ADAPTIVE GRADIENT METHODS FOR OVER-THE-AIR FEDERATED LEARNING

Chenhao Wang[§] *, *Zihan Chen*[†] *, *Howard H. Yang*[§], *and Tony Q. S. Quek*[†]

[§] *ZJU-UIUC Institute, Zhejiang University, Haining 314400, China*
[†] *Singapore University of Technology and Design, Singapore 487372, Singapore*

ABSTRACT

Federated learning (FL) provides a privacy-preserving approach to realizing networked intelligence. But the performance of FL is often constrained by the limited communication resources, especially in the context of a wireless system. To tackle this communication bottleneck, recent studies propose an analog over-the-air (A-OTA) FL paradigm which employs A-OTA computations in the model aggregation step, significantly enhancing system scalability. However, these existing architectures mainly conduct model training via (stochastic) gradient descent, while adaptive optimization methods, which have achieved notable success in deep learning, remain unexplored. In this paper, we establish a distributed training paradigm that incorporates adaptive gradient methods into the A-OTA FL framework, aiming to enhance the system's convergence performance. We derive an analytical expression for the convergence rate, capturing the effects of various system parameters on the convergence performances of the proposed method. We also perform several experiments to validate the efficacy of the proposed method.

Index Terms— Distributed optimization, wireless edge network, over-the-air computing, adaptive gradient descent.

1. INTRODUCTION

Federated learning (FL) is a promising technology that emerged from the intersection of artificial intelligence and edge computing, facilitating networked intelligence while preserving end-users data privacy [1–4]. However, FL requires frequent model parameter exchanges between the end-user devices (a.k.a. clients) and the server, which brings hefty communication overhead. For networks with limited communication resources, such communication bottleneck can throttle the scalability of the FL system and concurrently constrain the training efficiency of the system.

To address this issue, a line of recent studies [5–8] suggest using analog over-the-air (A-OTA) computations in the FL system, exploiting the superposition property of wireless waveforms for fast model aggregation. As a result, the A-OTA computing-based FL system archives significantly improved spectral efficiency, reduced access latency, and enhanced privacy protections for the clients [6, 8, 9].

However, A-OTA computations inevitably introduce distortion to the globally aggregated gradients [10–13]. Specifically, random channel fading and interference (i.e., channel noise) would be induced into the aggregated gradients at the edge server. The consequent noisy aggregated signal would result in performance degradation, such as slower convergence rate and unstable training performance.

To harness such uncertainty while maintaining the benefits of the analog over-the-air computations, numerous work has been proposed to improve the system performance, among which the majority of these approaches are built on FedAVG [1] or FedSGD [13] for gradient descent-based global model updates. But adaptive optimization methods, which are demonstrated with enhanced performance in non-FL setup compared to the standard gradient descent methods, are rarely explored in such FL scenarios. For instance, adaptive methods such as AdaGrad [14] and Adam [15] have been widely used in various fields with significant improvements in training effectiveness and robustness [16–18]. Adaptive algorithms are known to be interference-resistant in deep learning scenarios, which is a characteristic that also applies to A-OTA computing scenarios. Therefore, a natural question arises: *in the presence of noisy aggregated signals, does the adaptive optimization methods still benefit the performance of A-OTA FL (and/or, does such a training method even converge)?*

The present paper answers the above question by developing an A-OTA FL framework that incorporates adaptive gradient descent techniques. To the best of our knowledge, this is the first work that explores the adaptive gradient method for the A-OTA FL setting. Our theoretical analysis provides the upper bound performance for the convergence rate. The numerical results of extensive experiments also confirm the efficacy of our proposed framework.

2. SYSTEM MODEL

Let us consider a wireless edge network, as depicted in Fig. 1, that comprises one server and N clients. The clients communicate with the server through wireless transmissions over a shared spectrum. Every client $n \in \{1, ..., N\}$ has its own lo-

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LGJ22F010001. (*: Equal contribution; *Corresponding Author: Howard H. Yang.*)

cal dataset $\mathcal{D}_n = \left\{ \left(\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R} \right) \right\}_{i=1}^{m_n}$ with size $|\mathcal{D}_n| = m_n$. We assume the local datasets are statistically independent across the clients. The goal of all the entities in this system is to collaboratively learn a statistical model using data from all the clients without violating their data privacy. To be more concrete, the clients need to solve an optimization problem of the following form [1]:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \quad f(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} f_n(\boldsymbol{w}; \mathcal{D}_n), \tag{1}$$

where $f_n(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is the local empirical loss function of client n, constructed from its own dataset $\mathcal{D}_n$ and $\boldsymbol{w} \in \mathbb{R}^d$ is the global model parameter. We denote the optimal solution to (1) by $\boldsymbol{w}^*$, i.e.,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} f(\boldsymbol{w}). \tag{2}$$

To obtain $\boldsymbol{w}^*$ while concurrently preserving the data privacy of each client, the entities need to carry out the model training in an FL fashion. Specifically, the clients shall train their models locally and upload the intermediate gradients to the server. The server aggregates the clients' gradients and further improves the global model. Then, the server broadcasts the model to all the clients for another round of local training. Such interactions will repeat until the global model converges.

