$ variation such that the mean value for every 10 rounds remain the same as the nominal learning rate. We train a convolutional neural network (CNN) model with two 5×5 convolution layers (both with 64 channels), two fully connected layers (384 and 192 units respectively) with ReLU activation and a final output layer with softmax. The two convolution layers are both followed by 2×2 max pooling and a local response norm layer. In each round, the available clients are generated by the corresponding EH scheduler, and will participate in FL if its available energy is larger than one unit. Otherwise, the client will not participate in FL in the current round. We set $\Lambda = 5$ with a homogeneous arrival rate of all clients for the *Myopic* policy of [9].

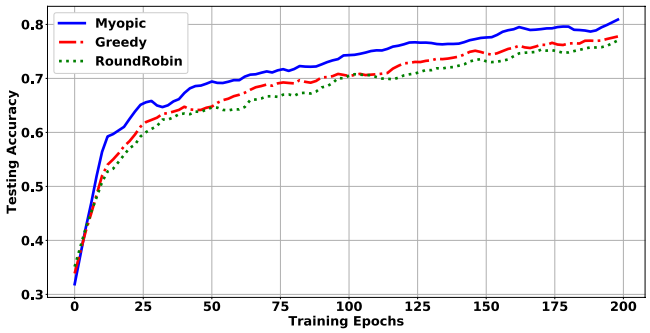

Fig. 2. Model convergence comparison of *Myopic* [9] with two baseline EH schedulers *Round Robin* and *Greedy* for EHFL.

Main result. The model convergence performances of EHFL under three EH schedulers are plotted in Fig. 2. The *Round Robin* policy cyclically schedule among all clients, while the *Greedy* policy always schedule the clients with non-empty energy queues. We can see that the *Myopic* policy has the best performance among the three scheduler, while *Round Robin* has the worst convergence.

5. CONCLUSIONS

We have carried out a novel convergence analysis of federated learning under an arbitrary sequence of participating clients for each learning round, for non-convex loss functions and both parallel and local SGD. The analysis revealed a unified client scheduling principle, which is to maintain a balanced number of clients participating in each round throughout the learning horizon. This result offers a principled guideline for the energy harvesting client scheduler design, and we have shown via a numerical experiment that a state-of-the-art energy harvesting scheduler that follows this guideline achieves better convergence performance for a standard real-world FL task.

6. REFERENCES

[1] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, 2020.

[2] Guangxu Zhu, Dongzhu Liu, Yuqing Du, Changsheng You, Jun Zhang, and Kaibin Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, 2020.

[3] Maria Gorlatova, John Sarik, Guy Grebla, Mina Cong, Ioannis Kymissis, and Gil Zussman, "Movers and shakers: Kinetic energy harvesting for the internet of things," *IEEE J. Select. Areas Commun.*, vol. 33, no. 8, pp. 1624–1639, 2015.

[4] Waleed Ejaz, Muhammad Naeem, Adnan Shahid, Alagan Anpalagan, and Minho Jo, "Efficient energy management for the internet of things in smart cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 84–91, 2017.

[5] Basak Guler and Aylin Yener, "Energy-harvesting distributed machine learning," *arXiv preprint arXiv:2102.05639*, 2021.

[6] Basak Guler and Aylin Yener, "Sustainable federated learning," *arXiv preprint arXiv:2102.11274*, 2021.

[7] Rami Hamdi, Mingzhe Chen, Ahmed Ben Said, Marwa Qaraqe, and H Vincent Poor, "Federated learning over energy harvesting wireless networks," *IEEE Internet Things J.*, 2021.

[8] Jing Yang and Sennur Ulukus, "Optimal packet scheduling in a multiple access channel with energy harvesting transmitters," *Journal of Communications and Networks*, vol. 14, no. 2, pp. 140–150, April 2012.

[9] Jing Yang, Xianwen Wu, and Jingxian Wu, "Optimal scheduling of collaborative sensing in energy harvesting sensor networks," *IEEE J. Select. Areas Commun.*, vol. 33, no. 3, pp. 512–523, March 2015.

[10] Jing Yang, Xianwen Wu, and Jingxian Wu, "Optimal online sensing scheduling for energy harvesting sensors with infinite and finite batteries," *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1578–1589, 2016.

[11] Silas L. Fong, Vincent Y. F. Tan, and Jing Yang, "Non-asymptotic achievable rates for energy-harvesting channels using save-and-transmit," *IEEE J. Select. Areas Commun.*, vol. 34, no. 12, pp. 3499–3511, 2016.

[12] Jing Yang, Omur Ozel, and Sennur Ulukus, "Broadcasting with an energy harvesting rechargeable transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 571–583, February 2012.

