I. Convergence Analysis

Before analyzing the convergence of the proposed FedWA mechanism, we first introduce the following assumptions (similar assumptions can be seen in [1]–[4]).

Definition 1 *(L-Lipschitz smoothness) $F_i(\mathbf{w})$ is L-Lipschitz smoothness for each of the participating nodes $i \in \mathcal{S}_t$, i.e., $\|\nabla F_i(\mathbf{w}) - \nabla F_i(\mathbf{w}')\| \leq L\|\mathbf{w}-\mathbf{w}'\|$ for any two parameter vectors $\mathbf{w}$, $\mathbf{w}'$.*

Definition 2 *The sequence of iterations $F(\mathbf{w}_t)$ is contained in an open set over which F is bounded below by a scalar F^*.*

Definition 3 *The stochastic gradient $\nabla F(\mathbf{w};\ \xi)$ computed from random samples ξ is an unbiased estimator of the true gradient for the parameter $\mathbf{w}$, i.e.,*

$$\mathbb{E}_{\xi}[\nabla F(\mathbf{w};\xi)] = \nabla F(\mathbf{w}) \tag{1}$$

Definition 4 *(Non-IID data) Let ξ_t^i be the sample that is randomly sampled from the client i's local data in t-th communication round. The variant of stochastic gradient in each client is bounded,*

$$\mathbb{E}\left\|\nabla F_i\left(\mathbf{w}_t^i;\xi_t^i\right) - \nabla F_i\left(\mathbf{w}_t^i\right)\right\|_2^2 \leq \sigma_i^2 \tag{2}$$

for any t, i, where σ_i^2 is a constant, and $\sigma_i^2 > 0$.

Next, we derive an upper bound on the expected average squared gradient norms, which can be viewed as a metric to measure the convergence rate for non-convex objective. Suppose that E is the number of epochs of each client, φ_t^i denotes the aggregation weight assigned to client i in communication round t, I is the total number of clients, and m_i means the number of local updates in one communication round. According to Assumptions 1, 2, 3, and 4, we can get

Theorem 1 *The expected average squared gradient norms of $F(\mathbf{w}_t)$ converges to a nonzero constant as $T \to \infty$ under a fixed learning rate.*

Proof. Firstly, based on Assumption 1, we can get that $F(\mathbf{w})$ is L-Lipschitz smoothness. In addition, for any mini-batch $\mathcal{B}^i$, the variance of stochastic gradient decreases by a factor of $b_i = \left|\mathcal{B}^i\right|$, namely,

$$\mathbb{E}\left\|\nabla F_i\left(\mathbf{w}_t^i;\xi_{B_t^i}\right) - \nabla F_i\left(\mathbf{w}_t^i\right)\right\|_2^2 \leq \frac{\sigma_i^2}{b_i}, \tag{3}$$

where B_t^i is the mini-batch sample sampled from the client i's local data in t-th communication round.

Suppose that our algorithm runs with a fixed learning rate $\eta_t = \eta$, which satisfies [5]

$$\sum_{i=1}^{I}\left[\frac{(m_i-2)(m_i+1)}{2} - \frac{1}{L^2\eta^2} + \frac{\varphi_t^i m_i}{L\eta}\right] \leq 0, \tag{4}$$

where $m_i = D_i E/b_i$, and $0 \leq \varphi_t^i \leq 1$.

Then, according to [5], the expected average squared gradient norms of F satisfies the following bound for all $T \in \mathbb{N}$,

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left\|\nabla F\left(\mathbf{w}_t\right)\right\|_2^2 \\ \leq\ &\frac{F\left(\mathbf{w}_1\right)-F^*}{T(G-A-C)} + \frac{L\eta^2}{2I(G-A-C)}\sum_{i=1}^{I}\left(\varphi_t^i\right)^2\beta_i m_i \\ &+\frac{L^2\eta^3}{12I(G-A-C)}\sum_{i=1}^{I}\varphi_t^i\beta_i\left(m_i-1\right)m_i\left(2m_i-1\right), \end{aligned} \tag{5}$$

where $\beta_i = \sigma_i^2/b_i$, $A = \frac{L^2\eta^3}{4I}\sum_{i=1}^{I} m_i\left(m_i-1\right)$, $G = \frac{\eta}{2I}\sum_{i=1}^{I}\left(m_i+1\right)$, and $C = \frac{L\eta^2}{2I}\sum_{i=1}^{I} m_i$. As φ_t^i is the aggregation weight assigned to client i in communication round t, and $0 \leq \varphi_t^i \leq 1$, it does not affect the convergence of the model. Therefore, we prove the Theorem. □

References

[1] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, On the convergence of fedavg on non-iid data, in Proc. of ICLR, 2020.

[2] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning, Siam Review, vol. 60, no. 2, pp. 223–311, 2018.

[3] H. Yu, S. Yang, and S. Zhu, Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning, in Proc. of AAAI, vol. 33, 2019, pp. 5693–5700.

[4] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, Sparsified SGD with memory, in Proc. of NeurIPS, 2018, pp. 4447–4458.

[5] J. Zhang et al., "Adaptive Federated Learning on Non-IID Data with Resource Constraint," in IEEE Transactions on Computers, doi: 10.1109/TC.2021.3099723.