Due to the limited spectral resources, the efficiency of such a federated training procedure is often throttled by the communication bottleneck, i.e., the server needs to select a portion of clients for parameter uploading in each communication round, which becomes cumbersome when the number of participants becomes large. In the following section, we introduce a model training framework that addresses this bottleneck using the analog over-the-air computing method. Additionally, it is devised on the basis of the adaptive gradient descent method, which has the potential to accelerate the model training process. Owing to these two attributes, we coin our method as *adaptive over-the-air federated learning (ADOTA-FL)*.

3. ADAPTIVE OVER-THE-AIR FEDERATED LEARNING

This section elaborates on the training process of ADOTA-FL. In particular, this method employs analog over-the-air computations in the global gradient aggregation stage, substantially reducing access latency and facilitating (theoretically unlimited) algorithm scability. In the global model improvement stage, it leverages ideas from AdaGrad methods to potentially speed up convergence. The general steps are summarized in Algorithm 1. And more details are provided below.

1) Local Model Training: Without loss of generality, we assume the system has progressed to the k-th round of global

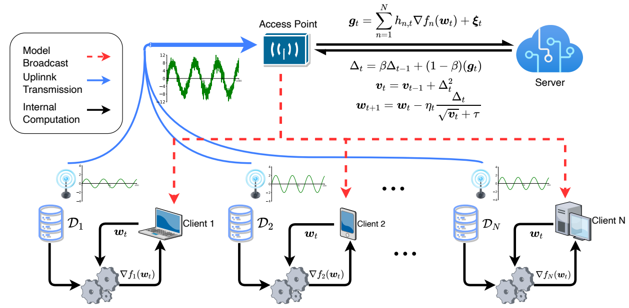

Fig. 1. An overview of the adaptive algorithm for over-the-air federated learning.

training, upon which the clients just received the global model parameters $\boldsymbol{w}_k$ from the edge server.[1] Then, every client n updates its evaluates its local gradient $\nabla f_n(\boldsymbol{w}_k)$ by taking the global model as an input.

2) Analog Gradient Aggregation: In this system, the clients use analog transmissions to upload their locally trained gradients. Specifically, once $\nabla f_n(\boldsymbol{w}_k)$ is computed, client n modulates this gradient vector entry-by-entry onto the magnitudes of a common set of orthogonal baseband waveforms, arriving at the following analog signal

$$x_n(t) = \langle \boldsymbol{s}(t), \nabla f_n(\boldsymbol{w}_k) \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two vectors and $\boldsymbol{s}(t) = (s_1(t), ..., s_d(t)), t \in [0, \tau]$ has its entries satisfying

$$\int_0^\tau s_i^2(t)dt = 1, \;\; i = 1, 2, ..., d, \tag{4}$$

$$\int_0^\tau s_i(t)s_j(t)dt = 0, \;\; i \neq j. \tag{5}$$

Once the analog waveforms $\{x_n(t)\}_{n=1}^N$ have been assembled, the clients transmit them simultaneously to the edge server. Thanks to the superposition property of electromagnetic waves, the signal received at the edge server obey the following form:

$$y(t) = \sum_{n=1}^{N} h_{t,n} P_n x_n(t) + \xi(t) \tag{6}$$

where $h_{t,n}$ is the channel fading experienced by client n, P_n is the corresponding transmit power, and $\xi(t)$ denotes the additive noise. In this work, we assume the channel fading is i.i.d. across clients, with mean and variance being μ_{c} and σ_{c}^2, respectively. Besides, the transmit power of each client is set to compensate for the large-scale path loss. Additionally, we assume the noise follows a Gaussian distribution with variance σ_{n}^2. This received signal will be passed through a bank of

[1]Since the server can broadcast its signal at a high transmit power, we assume the global model can be successfully received by all the clients.

match filters, with each branch tuning to $s_i(t), i = 1, 2, .., d$. On the output side, the server obtains the following vector:

$$\boldsymbol{g}_t = \frac{1}{N}\sum_{n=1}^{N} h_{t,n}\nabla f_n(\boldsymbol{w}_t) + \boldsymbol{\xi}_t, \tag{7}$$

in which $\boldsymbol{\xi}_t$ is a d-dimensional random vector with each entry being i.i.d. and follows the Gaussian distribution. It is worthwhile to stress that the vector given in (7) is a distorted version of the globally aggregated gradient.

