On the Necessity of Aligning Gradients for Wireless Federated Learning

Wei-Ting Chang Mohamed Seif Eldin Mohamed Ravi Tandon
Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, 85721
Email: {*wchang, mseif, tandonr*}@email.arizona.edu

***Abstract*—In this paper, we consider the problem of wireless federated learning, where the users wish to jointly train a machine learning model with the help of a parameter server. During the training, the local gradients from the users are aggregated over a wireless channel. Typically, coefficients of the local gradients are aligned by power control techniques to ensure that the estimated aggregated gradient is an unbiased estimator of the true gradient. However, schemes that align gradients require coordination, can be challenging to implement in practice, and often lead to degraded performance due to heterogeneity of users' channels. In this paper, we show that alignment of gradients for wireless FL is not always necessary for convergence. Specifically, we consider non-convex loss functions, and derive conditions under which misaligned wireless gradient aggregation still converges to a stationary point. We also present experimental results to show that transmitting at full power can outperform aligned gradient aggregation depending on the heterogeneity of users' channels.**

***Index Terms*—Federated learning, Wireless Aggregation, Stochastic Gradient Descent.**

I. Introduction

There has been a significant recent interest on the topic of wireless federated learning (FL), in part due to a) the increase of computational capabilities of mobile devices and b) the fact that superposition property of wireless channel can naturally facilitate bandwidth efficient aggregation of gradients/models over the air. Several works [1]–[7]) under the umbrella of wireless FL have been proposed.

In this paper, we consider the Federated stochastic gradient descent (FedSGD) algorithm for FL, where in each training iteration, the local gradients from the participating users need to be aggregated for model updates. We focus on wireless analog aggregation of gradients for FedSGD, in which the local gradient of each user is rescaled (to satisfy power constraints and/or mitigate channel impairments). The rescaled gradients are then transmitted directly over the air by all users simultaneously. Since no error-control codes is used, the superposition nature of the wireless medium aggregates the gradients from users on the fly, which makes analog schemes more bandwidth efficient compared to digital ones. There have been several recent works (e.g., [3], [4], [8]–[10]) focusing on the design of analog schemes for wireless FL that hinge on aligning the gradients. More specifically, these schemes requires two types of alignments: 1) temporal alignment; that requires strict time synchronization between users and 2) power alignment; where it requires accurate channel gain for pre-equalization. For the former, practical synchronization mechanisms (e.g., timing advance [6]) can be adopted for aggregation with negligible error. For the latter, power alignment schemes can be challenging for two reasons: 1) it may not be feasible to employ channel pre-equalization for some users due to power constraints, and 2) it assumes perfect local channel state information (CSI) which can be challenging in practice.

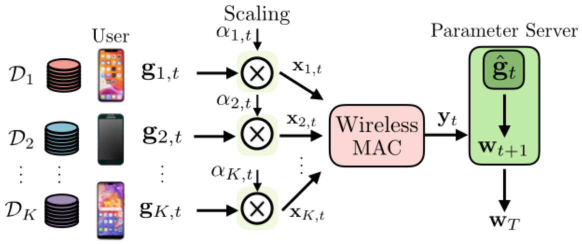

Fig. 1. Illustration of the wireless FedSGD framework: Users collaborate with the PS to jointly train a machine learning model over a fading MAC.

To tackle the issue of power alignment, several power control strategies have recently been proposed: a truncation-based strategy was proposed in [3] for improving the signal-to-noise ratio (SNR) for unbiased gradient aggregation, [4], [11] proposed to jointly design users' transmit powers and a denoising factor at the PS to minimize the mean-squared-error (MSE) between the estimated gradient and the true gradient at each iteration. The main concern in these works is that they do not provide theoretical guarantees on the convergence rates on their FL algorithms. Specifically, the underlying analysis is only tailored for minimizing the MSE and does not guarantee convergence of the learning algorithms in general.