[13] Tamoghna Ojha, Sudip Misra, and Narendra Singh Raghuwanshi, "Internet of things for agricultural applications: The state of the art," *IEEE Internet Things J.*, 2021.

[14] Sebastian U Stich, "Local SGD converges fast and communicates little," in *International Conference on Learning Representations*, 2019.

[15] Jianyu Wang and Gauri Joshi, "Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms," in *ICML Workshop on Coding Theory for Machine Learning*, 2019.

[16] Cong Shen, Jing Yang, and Jie Xu, "Online appendix for the technical proofs," `http://www.ece.virginia.edu/~cs7dt/appendix.pdf`, Oct. 2021.

[17] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I. Jordan, "Perturbed iterate analysis for asynchronous stochastic optimization," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2202–2229, 2017.

[18] Alex Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., University of Toronto, April 2009.

[19] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2021.

A. PROOF OF THEOREM 1

Proof: The server model update at the end of round t is

$$x_{t+1} = \frac{1}{n_t}\sum_{i=1}^{n_t} x_{t+1}^i = \frac{1}{n_t}\sum_{i=1}^{n_t}\left(x_t - \eta_t \nabla f_i(x_t)\right) = x_t - \frac{\eta_t}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t).$$

We can evaluate the average loss with respect to model x_{t+1} as

$$\begin{aligned}
\mathbb{E}f(x_{t+1}) &= \mathbb{E}f\left(x_t - \frac{\eta_t}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t)\right) \\
&\leq \mathbb{E}f(x_t) - \frac{\eta_t}{n_t}\mathbb{E}\left\langle \nabla f(x_t), \sum_{i=1}^{n_t}\nabla f_i(x_t)\right\rangle + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t)\right\|^2 \\
&= \mathbb{E}f(x_t) - \eta_t\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t)\right\|^2. \qquad (5)
\end{aligned}$$

We analyze the last term in Eqn. (5), and have

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t) - \nabla f(x_t) + \nabla f(x_t)\right\|^2 \\
&= \mathbb{E}_{x_t}\left[\mathbb{E}_\xi\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t) - \nabla f(x_t) + \nabla f(x_t)\right\|^2 |x_t\right] \\
&= \mathbb{E}_{x_t}\left[\mathbb{E}_\xi\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}(\nabla f_i(x_t) - \nabla f(x_t))\right\|^2 + \|\nabla f(x_t)\|^2 |x_t\right] \\
&\leq \frac{\sigma^2}{n_t} + \mathbb{E}\|\nabla f(x_t)\|^2. \qquad (6)
\end{aligned}$$

Plugging Eqn. (6) back to (5) leads to

$$\begin{aligned}
\mathbb{E}f(x_{t+1}) &\leq \mathbb{E}f(x_t) - \eta_t\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2}\mathbb{E}\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\nabla f_i(x_t)\right\|^2 \\
&\leq \mathbb{E}f(x_t) - \eta_t\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L\eta_t^2}{2}\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L\eta_t^2\sigma^2}{2n_t} \\
&= \mathbb{E}f(x_t) - \left(\eta_t - \frac{L\eta_t^2}{2}\right)\mathbb{E}\|\nabla f(x_t)\|^2 + \frac{L\eta_t^2\sigma^2}{2n_t}. \qquad (7)
\end{aligned}$$

Eqn. (7) is equivalent to

$$\left(\eta_t - \frac{L\eta_t^2}{2}\right)\mathbb{E}\|\nabla f(x_t)\|^2 \leq \mathbb{E}f(x_t) - \mathbb{E}f(x_{t+1}) + \frac{Ln_t^2\sigma^2}{2n_t}, \qquad (8)$$

and we can further sum Eqn. (8) from 0 to $T-1$ and average, resulting in

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\left(\eta_t - \frac{L\eta_t^2}{2}\right)\mathbb{E}\|\nabla f(x_t)\|^2 &\leq \frac{1}{T}\left(f(x_0) - \mathbb{E}f(x_T)\right) + \frac{1}{T}\sum_{t=0}^{T-1}\frac{L\eta_t^2\sigma^2}{2n_t} \\
&\leq \frac{1}{T}\left(f(x_0) - f^*\right) + \frac{L\sigma^2}{2T}\sum_{t=0}^{T-1}\frac{\eta_t^2}{n_t}. \qquad (9)
\end{aligned}$$

The condition $0 < \eta \leq \frac{1}{L}\sqrt{\frac{T}{n_{\max}}}$ implies the function $\eta_t - \frac{L\eta_t^2}{2}$ is both positive and monotonically increasing with η_t. Hence