3) Global Model Update: Using $\boldsymbol{g}_t$, the server can update the global model via adaptive optimization methods. However, due to channel fading and interference noise perturbations, the globally aggregated gradient may experience a significant distortion, and the general adaptive method may not perform well. To abbreviate the impact of such distortions, we store and update an intermediate global model as follows:

$$\Delta_t = \beta\Delta_{t-1} + (1-\beta)\,\boldsymbol{g}_t, \tag{8}$$

where β is a controlling factor that adjusts the portion between the historical information and the newly acquired information. Notably, the operation (8) introduces a momentum-like approach to smooth out the fluctuation in the aggregated gradient. As such, a smaller value of β leads to a faster convergence but also incurs stronger volatility.

Aided by Δ_t, we construct a vector $\boldsymbol{v}_t$ as follows:

$$\boldsymbol{v}_t = \boldsymbol{v}_{t-1} + \Delta_t^2, \tag{9}$$

where in Δ_t^2, the square is performed entry-wise. The vector $\boldsymbol{v}_t$ is used as the updated global model's denominator, where it adjusts the learning rate corresponding to each entry in the global model.

Finally, using $\boldsymbol{v}_t$ and Δ_t, we update the global model as:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t\frac{\Delta_t}{\sqrt{\boldsymbol{v}_t}+\tau}, \tag{10}$$

where η_t is the learning rate factor, and τ is a constant that prevents ill-conditioning. It is important to note that the operation $\sqrt{\boldsymbol{v}_t}$ is taken with respect to each entry of $\boldsymbol{v}_t$.

The new global model $\boldsymbol{w}_{t+1}$ will be broadcasted to all the clients for the next round of local computing. And the clients and server will repeat this process for multiple rounds until the global model converges.

4. ANALYSIS

In this section, we analyze the convergence rate of the proposed ADOTA-FL method. We assume each client's local objective function $f_n(\cdot)$ is convex. As such, the global objective function $f(\cdot)$ is also convex. We denote by $\boldsymbol{g}_{t,n} = h_{t,n}\nabla f_n(\boldsymbol{w}_t) + \boldsymbol{\xi}_t$ as the noisy gradient of client n in the t-th global iteration.

The following assumptions are adopted in our proof.

Algorithm 1 ADOTA-FL

Input: Initial delay vector $\boldsymbol{v}_{-1}$, initial global model $\boldsymbol{w}_0$, T, η_0

for $t = 0, 1, 2, ..., T-1$ **do**
  for each client $n \in N$ **in parallel do**
    # Train model locally and upload gradients
    $\nabla f_n(\boldsymbol{w}_t) \leftarrow \text{CLIENTUPDATE}(t, \boldsymbol{w}_t)$
  $\boldsymbol{g}_t = \frac{1}{N}\sum_{n=1}^{N} h_{t,n}\nabla f_n(\boldsymbol{w}_t) + \boldsymbol{\xi}_t$
  $\Delta_t = \beta\Delta_{t-1} + (1-\beta)\,\boldsymbol{g}_t$
  $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} + \Delta_t^2$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t\frac{\Delta_t}{\sqrt{\boldsymbol{v}_t}+\tau}$

Output: w_T

Assumption 1. *There exists a constant C such that $\|\boldsymbol{w}_t - \boldsymbol{w}^*\|_2 \le C$ for any $t \in \{1, \ldots, T\}$.*

Assumption 2. *There exists a constant D such that $\|\frac{\boldsymbol{g}_{t,n}}{\sqrt{\boldsymbol{v}_t}+\tau}\|_2 \le D$ for any $t \in \{1, \ldots, T\}$ and $n \in \{1, \ldots, N\}$.*

For any verctor $\boldsymbol{a}$, $\boldsymbol{a}^j$ denotes the j-th coordinate of $\boldsymbol{a}$. For example, $\boldsymbol{g}_t^j$ means the j-th coordinate of gradient $\boldsymbol{g}_t$, and $\boldsymbol{v}_t^j$ means the j-th coordinate of gradient $\boldsymbol{v}_t$. For the ease of expository, we assume $\beta = 0$, though our analysis can be directly extended to $\beta > 0$. For details on the proof, see Appendix A.1.