Main Contributions: In this work, we study the problem of wireless FL for smooth, non-convex loss functions with a goal of answering the following question: *is alignment of gradients for analog aggregation necessary for convergence?* For non-convex losses, we first derive the convergence rate of channel inversion based alignment scheme to a stationary point. While for this scheme, the gradient estimate is unbiased, the convergence rate is limited by the worst SNR across all users. We then study the full power transmission scheme (in which all users transmit the gradients at full power), and obtain the resulting convergence rate of this scheme. When the local

This work has been supported in part by NSF Grants CAREER 1651492, CNS 1715947, CCF 2100013 and the 2018 Keysight Early Career Professor Award.

gradients are assumed to be i.i.d., we prove that the full power scheme always outperforms the alignment scheme. We also study a variant of the full power scheme in which the gradient estimate may be biased, and we show that a careful choice of the bias can further improve the convergence rate. Through our convergence results and experiments, we show that when SNRs across users vary significantly, full power transmission schemes (both unbiased and biased ones) can outperform alignment based schemes.

II. System Model

Wireless Channel Model: We consider a single-antenna Wireless FL with n users and a central PS. Users are connected to the PS through a fading MAC as shown in Fig. 1. The input-output relationship at the t-th block is

$$\mathbf{y}_t = \sum_{k=1}^{K} h_{k,t}\mathbf{x}_{k,t} + \mathbf{n}_t, \tag{1}$$

where $\mathbf{x}_{k,t} \in \mathbb{R}^d$ is transmitted signal by user k at the t-th block, and $\mathbf{y}_t$ is the received signal at the PS. Here, $h_{k,t} \in \mathbb{R}$ is the channel coefficient between the k-th user and the PS at iteration t. We assume a block flat-fading channel, where channel coefficients remain constant within the duration of a communication block. Each user is assumed to know its local channel gain, whereas we assume that the PS has global channel state information. Each user can transmit subject to average power constraint i.e., $\mathbb{E}\left[\|\mathbf{x}_{k,t}\|_2^2\right] \le P_k$. $\mathbf{n}_t \in \mathbb{R}^d$ is the channel noise whose elements are independent and identically distributed (i.i.d.) according to $\mathcal{N}(0, N_0)$.

Federated Learning Problem: Each user k has a private local dataset $\mathcal{D}_k$ with D_k data points, denoted as $\mathcal{D}_k = \{(\mathbf{u}_i^{(k)}, v_i^{(k)})\}_{i=1}^{D_k}$, where $\mathbf{u}_i^{(k)}$ is the i-th data point and $v_i^{(k)}$ is the corresponding label at user k. The local loss function at user k is given by $f_k(\mathbf{w}) = \frac{1}{D_k}\sum_{i=1}^{D_k} f(\mathbf{w}; \mathbf{u}_i^{(k)}, v_i^{(k)}) + \Omega R(\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^d$ is the parameter vector to be optimized, $R(\mathbf{w})$ is a regularization function and $\Omega \ge 0$ is a regularization hyperparameter. Users communicate with the PS through the fading MAC described above in order to train a model by minimizing the loss function $F(\mathbf{w})$, i.e., $\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w}) \triangleq \frac{1}{\sum_{k=1}^{K} D_k}\sum_{k=1}^{K} D_k f_k(\mathbf{w})$. The minimization of $F(\mathbf{w})$ is carried out iteratively through a distributed stochastic gradient descent (SGD) algorithm. More specifically, in the t-th training iteration, the PS broadcasts the global parameter vector $\mathbf{w}_t$ to all users. Each user k computes his local gradient using stochastic mini batch $\mathcal{B}_k \subseteq \mathcal{D}_k$, with size b_k (i.e., $|\mathcal{B}_k| = b_k$), i.e.,

$$\mathbf{g}_{k,t} = \frac{1}{b_k}\sum_{i \in \mathcal{B}_k} \nabla f_k(\mathbf{w}_t; (\mathbf{u}_i^{(k)}, v_i^{(k)})) + \Omega \nabla R(\mathbf{w}_t), \tag{2}$$

where $\mathbf{g}_{k,t}$ is the stochastic gradient estimate of user k. Upon receiving $\mathbf{y}_t$, the PS performs post-processing on $\mathbf{y}_t$ to obtain $\hat{\mathbf{g}}_t$, the estimate of the true gradient $\mathbf{g}_t$ which is defined as,

$$\mathbf{g}_t = \frac{1}{\sum_{k=1}^{K} D_k}\sum_{k=1}^{K} D_k \nabla f_k(\mathbf{w}_t). \tag{3}$$

The global parameter $\mathbf{w}_t$ is updated using the estimated gradient $\hat{\mathbf{g}}_t$ (shown later) according to $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \hat{\mathbf{g}}_t$, where η_t is the learning rate of the distributed GD algorithm at iteration t. The iteration process continues until convergence.