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left(\eta_{\min} - \frac{L\eta_{\min}^2}{2}\right)\mathbb{E}\|\nabla f(x_t)\|^2 &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left(\eta_t - \frac{L\eta_t^2}{2}\right)\mathbb{E}\|\nabla f(x_t)\|^2 \\ &\leq \frac{1}{T}\left(f(x_0) - f^*\right) + \frac{L\sigma^2\eta^2}{2T} \qquad (10)\end{aligned}$$

where (10) comes from plugging $\eta_t = \eta\sqrt{\frac{n_t}{T}}$ in (9). Dividing the t-independent $\left(\eta_{\min} - \frac{L\eta_{\min}^2}{2}\right)$ from both sides of Eqn. (10) and plugging in $\eta_{\min} = \eta\sqrt{\frac{n_{\min}}{T}}$ complete the proof. ■

B. PROOF OF THEOREM 2

Proof:

Some preparation is necessary to facilitate this proof. First of all, the following lemma from [19, Lemma 3] is useful.

Lemma 1 *For $\eta_t \leq 1/(8KL)$ we have*

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{E}\left\|x_t - x_{t,\tau}^i\right\|^2 \leq 5K\sigma^2\eta_t^2 + 30K^2\eta_t^2\left\|\nabla f(x_t)\right\|^2.$$

Next we define some new variables to simplify the derivation. We denote

$$\begin{aligned}\nabla g_t^i &:= \sum_{\tau=0}^{K-1}\nabla f_i\left(x_{t,\tau}^i\right) \\ \nabla g_t &:= \frac{1}{n_t}\sum_{i=1}^{n_t}\nabla g_t^i = \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\nabla f_i\left(x_{t,\tau}^i\right) \\ \nabla \bar{g}_t &:= \mathbb{E}_{\{\xi_i\}}\nabla g_t = \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\nabla f\left(x_{t,\tau}^i\right)\end{aligned}$$

Proving (4)**.** We start with

$$\begin{aligned}f(x_{t+1}) &= f\left(\frac{1}{n_t}\sum_{i=1}^{n_t}x_{t+1}^i\right) \\ &= f\left(x_t - \eta_t\nabla g_t\right) \\ &\leq f(x_t) - \langle\nabla f(x_t), \eta_t\nabla g_t\rangle + \frac{L}{2}\eta_t^2\left\|\nabla g_t\right\|^2 \\ &= f(x_t) - \eta_t K\left\|\nabla f(x_t)\right\|^2 + \eta_t\left\langle\nabla f(x_t), K\nabla f(x_t) - \nabla g_t\right\rangle + \frac{L\eta_t^2}{2}\left\|\nabla g_t\right\|^2. \qquad (11)\end{aligned}$$

The next steps are to separately analyze the expectation of the last two terms in Eqn. (11). We first have

$$\begin{aligned}
\mathbb{E}\left\langle \nabla f(x_t), K\nabla f(x_t) - \nabla g_t \right\rangle &= \mathbb{E}\left\langle \nabla f(x_t), \frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\left(\nabla f(x_t) - \nabla f(x_{t,\tau}^i)\right)\right\rangle \\
&\overset{(b1)}{=} \frac{K}{2}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 + \frac{1}{2Kn_t^2}\mathbb{E}\left\|\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\left(\nabla f(x_t) - \nabla f(x_{t,\tau}^i)\right)\right\|^2 - \frac{1}{2K}\mathbb{E}\left\|\nabla \bar{g}_t\right\|^2 \\
&\overset{(b2)}{\leq} \frac{K}{2}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 + \frac{1}{2n_t}\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|\nabla f(x_t) - \nabla f(x_{t,\tau}^i)\right\|^2 - \frac{1}{2K}\mathbb{E}\left\|\nabla \bar{g}_t\right\|^2 \\
&\leq \frac{K}{2}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 + \frac{L^2}{2n_t}\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\mathbb{E}\left\|x_t - x_{t,\tau}^i\right\|^2 - \frac{1}{2K}\mathbb{E}\left\|\nabla \bar{g}_t\right\|^2 \\
&\overset{(b3)}{\leq} \frac{K}{2}\left(1 + 30K^2L^2\eta_t^2\right)\mathbb{E}\left\|\nabla f(x_t)\right\|^2 + \frac{5K^2L^2\sigma^2\eta_t^2}{2} - \frac{1}{2K}\mathbb{E}\left\|\nabla \bar{g}_t\right\|^2
\end{aligned} \tag{12}$$

where (b1) is because $\langle x, y\rangle = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x-y\|^2$, (b2) is due to Cauchy-Schwartz, and (b3) is from Lemma 1.