Lemma 1. *Let Assumptions 1, 2 hold, then we can get the auxiliary lemma by induction:*

$$\sum_{t=1}^{T}\mathbb{E}\left[\frac{\left(\boldsymbol{g}_t^j\right)^2}{\sqrt{t\boldsymbol{v}_t^j}+\sqrt{t}\tau}\right] \le \mathbb{E}\left[\sqrt{T\boldsymbol{v}_T^j}+\sqrt{T}\tau\right]. \tag{11}$$

In order to reflect the effects caused by channel fading and interference in the final results, we consider an auxiliary factor. We denote the auxiliary factor as

$$\tilde{\boldsymbol{v}}_t = \tilde{\boldsymbol{v}}_{t-1} + (\nabla f(\boldsymbol{w}_t))^2, \tag{12}$$

so $\tilde{\boldsymbol{v}}_T^j$ means the j-th coordinate of $\tilde{\boldsymbol{v}}_T$. Then we introduce this factor into the final result

Theorem 1. *Let assumptions 1 and 2 hold. After running Algorithm 1 for T iterations with initial learning rate η_1, we have*

$$\min_{t\in\{1,\ldots,T\}}\mathbb{E}[f(\boldsymbol{w}_t)] - f^* \le \frac{G}{\sqrt{T}}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{(\mu_c^2+\sigma_c^2)\,\tilde{\boldsymbol{v}}_T^j+\sigma_n}+\tau\right], \tag{13}$$

where $f^* := f(\boldsymbol{w}^*)$ and $G = \frac{D^2}{2\eta_1} + \frac{\eta_1}{2}$.

Proof. We omit the proof due to space limit. For details on the proof, see Appendix A.1. □

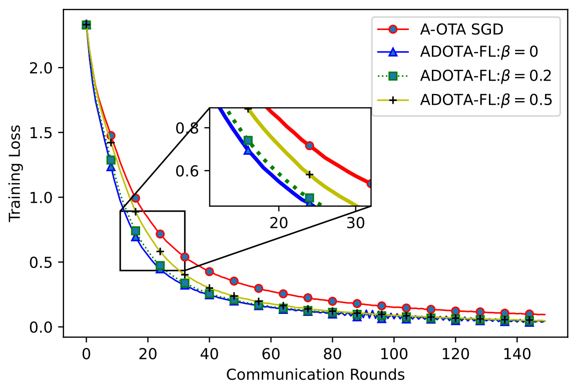

Fig. 2. Performance comparison for training loss on CIFAR-10 with IID data

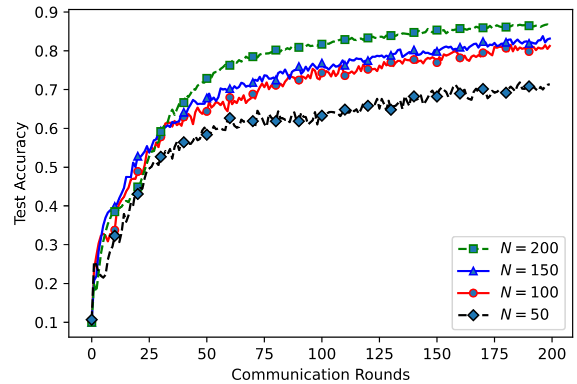

Fig. 3. Performance comparison for test accuracy with different total number of clients N on CIFAR-10 with IID data.

ADOTA-FL's upper bound is provided by Theorem 1, which suggests that the algorithm is convergent. The upper bound indicates that channel fading and interference work together to affect the algorithm's convergence.

5. NUMERICAL RESULT

We conduct several experiments in this section to examine the performance of ADOTA-FL. We explore the impacts of parameters pertaining to the algorithm (e.g., hyperparameters) and system (e.g., the number of participating clients in the network) on the performance of the A-OTA federated learning framework. More details are given in the sequel.

5.1. Experimental setup

To evaluate the performance of the proposed framework, we carry out the experiments of image classification tasks on the CIFAR-10 [19] with ResNet-18 [20]. To simulate the perturbations introduced by analog transmissions, we modify the gradient aggregation process to include multiplicative fading effects that follow the Rayleigh distribution and additive noise with the zero-mean Gaussian distribution. All the reported results are averaged with 3 trials. Unless otherwise specified, we use the total number of clients $N = 100$. Both the IID and non-IID data settings are evaluated, in which non-IID data are partition with widely used Dirichlet distribution with parameter $\alpha = 0.1$.