III. Main Results & Discussions

In this section, we present two schemes considered in this paper. We show the convergence rate of the schemes, and show that alignment of the local gradients is not necessary.

A. FL Transmission Scheme over Fading MAC

The transmitted signal of the k-th user at iteration t is

$$\mathbf{x}_{k,t} = \frac{\sqrt{\alpha_{k,t}P_k}}{L}\mathbf{g}_{k,t}, \tag{4}$$

where we assume that the norm of gradient vectors are bounded by L, i.e., $\|\mathbf{g}_{k,t}\| \le L$. The scaling factor $\alpha_{k,t} \in [0,1]$ denotes a fraction of the maximum transmit power P_k at iteration t. From (1) and (4), the received signal at the PS is given as

$$\mathbf{y}_t = \sum_{k=1}^{K} \frac{h_{k,t}\sqrt{\alpha_{k,t}P_k}}{L}\mathbf{g}_{k,t} + \mathbf{n}_t.$$

We next present two transmission schemes which will be analyzed in the next Section.

1. Alignment: In this scheme, all users pick the coefficients $\alpha_{k,t}$s such that the transmitted local gradients are aligned at the PS, i.e., $c_t^{\text{align}} = \frac{h_{k,t}\sqrt{\alpha_{k,t}P_k}}{L}, \forall k$, where c_t^{align} is an alignment constant. User k picks $\alpha_{k,t}$ as $\alpha_{k,t} = \frac{(c_t^{\text{align}})^2 L^2}{h_{k,t}^2 P_k}$. Using the fact that $\alpha_{k,t} \le 1$, the alignment constant c_t^{align} can be upper bounded as

$$c_t^{\text{align}} \le \frac{\sqrt{\min_i h_{i,t}^2 P_i}}{L}. \tag{5}$$

In order to maximize the SNR of the received signal, we pick c_t^{align} as $c_t^{\text{align}} = \frac{\sqrt{\min_i h_{i,t}^2 P_i}}{L}$. Therefore, we obtain $\alpha_{k,t} = \frac{\min_i h_{i,t}^2 P_i}{h_{k,t}^2 P_k}$. It is worth noting that the alignment scheme is effectively limited by the user with the worst effective SNR, i.e., $\min_i h_{i,t}^2 P_i$. Upon receiving the local gradients estimates, the PS performs post-processing on $\mathbf{y}_t$ as follows

$$\hat{\mathbf{g}}_t = \frac{1}{K c_t^{\text{align}}}\mathbf{y}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_{k,t} + \frac{1}{K c_t^{\text{align}}}\mathbf{n}_t, \tag{6}$$

where c_t^{align} is a post-processing scaling factor at iteration t.

2. Full Power: The second scheme that we study is the full power scheme, where all users pick $\alpha_{k,t} = 1, \forall k$. The PS performs post-processing on $\mathbf{y}_t$ as follows

$$\hat{\mathbf{g}}_t = \frac{1}{c_t^{\text{F.P.}}}\mathbf{y}_t = \frac{1}{c_t^{\text{F.P.}}}\sum_{k=1}^{K}\underbrace{\frac{h_{k,t}\sqrt{P_k}}{L}}_{\psi_{k,t}}\mathbf{g}_{k,t} + \frac{1}{c_t^{\text{F.P.}}}\mathbf{n}_t,$$

where $c_t^{\text{F.P.}}$ is a post-processing scaling factor at iteration t. Depending on the choice of $c_t^{\text{F.P.}}$, we can get obtain both unbiased and biased estimators for full power scheme with an additional assumption that the local gradients across users are i.i.d. with bounded second moments, i.e., $\mathbb{E}\left[\mathbf{g}_{k,t}\right] = \mathbf{g}_t$, and $\mathbb{E}\left[\|\mathbf{g}_{k,t}\|^2\right] \leq (1+\epsilon)\|\mathbf{g}_t\|^2$, and the variance of the estimated gradient is $\mathrm{Var}\left(\hat{\mathbf{g}}_t\right) = \left(\epsilon/\left(c_t^{\text{F.P.}}\right)^2\right)\sum_{k=1}^{K}\psi_{k,t}^2\|\mathbf{g}_t\|^2 + dN_0/\left(c_t^{\text{F.P.}}\right)^2$. To obtain an unbiased estimator, we let

$$c_t^{\text{F.P.}} = c_t^{\text{F.P., unbiased}} \triangleq \sum_{k=1}^{K}\psi_{k,t}, \tag{7}$$