We then evaluate the expectation of the last term of Eqn. (11).

$$\begin{aligned}
\mathbb{E}\left\|\nabla g_t\right\|^2 &= \mathbb{E}\left\|\frac{1}{n_t}\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\nabla f_i(x_{t,\tau}^i)\right\|^2 \\
&\overset{(b4)}{=} \frac{1}{n_t}\mathbb{E}\left\|\sum_{i=1}^{n_t}\sum_{\tau=0}^{K-1}\left(\nabla f(x_{t,\tau}^i) - \nabla f_i(x_{t,\tau}^i)\right)\right\|^2 + \mathbb{E}\left\|\nabla \bar{g}_t\right\|^2 \\
&\leq \frac{K\sigma^2}{n_t} + \mathbb{E}\left\|\nabla \bar{g}_t\right\|^2
\end{aligned} \tag{13}$$

where (b4) uses the fact that the SGD sampling error is independent of other random variables.

Putting both Eqns. (12) and (13) back to the expectation of Eqn. (11), we have

$$\begin{aligned}
\mathbb{E}f(x_{t+1}) &\leq \mathbb{E}f(x_t) - \frac{K}{2}\left(\eta_t - 30K^2L^2\eta_t^3\right)\mathbb{E}\left\|\nabla f(x_t)\right\|^2 + \frac{5K^2L^2\sigma^2\eta_t^3}{2} + \frac{KL\sigma^2\eta_t^2}{2n_t} + \left(\frac{L\eta_t^2}{2} - \frac{\eta_t}{2K}\right)\mathbb{E}\left\|\nabla \bar{g}_t\right\|^2 \\
&\overset{(b5)}{\leq} \mathbb{E}f(x_t) - \frac{K}{2}\left(\eta_t - 30K^2L^2\eta_t^3\right)\mathbb{E}\left\|\nabla f(x_t)\right\|^2 + \frac{5K^2L^2\sigma^2\eta_t^3}{2} + \frac{KL\sigma^2\eta_t^2}{2n_t}
\end{aligned} \tag{14}$$

where (b5) is because for the choice of $\eta \leq \frac{1}{2KL}\sqrt{\frac{1}{30n_{\max}}}$ we can guarantee $\eta_t \leq 1/\left(2\sqrt{30}KL\right) < 1/\left(KL\right)$, and thus

$$\frac{L\eta_t^2}{2} - \frac{\eta_t}{2K} \leq 0.$$

Now, rearranging terms of both sides in Eqn. (14) and averaging over $t = 0$ to $t = T-1$ leads to

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\frac{K}{2}\left(\eta_t - 30K^2L^2\eta_t^3\right)\mathbb{E}\left\|\nabla f(x_t)\right\|^2 &\leq \frac{f(x_0) - \mathbb{E}f(x_T)}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{KL\sigma^2\eta_t^2}{2n_t} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{5K^2L^2\sigma^2\eta_t^3}{2} \\
&\leq \frac{f(x_0) - f^*}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{KL\sigma^2\eta_t^2}{2n_t} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{5K^2L^2\sigma^2\eta_t^3}{2}.
\end{aligned} \tag{15}$$

When $\eta_t \leq 1/\left(2\sqrt{30}KL\right)$, we have

$$\eta_t - 30K^2L^2\eta_t^3 \geq \eta_t\left(1 - \sqrt{30}KL\eta_t\right) \geq \eta_{\min} - \sqrt{30}KL\eta_{\min}^2.$$

Then, Eqn. (15) can be further bounded as

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(x_t)\right\|^2 \quad &\leq \quad \frac{\frac{2}{K}\left(f(x_0)-f^*\right)}{T\left(\eta_{\min}-\sqrt{30}KL\eta_{\min}^2\right)}+\frac{5KL^2\sigma^2}{T\left(\eta_{\min}-\sqrt{30}KL\eta_{\min}^2\right)}\sum_{t=0}^{T-1}\eta_t^3 \\ &\quad +L\sigma^2\frac{1}{T\left(\eta_{\min}-\sqrt{30}KL\eta_{\min}^2\right)}\sum_{t=0}^{T-1}\frac{\eta_t^2}{n_t}. \end{aligned} \tag{16}$$

Plugging in $\eta_t=\eta\sqrt{n_t/T}$, $\eta_{\min}=\eta\sqrt{n_{\min}/T}$, and using

$$\sum_{t=0}^{T-1}\eta_t^3\leq\frac{1}{\sqrt{T}}\eta^3 n_{\max}^{\frac{3}{2}}$$

lead to Eqn. (4), and the proof is complete. ■