5.2. Performance evaluation

Convergence performance: Fig. 2 compares the training loss performances under A-OTA SGD [13] and ADOTA-FL, where the impacts of different controlling factors β are also investigated. It can be observed that ADOTA-FL attains a faster convergence rate than the A-OTA SGD for a varying value of β, verifying the effectiveness of employing adaptive optimization methods in the framework of A-OTA FL systems. Moreover, we notice that the convergence speed of ADOTA-FL is slower but more stable for higher β values. Conversely, as the β value decreases, the convergence speed increases, but the fluctuation rises as well. This observation is consistent with what we expected because β is related to historical information, and the larger the β, the more information will be retained during the model update process, which will result in slower convergence but more stable convergence overall.

Performance with different system scale: In Fig. 3, we depict the test accuracy under ADOTA-FL as a function of communication rounds under different values of client number N. This figure unveils a phenomenon unique to A-OTA FL systems, that an increase in the number of participating clients can improve the system's performance [13]. This is because, using A-OTA computing, all the clients can simultaneously upload their locally computed gradients in each communication round. In contrast to the conventionally adopted digital communication-based FL that requires scheduling [3], it significantly boosts up the amount of information aggregated in every global iteration, thus enhancing the generalization performance. To evaluate the performance in the presence of data heterogeneity, we also compare the performance with non-IID CIFAR-10 settings. The best test accuracies corresponding to $N = 200, 100$, and 50 achieve 0.8092 ± 0.006, 0.7831 ± 0.007, and 0.7491 ± 0.03 respectively, where a larger N reveals a positive influence on the system performance.

6. CONCLUSION

In this paper, we proposed ADOTA-FL, which uses an adaptive optimization method to improve the convergence performance of OTA computing-based FL systems. We derived the convergence rate of ADOTA-FL, accounting for effects from both agorithmic and system perspectives. We also carried out several experiments to demonstrate the efficacy of the proposed method and explored the effects of different system parameters on the ADOTA-FL performance. The robustness performance would be further explored in our future work.

7. REFERENCES

[1] D. Ramage S. Hampson H. B. McMahan, E. Moore and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[2] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[3] Howard H. Yang, Zuozhu Liu, Tony Q. S. Quek, and H. Vincent Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, 2020.

[4] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[5] Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.

[6] Guangxu Zhu, Jie Xu, and Kaibin Huang, "Over-the-air computing for 6g - turning air into a computer," *CoRR*, vol. abs/2009.02181, 2020.

[7] Mohammad Mohammadi Amiri and Deniz Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, Mar. 2020.

[8] Huayan Guo, Yifan Zhu, Haoyu Ma, Vincent KN Lau, Kaibin Huang, Xiaofan Li, Huabin Nong, and Mingyu Zhou, "Over-the-air aggregation for federated learning: Waveform superposition and prototype validation," *J. of Commun. and Inf. Networks*, vol. 6, no. 4, pp. 429–442, 2021.

[9] Guangxu Zhu, Yong Wang, and Kaibin Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[10] Howard H Yang, Zihan Chen, Tony Q.S. Quek, and H Vincent Poor, "Revisiting analog over-the-air machine learning: The blessing and curse of interference," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 406–419, 2021.

[11] Zezhong Zhang, Guangxu Zhu, Rui Wang, Vincent K. N. Lau, and Kaibin Huang, "Turning channel noise into an accelerator for over-the-air principal component analysis," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 7926–7941, 2022.

[12] Guangxu Zhu, Jie Xu, Kaibin Huang, and Shuguang Cui, "Over-the-air computing for wireless data aggregation in massive iot," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 57–65, 2021.

[13] Tomer Sery and Kobi Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process*, vol. 68, pp. 2897–2911, 2020.

[14] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research (JMLR)*, vol. 12, no. 61, pp. 2121–2159, 2011.

[15] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[16] Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan, "Adaptive federated optimization," in *International Conference on Learning Representations*, 2020.

[17] Rachel Ward, Xiaoxia Wu, and Leon Bottou, "Adagrad stepsizes: Sharp convergence over nonconvex landscapes," *The Journal of Machine Learning Research (JMLR)*, vol. 21, no. 1, pp. 9047–9076, 2020.

[18] Smit Mehta, Chirag Paunwala, and Bhaumik Vaidya, "Cnn based traffic sign classification using adam optimizer," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. IEEE, 2019, pp. 1293–1298.