We note that any other choice of $c_t^{\text{F.P.}}$ leads to biased estimators of the full gradient. For the case when the estimated gradient is a biased estimator, we define bias as follows, $\mathbf{b}_t = \mathbf{g}_t - (1/c_t^{\text{F.P.}})\sum_{k=1}^{K}\psi_{k,t}\mathbf{g}_{k,t}$, and treat the channel noise term separately. For simplicity, we let $\widetilde{\mathbf{g}}_t = \sum_{k=1}^{K}\psi_{k,t}\mathbf{g}_{k,t}$. We can observe that there is a clear tradeoff between the bias and variance of the gradient estimate $\hat{\mathbf{g}}_t$. One could potentially reduce the variance and speed up the convergence by introducing bias.

B. Convergence rate

We next analyze the convergence rate of the schemes considered when the global loss function is *non-convex* and μ-smooth. Under these assumptions, we want to show that the average expected norm squared of the true gradient diminishes as the number of iterations T increases, which indicates that FedSGD converges to a stationary point. We first look at the case when the estimated gradient $\hat{\mathbf{g}}_t$ is unbiased.

Theorem 1. *Suppose the loss function F is non-convex, μ-smooth with respect to $\mathbf{w}$ and the local gradients are i.i.d. For a learning rate of $\eta = \min(1/\mu, \widetilde{\eta})$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \leq \frac{1}{T}\left[\frac{2R}{\eta} + \frac{\eta\mu d}{K^2}\sum_{t=0}^{T-1}\frac{N_0}{C_t^2}\right], \tag{8}$$

where $R = E\left[f(\mathbf{w}_0)\right] - f^*$ and $\widetilde{\eta} = \sqrt{2R/B}, B = (d\mu/K^2)\sum_{t=0}^{T-1}N_0/C_t^2$, where C_t can be either c_t^{align} in (5) or $c_t^{F.P.,\ unbiased}$ in (7) depending on the scheme.

We note that the bound in (8) behaves as $\mathcal{O}(1/\sqrt{T})$. In addition, one can check that bound (8) is a monotonic increasing function of B. Since $c_t^{\text{align}} < c_t^{\text{F.P., unbiased}}$, we conclude that under the assumption of i.i.d. local gradients, the unbiased full power scheme can outperform the alignment scheme. We next present the result for the case when the estimated gradient is a biased estimator of the true gradient.

Theorem 2. *Suppose the loss function F is non-convex and μ-smooth with respect to $\mathbf{w}$. Then for a learning rate of $\eta = \min\{1/\mu, \tilde{\eta}\}$, we have the following bound:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \leq \frac{1}{T\min_t m_t}\left(\frac{2R}{\eta} + \eta\mu d\sum_{t=0}^{T-1}\frac{N_0}{\left(c_t^{F.P.}\right)^2}\right), \tag{9}$$

for any

$$c_t^{F.P.} > \frac{\left(\sum_{k=1}^{K}h_{k,t}\sqrt{P_k}\right)^2 + \epsilon\sum_{k=1}^{K}h_{k,t}^2P_k}{2L\sum_{k=1}^{K}h_{k,t}\sqrt{P_k}} \triangleq c_{t,LB}^{F.P.}, \tag{10}$$

where $R = E\left[f(\mathbf{w}_0)\right] - f^*$, $\tilde{\eta} = \sqrt{2R/D}, D = \mu d\sum_{t=0}^{T-1}N_0/(c_t^{F.P.})^2$, and

$$m_t = \frac{2}{c_t^{F.P.}}\sum_{k=1}^{K}\psi_{k,t} - \left(\frac{1}{c_t^{F.P.}}\sum_{k=1}^{K}\psi_{k,t}\right)^2 - \frac{\epsilon}{\left(c_t^{F.P.}\right)^2}\sum_{k=1}^{K}\psi_{k,t}^2.$$

The bound in (9) also behaves as $\mathcal{O}(1/\sqrt{T})$. When $c_t^{\text{F.P.}} = c_t^{\text{align}}$, the convergence bound presented in Theorem 2 recovers the result in Theorem 1 up to a constant term depending on ϵ. We also observe that the term $\min_t m_t$ depends on the choice of $c_t^{\text{F.P.}}$, hence we can get a better constant (compared to Theorem 1) by increasing $\min_t m_t$. It can be readily shown that the biased scheme in Theorem 2 performs better than unbiased schemes by picking any $c_t^{\text{F.P.}}$ in the region $\mathcal{R}$, where $\mathcal{R} \triangleq \{c_t^{\text{F.P.}} : (4-K)^2(c_t^{\text{F.P.}})^2 - 2\alpha_t c_t^{\text{F.P.}} + \alpha_t^2 > 0\ \&\ c_t^{\text{F.P.}} > c_{t,\text{LB}}^{\text{F.P.}}\}$, where $\alpha_t = \sum_{k=1}^{K}\psi_{k,t} + \epsilon$. $\mathcal{R}$ is obtained by comparing the right-hand sides of bounds in (8) and (9). The term m_t shows up due to the bias and appears in the denominator in the proof of Theorem 2. Since $c_t^{\text{F.P.}}$ controls the bias, hence m_t, it needs to satisfy (10) to ensure that $\min_t m_t > 0$. This observation reveals that there are regimes in which biased gradient aggregation can outperform unbiased gradient aggregation (as highlighted in the experiments section as well).

Remark 1. *We note that biased SGD in the centralized setting has been studied in [12]. The key challenge in the wireless federated learning setting is that the channels and the power constraints need to be considered while designing transmission schemes and deriving convergence rates. It significantly changes the proof of Theorem 2 and allows us to gain insight on how post-processing should be done for fast convergence.*

IV. Experiments

In this section, we evaluate the performance of the wireless FL schemes through experiments. We consider image classification task on MNIST dataset with LeNet-5 architecture. MNIST dataset consists of $60,000$ training samples, and $10,000$ testing samples. The training samples are divided evenly and distributed randomly across $K = 10$ users. Cross-entropy and SGD optimizer are used for training with a learning rate of $\eta = 0.01$. We consider two settings, homogeneous and heterogeneous. In homogeneous setting, all users have the same channel statistics and transmit powers, i.e., $h_{k,t} \sim \mathcal{N}(0.5, 0.01)$ and $P_k = 5$ dB for all k. In heterogeneous setting, we split users into two groups. Without loss of generality, we place the first five users in the first group and the rest in the second group. The first group consists of five users with good channel conditions and high transmit power, i.e., $h_{k,t} \sim \mathcal{N}(1, 0.01)$ and $P_k = 20$ dB. The second group consists of the other five users with poor channel conditions and low transmit power, i.e., $h_{k,t} \sim \mathcal{N}(0.004, 10^{-6})$ and $P_k = 0$ dB. For both settings, $N_0 = 1$. The considered schemes

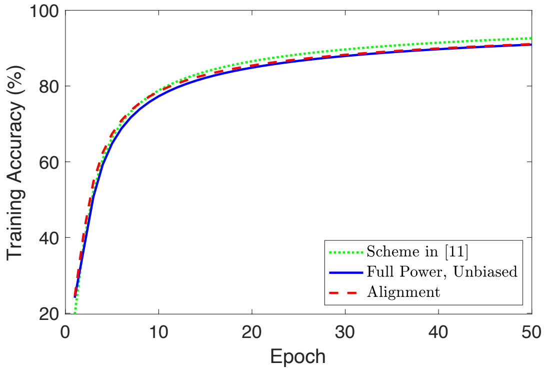

Fig. 2. For the case when channels and powers are homogeneous for all users, unbiased full power scheme performs as well as alignment scheme for MNIST image classification.

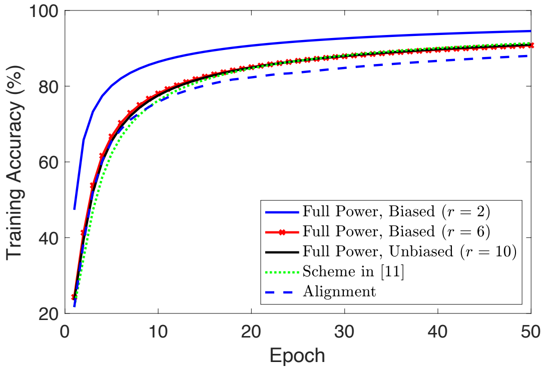

Fig. 3. For the case when there is heterogeneity in channel conditions and power constraints, with proper choices of $c_t^{\text{F.P.}}$'s, biased full power scheme outperforms both unbiased full power and alignment schemes.