[19] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," Tech. Rep. 0, University of Toronto, Toronto, Ontario, 2009.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

A. APPENDIX

A.1. Proof of Lemma 1

We prove the following inequality by induction:

$$\sum_{t=1}^{T} \mathbb{E}\left[\frac{\left(g_j^t\right)^2}{\sqrt{tv_j^t}+\sqrt{t}\epsilon^t}\right] \leq \mathbb{E}\left[\sqrt{Tv_j^T}+\sqrt{T}\epsilon^T\right] \tag{14}$$

For $t=1$

$$\begin{aligned} &\mathbb{E}\left[\frac{\left(g_j^1\right)^2}{\sqrt{v_j^1}+\epsilon}\right] = \mathbb{E}\left[\frac{\left(g_j^1\right)^2}{\sqrt{\left(g_j^1\right)^2}+\epsilon}\right] \\ &\leq \mathbb{E}\left[\sqrt{\left(g_j^1\right)^2+\epsilon}\right] \leq 2\mathbb{E}\left[\sqrt{\left(g_j^1\right)^2+\epsilon}\right] \end{aligned} \tag{15}$$

Suppose that the conclusion holds when $t=T-1$, i.e., for any $j \in [d]$,

$$\sum_{t=1}^{T-1} \mathbb{E}\left[\frac{\left(g_j^t\right)^2}{\sqrt{tv_j^t}+\sqrt{t}\epsilon^t}\right] \leq \mathbb{E}\left[\sqrt{(T-1)v_j^{T-1}}+\sqrt{T-1}\epsilon\right]. \tag{16}$$

In addition, combined with the fact that $v_j^T = v_j^{T-1} + \left(g_j^T\right)^2$ and $\sqrt{T}\epsilon \geq \sqrt{T-1}\epsilon^{T-1}$, we have

$$\begin{aligned} &\sqrt{(T-1)v_j^{T-1}}+\sqrt{T-1}\epsilon \leq \\ &\sqrt{(T-1)v_j^T-(T-1)\left(g_j^T\right)^2}+\sqrt{T}\epsilon \end{aligned} \tag{17}$$

Using $\sqrt{x-c} \leq \sqrt{x} - \frac{c}{2\sqrt{x}} (x>c)$, we have

$$\begin{aligned} &\sqrt{(T-1)v_j^T-(T-1)\left(g_j^T\right)^2}+\sqrt{T}\epsilon \\ &\leq \sqrt{(T-1)v_j^T} - \frac{(T-1)\left(g_j^T\right)^2}{2\sqrt{(T-1)v_j^T}}+\sqrt{T}\epsilon \\ &\leq \sqrt{Tv_j^T} - \frac{(T-1)\left(g_j^T\right)^2}{2\left(\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right)}+\sqrt{T}\epsilon \end{aligned} \tag{18}$$

Hence

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{E}\left[\frac{\left(g_j^t\right)^2}{\sqrt{tv_j^t}+\sqrt{t}\epsilon}\right] \\ &\leq \mathbb{E}\left[\sqrt{Tv_j^T} - \frac{(T-1)\left(g_j^T\right)^2}{2\left(\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right)}+\sqrt{T}\epsilon\right] \\ &+\mathbb{E}\left[\frac{\left(g_j^T\right)^2}{\sqrt{Tv_j^T}+\sqrt{T}\epsilon}\right] \\ &\leq \mathbb{E}\left[\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right] + \left(1-\frac{T-1}{2}\right)\mathbb{E}\left[\frac{\left(g_j^T\right)^2}{\sqrt{Tv_j^T}+\sqrt{T}\epsilon}\right] \\ &\leq \mathbb{E}\left[\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right] \end{aligned} \tag{19}$$

A.2. Proof of Theorem 1

We employ the norm as

$$\|\cdot\|_M = \sqrt{\langle \cdot, M\cdot\rangle}$$

for symmeric and positive definite $M \in \mathbb{R}^d$. If $M \in \mathbb{R}^d$ and has entries being non-negative, then diagonalize it.

To simplify the notation, we denote $\sqrt{\nu_t}+\varepsilon_t$ as A_t, and $\eta^t = \frac{\eta}{\sqrt{t}}$, which means η^t is a decay factor

Based on the updating rule, we have

$$\begin{aligned} &\left\|w^{t+1}-w^*\right\|_{A^t}^2 \\ &=\left\|w^t-\eta^t\left(A^t\right)^{-1}g^t-w^*\right\|_{A^t}^2 \\ &=\left\|w^t-w^*\right\|_{A^t}^2+\left\|\eta^t\left(A^t\right)^{-1}g^t\right\|_{A^t}^2-2\left\langle w^t-w^*,\eta^t g^t\right\rangle \\ &=\left\|w^t-w^*\right\|_{A^t}^2-2\eta^t\left\langle g^t,w^t-w^*\right\rangle+\left(\eta^t\right)^2\left\langle g^t,\left(A^t\right)^{-1}g^t\right\rangle \end{aligned} \tag{20}$$