are compared to the scheme proposed in [11], which minimizes the MSE between true gradients and estimated gradients by optimizing power control and post-processing factors using the statistics of the gradients and channel conditions. Under homogeneous setting, it can be seen in Fig. 2 that the scheme in [11] has the best performance among the three due to optimized power control and post-processing factors. However, alignment and full power schemes achieve comparable performance without the need of optimization. In addition, it can be seen that full power performs as well as the alignment scheme. Under heterogeneous setting, $c_t^{\text{F.P}}$'s are chosen by selecting the top r users who have the largest $\psi_{k,t}$, i.e., $c_t^{\text{F.P}} = \sum_{k=1}^{r} \psi_{k,t}$. We can observe in Fig. 3 that alignment has the worst performance. We note that by minimizing the MSE between true and estimated gradients, the scheme in [11] tries to minimize the bias as well. Therefore, the scheme in [11], unbiased full power scheme, and biased full power scheme with $r = 6$ achieve similar performance. In addition, biased full power scheme with $r = 2$ achieve the best performance. We can also observe that the performance is better when r is small. This is due to the fact that when r is small, $1/c_t^{\text{F.P}}$ is large, hence, the gradients are magnified. However, when r passes 5, $\psi_{k,t}, k = 6, \ldots, 10$ are too small and do not contribute significantly to the post-processing factor $c_t^{\text{F.P}}$. Therefore, the performances for $r = 6$ and $r = 10$ do not change much.

V. Conclusion

In this work, we studied the problem of wireless federated learning and considered two analog transmission schemes, alignment and full power schemes for FedSGD. We derived convergence bounds for the case when the global loss function is non-convex and smooth. Our first finding is that under the assumption of i.i.d. data, full power transmission scheme can outperform the alignment scheme, thereby highlighting that one is not always necessarily limited by worst users' SNR when performing analog aggregation. Secondly, we also show that there are regimes (in terms of channel conditions across users), where biased gradient aggregation can outperform unbiased gradient aggregation, especially when the channels across users are heterogenous. Generalizing the analysis to non-i.i.d. setting and validating the results using other datasets for broader studies are very interesting future directions.

References

[1] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[2] W.-T. Chang and R. Tandon, "MAC aware quantization for distributed gradient descent," in *IEEE Global Communications Conference (GLOBECOM)*, 2020, pp. 1–6.

[3] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2019.

[4] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[5] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2897–2911, 2020.

[6] Y. Shao, D. Gunduz, and S. C. Liew, "Federated edge learning with misaligned over-the-air computation," *arXiv preprint arXiv:2102.13604*, 2021.

[7] E. Ozfatura, K. Ozfatura, and D. Gündüz, "Time-correlated sparsification for communication-efficient federated learning," *arXiv preprint arXiv:2101.08837*, 2021.

[8] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2604–2609.

[9] M. Seif, W.-T. Chang, and R. Tandon, "Privacy amplification for federated learning via user sampling and wireless aggregation," *arXiv preprint arXiv:2103.01953*, 2021.

[10] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170–185, 2020.

[11] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *to appear in IEEE Transactions on Wireless Communications*, 2021.

[12] A. Ajalloeian and S. U. Stich, "Analysis of SGD with biased gradient estimators," *arXiv preprint arXiv:2008.00051*, 2020.

Appendix A: Proof of Theorem 1

Starting with the μ-smooth assumption and taking expectation over noise and randomness of SGD, we obtain the following inequalities:

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{w}_{t+1})\right] \le f(\mathbf{w}_t) - \eta \mathbf{g}_t^T \mathbb{E}\left[\hat{\mathbf{g}}_t\right] + \frac{\eta^2\mu}{2}\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|^2\right] \\
&\overset{(a)}{=} f(\mathbf{w}_t) - \eta\|\mathbf{g}_t\|^2 + \frac{\eta^2\mu}{2}\left[\|\mathbf{g}_t\|^2 + \frac{1}{C_t^2K^2}\mathbb{E}\left[\|\mathbf{n}_t\|^2\right]\right] \\
&\overset{(b)}{=} f(\mathbf{w}_t) - \frac{\eta}{2}\|\mathbf{g}_t\|^2 + \frac{\eta^2\mu d N_0}{2C_t^2K^2},
\end{aligned}$$

where (a) follows from the facts that $\hat{\mathbf{g}}_t$ is an unbiased estimator of $\mathbf{g}_t$, noise is zero mean and by choosing $\eta \le 1/\mu$;

and (b) follows from the fact that each element of the noise vector has variance N_0. We then take expectation over the randomness of the model and apply telescoping sum.