Rearranging terms gives

$$\begin{aligned} \left\langle g^t,w^t-w^*\right\rangle &= \frac{\left\|w^t-w^*\right\|_{A^t}^2-\left\|w^{t+1}-w^*\right\|_{A^t}^2}{2\eta^t} \\ &+\frac{\eta^t}{2}\left\langle g^t,\left(A^t\right)^{-1}g^t\right\rangle. \end{aligned} \tag{21}$$

Take expectation on both sides conditioned on w^t

$$\begin{aligned} \left\langle \nabla F\left(w^t\right),w^t-w^*\right\rangle &= \frac{\mathbb{E}_t\left[\left\|w^t-w^*\right\|_{A^t}^2\right]-\mathbb{E}_t\left[\left\|w^{t+1}-w^*\right\|_{A^t}^2\right]}{2\eta^t} \\ &+\frac{\eta^t}{2}\mathbb{E}_t\left[\left\langle g^t,\left(A^t\right)^{-1}g^t\right\rangle\right], \end{aligned} \tag{22}$$

where we have used the fact that N is a zero-mean Gaussian variable independent of g^t, w^t. Taking expectation on both sides and using the convexity of $F(\cdot)$:

$$\begin{aligned} &\mathbb{E}\left[F\left(w^t\right)\right]-F\left(w^*\right) \\ &\leq \frac{\mathbb{E}\left[\left\|w^t-w^*\right\|_{A^t}^2\right]-\mathbb{E}\left[\left\|w^{t+1}-w^*\right\|_{A^t}^2\right]}{2\eta^t} \\ &+\frac{\eta^t}{2}\mathbb{E}\left[\left\langle g^t,\left(A^t\right)^{-1}g^t\right\rangle\right] \end{aligned} \tag{23}$$

Applying telescope sum, we have

$$\begin{aligned} &\sum_{t=1}^{T}\left(\mathbb{E}\left[F\left(w^t\right)\right]-F\left(w^*\right)\right) \leq \frac{\left\|w^1-w^*\right\|_{A^1}^2}{2\eta_1} \\ &+\sum_{t=2}^{T}\left(\frac{\mathbb{E}\left[\left\|w^t-w^*\right\|_{A^t}^2\right]}{2\eta^t}-\frac{\mathbb{E}\left[\left\|w^t-w^*\right\|_{A^{t-1}}^2\right]}{2\eta^{t-1}}\right) \\ &+\sum_{t=1}^{T}\frac{\eta^t}{2}\mathbb{E}\left[\left\langle g^t,\left(A^t\right)^{-1}g^t\right\rangle\right] \end{aligned} \tag{24}$$

Using $\eta^t=\frac{\eta}{\sqrt{t}}$, we now have

$$\begin{aligned} &\sum_{t=1}^{T}\left(\mathbb{E}\left[F\left(w^t\right)\right]-F\left(w^*\right)\right) \\ &\leq \frac{\left\|w^1-w^*\right\|_{A^1}^2}{2\eta}+\underbrace{\sum_{t=2}^{T}\frac{\mathbb{E}\left[\left\|w^t-w^*\right\|_{\sqrt{t}A^t-\sqrt{t-1}A^{t-1}}^2\right]}{2\eta}}_{T_1}+ \\ &\underbrace{\sum_{t=1}^{T}\frac{\eta}{2\sqrt{t}}\mathbb{E}\left[\left\langle g^t,\left(A^t\right)^{-1}g^t\right\rangle\right]}_{T_2} \end{aligned} \tag{25}$$

We first bound T_1. Based on the relations between v^t and v^{t-1} , apparently for any j,t we have

$$\sqrt{t}A_j^t \geq \sqrt{t-1}A_j^{t-1}. \tag{26}$$

Hence,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=2}^{T}\left\|w^t-w^*\right\|_{\sqrt{t}A^t-\sqrt{t-1}A^{t-1}}^2\right] \\ &=\mathbb{E}\left[\sum_{t=2}^{T}\sum_{j=1}^{d}\left(w_j^t-w_j^*\right)^2\left(\sqrt{tv_j^t}+\sqrt{t}\epsilon-\sqrt{(t-1)v_j^{t-1}}-\sqrt{t-1}\epsilon\right)\right] \\ &=\mathbb{E}\left[\sum_{j=1}^{d}\sum_{t=2}^{T}\left(w_j^t-w_j^*\right)^2\left(\sqrt{tv_j^t}+\sqrt{t}\epsilon-\sqrt{(t-1)v_j^{t-1}}-\sqrt{t-1}\epsilon\right)\right] \\ &\leq \mathbb{E}\left[\sum_{j=1}^{d}D^2\sum_{t=2}^{T}\left(\sqrt{tv_j^t}+\sqrt{t}\epsilon-\sqrt{(t-1)v_j^{t-1}}-\sqrt{t-1}\epsilon\right)\right] \\ &=\mathbb{E}\left[\sum_{j=1}^{d}D^2\left(\sqrt{Tv_j^T}+\sqrt{T}\epsilon-\sqrt{v_j^1}-\epsilon\right)\right] \end{aligned} \tag{27}$$