$$\mathbb{E}\left[f(\mathbf{w}_T)\right] \le \mathbb{E}\left[f(\mathbf{w}_0)\right] - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] + \frac{\eta^2\mu d}{2K^2}\sum_{t=0}^{T-1}\frac{N_0}{C_t^2}.$$

Lower bounding $\mathbb{E}\left[f(\mathbf{w}_T)\right]$ with f^* and rearranging, we get,

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le \mathbb{E}\left[f(\mathbf{w}_0)\right] - f^* + \frac{\eta^2\mu d}{2K^2}\sum_{t=0}^{T-1}\frac{N_0}{C_t^2}.$$

Letting $R = \mathbb{E}\left[f(\mathbf{w}_0)\right] - f^*$ and rearranging, we have,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le \frac{1}{T}\left(\frac{2R}{\eta} + \frac{\eta\mu d}{K^2}\sum_{t=0}^{T-1}\frac{N_0}{C_t^2}\right)$$

We can then optimize the learning rate by minimizing the bound. This completes the proof of Theorem 1.

Appendix B: Proof of Theorem 2

Before proving Theorem 2, we first introduce an upper bound on the expectation of the squared norm of the bias through the following Lemma.

Lemma 1. *Under the assumptions that the local gradients are i.i.d. with bounded second moment, the expectation of the squared norm of the bias can be bounded using the squared norm of the true gradient as follows,*

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{b}_t\|^2\right]\\ &\le \left(1 + \left(\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)^2 - 2\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}} + \epsilon\sum_{k=1}^{K}\left(\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)^2\right)\|\mathbf{g}_t\|^2\end{aligned}$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{b}_t\|^2\right] = \mathbb{E}\left[\left\|\mathbf{g}_t - \frac{1}{c_t^{\text{F.P.}}}\widetilde{\mathbf{g}}_t\right\|^2\right]\\
&= \|\mathbf{g}_t\|^2 - 2\mathbf{g}_t^T\mathbb{E}\left[\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\mathbf{g}_{k,t}\right] + \mathbb{E}\left[\left\|\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\mathbf{g}_{k,t}\right\|^2\right]\\
&\overset{(a)}{=} \|\mathbf{g}_t\|^2 - 2\left(\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)\|\mathbf{g}_t\|^2\\
&+\mathbb{E}\left[\sum_{k=1}^{K}\left(\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)^2\|\mathbf{g}_{k,t}\|^2 + \sum_{k=1}^{K}\sum_{k'\neq k}\frac{\psi_{k,t}\psi_{k',t}}{\left(c_t^{\text{F.P.}}\right)^2}\mathbf{g}_{k,t}^T\mathbf{g}_{k',t}\right]\\
&= \left(1 + \sum_{k=1}^{K}\sum_{k'\neq k}\frac{\psi_{k,t}\psi_{k',t}}{\left(c_t^{\text{F.P.}}\right)^2} - 2\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)\|\mathbf{g}_t\|^2\\
&\qquad + \sum_{k=1}^{K}\left(\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)^2\mathbb{E}\left[\|\mathbf{g}_{k,t}\|^2\right]\\
&\overset{(b)}{\le} \left(1 + \left(\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)^2 - 2\sum_{k=1}^{K}\frac{\psi_{k,t}}{c_t^{\text{F.P.}}} + \epsilon\sum_{k=1}^{K}\left(\frac{\psi_{k,t}}{c_t^{\text{F.P.}}}\right)^2\right)\|\mathbf{g}_t\|^2\\
&\triangleq (1 - m_t)\|\mathbf{g}_t\|^2,
\end{aligned}$$

where (a) and (b) follow from the assumptions that the gradients across users are i.i.d. with bounded second moments. □

We now are ready to prove Theorem 2. Starting with the smoothness assumption and taking expectation over noise and randomness of SGD, we obtain the following inequalities:

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{w}_{t+1})\right] \le f(\mathbf{w}_t) - \eta\mathbf{g}_t^T\mathbb{E}\left[\hat{\mathbf{g}}_t\right] + \frac{\eta^2\mu}{2}\mathbb{E}\left[\|\hat{\mathbf{g}}_t\|^2\right]\\
&\overset{(a)}{=} f(\mathbf{w}_t) - \frac{\eta}{c_t^{\text{F.P.}}}\mathbf{g}_t^T\mathbb{E}\left[\widetilde{\mathbf{g}}_t\right] + \frac{\eta^2\mu}{2}\left[\frac{\mathbb{E}\left[\|\widetilde{\mathbf{g}}_t\|^2\right]}{\left(c_t^{\text{F.P.}}\right)^2} + \frac{dN_0}{\left(c_t^{\text{F.P.}}\right)^2}\right]\\
&\overset{(b)}{\le} f(\mathbf{w}_t) - \frac{\eta}{c_t^{\text{F.P.}}}\mathbf{g}_t^T\mathbb{E}\left[\widetilde{\mathbf{g}}_t\right] + \frac{\eta}{2\left(c_t^{\text{F.P.}}\right)^2}\mathbb{E}\left[\|\widetilde{\mathbf{g}}_t\|^2\right] + \frac{\eta^2\mu dN_0}{2\left(c_t^{\text{F.P.}}\right)^2}\\
&= f(\mathbf{w}_t) - \eta\mathbf{g}_t^T\mathbb{E}\left[\mathbf{g}_t - \mathbf{b}_t\right] + \frac{\eta}{2}\mathbb{E}\left[\|\mathbf{g}_t - \mathbf{b}_t\|^2\right] + \frac{\eta^2\mu dN_0}{2\left(c_t^{\text{F.P.}}\right)^2}\\
&= f(\mathbf{w}_t) + \frac{\eta}{2}\left(-\|\mathbf{g}_t\|^2 + \mathbb{E}\left[\|\mathbf{b}_t\|^2\right]\right) + \frac{\eta^2\mu dN_0}{2\left(c_t^{\text{F.P.}}\right)^2}\\
&\overset{(c)}{\le} f(\mathbf{w}_t) - \frac{\eta}{2}m_t\|\mathbf{g}_t\|^2 + \frac{\eta^2\mu dN_0}{2\left(c_t^{\text{F.P.}}\right)^2},
\end{aligned}$$

where (a) follows due to zero mean noise vector with element that has variance N_0; (b) follows from choosing $\eta \le 1/\mu$; (c) follows from Lemma 1. We then take expectation over the model and obtain,

$$\mathbb{E}\left[f(\mathbf{w}_{t+1})\right] \le \mathbb{E}\left[f(\mathbf{w}_t)\right] - \frac{\eta}{2}m_t\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] + \frac{\eta^2\mu dN_0}{2\left(c_t^{\text{F.P.}}\right)^2}.$$

We then again apply telescoping sum and lower bound $\mathbb{E}\left[f(\mathbf{w}_T)\right]$ with f^*. With some rearranging, we get,

$$\frac{\eta}{2}\sum_{t=0}^{T-1}m_t\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le \mathbb{E}\left[f(\mathbf{w}_0)\right] - f^* + \frac{\eta^2\mu d}{2}\sum_{t=0}^{T-1}\frac{N_0}{\left(c_t^{\text{F.P.}}\right)^2}.$$

We can then lower bound m_t by $\min_t m_t$, and subsequently divide both sides by $\frac{\eta(\min_t m_t)}{2T}$ to get,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \le \frac{1}{T\min_t m_t}\left(\frac{2R}{\eta} + \eta\mu d\sum_{t=0}^{T-1}\frac{N_0}{\left(c_t^{\text{F.P.}}\right)^2}\right).$$

We get the expression in Theorem 2 by choosing $\eta = 1/\mu$. However, this is only true if $\min_t m_t$ is positive. Therefore, we can obtain a condition on $c_t^{\text{F.P.}}$ to ensure that $\min_t m_t > 0$:

$$c_t^{\text{F.P.}} > \frac{\left(\sum_{k=1}^{K}\psi_{k,t}\right)^2 + \epsilon\sum_{k=1}^{K}\psi_{k,t}^2}{2\sum_{k=1}^{K}\psi_{k,t}}.$$

This completes the proof of Theorem 2.