We next bound T_2. According to Lemma 1, we have

$$\sum_{t=1}^{T}\mathbb{E}\left[\frac{\left(g_j^t\right)^2}{\sqrt{tv_j^t+\sqrt{t}\epsilon^t}}\right] \leq \mathbb{E}\left[\sqrt{Tv_j^T+\sqrt{T}\epsilon^T}\right] \tag{28}$$

so we can bound T_2 as follows.

$$\begin{aligned} T_2 &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta^t}{2}\sum_{j=1}^{d}\frac{\left(g_j^t\right)^2}{\sqrt{v_j^t}+\epsilon}\right]=\frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{d}\frac{\left(g_j^t\right)^2}{\sqrt{tv_j^t}+\sqrt{t}\epsilon}\right] \\ &\leq \frac{\eta}{2}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right]. \end{aligned} \tag{29}$$

Noting that

$$\frac{\left\|w^1-w^*\right\|_{A^1}^2}{2\eta} \leq \frac{D^2}{2\eta}\sum_{j=1}^{d}\left(\sqrt{v_j^1}+\epsilon\right), \tag{30}$$

combined with the bounds of T_1, T_2 yields

$$\sum_{t=1}^{T}\left(\mathbb{E}\left[F\left(w^t\right)-F\left(w^*\right)\right]\right) \leq \left(\frac{D^2}{2\eta}+\frac{\eta}{2}\right)\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right], \tag{31}$$

which implies that

$$\begin{aligned} \min_{t\in[T]}\mathbb{E}\left[F\left(w^t\right)\right]-F\left(w^*\right) &\leq \left(\frac{D^2}{2\eta}+\frac{\eta}{2}\right)\frac{1}{T}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{Tv_j^T}+\sqrt{T}\epsilon\right] \\ &\leq \left(\frac{D^2}{2\eta}+\frac{\eta}{2}\right)\frac{1}{\sqrt{T}}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{v_j^T}+\epsilon\right] \end{aligned} \tag{32}$$

Further note that

$$\begin{aligned} v_T &= v_{T-1}+g_T^2 \\ &= v_{T-2}+g_{T-1}^2+g_T^2 \quad \ldots \end{aligned} \tag{33}$$

Since

$$g_k=\frac{1}{N}\sum_{n=1}^{N}h_{n,k}\cdot\nabla g_n\left(\omega_k\right)+\xi_k \tag{34}$$

We average out randomness from fading and noise from $\mathbb{E}\left[\sqrt{Tv_j^T}\right]$

since $f(\hat{x})=\sqrt{x}$ is convex, we have

$$\mathbb{E}_{h,\xi}\left[\sqrt{Tv_T}\right] \leqslant \mathbb{E}\left[\sqrt{T\mathbb{E}_{h,\xi}\left(v_T\right)}\right] \tag{35}$$

So

$$\begin{aligned} \mathbb{E}\left[\left(g_k^j\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}h_{n,k}\left(\nabla f_n\left(w_k\right)\right)_j+\xi_k^j\right)^2\right] \\ &\leq \left(\mu_c^2+\sigma_c^2\right)\left(\frac{1}{N}\sum_{n=1}^{N}\nabla F(w_k)\right)_j^2+\sigma_n^2 \\ &= \left(\mu_c^2+\sigma_c^2\right)\left(\nabla F(w_k)\right)_j^2+\sigma_n^2 \end{aligned} \tag{36}$$

We denote an auxiliary factor as

$$\tilde{v}_t = \tilde{v}_{t-1} + \left(\nabla f\left(w_t\right)\right)^2 \tag{37}$$

where $f(\omega) = \frac{1}{N}\sum_{n=1}^{N} f_n(w)$. Then

$$\begin{aligned}
&\min_{t\in[T]} \mathbb{E}\left[F\left(w^t\right)\right] - F\left(w^*\right) \\
&\leq \left(\frac{D^2}{2\eta} + \frac{\eta}{2}\right)\frac{1}{T}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{T v_j^T} + \sqrt{T}\epsilon\right] \\
&\leq \left(\frac{D^2}{2\eta} + \frac{\eta}{2}\right)\frac{1}{\sqrt{T}}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\left(\mu_c^2 + \sigma_c^2\right)\tilde{v}_j^T} + \epsilon\right]
\end{aligned} \tag{38}$$