Private Federated Learning with Dynamic Power Control via Non-Coherent Over-the-Air Computation

Anbang Zhang [1 2] **Shuaishuai Guo** [1 2] **Shuai Liu** [1]

Abstract

To further preserve model weight privacy and improve model performance in Federated Learning (FL), FL via Over-the-Air Computation (AirComp) scheme based on dynamic power control is proposed. The edge devices (EDs) transmit the signs of local stochastic gradients by activating two adjacent orthogonal frequency division multiplexing (OFDM) subcarriers, and majority votes (MVs) at the edge server (ES) are obtained by exploiting the energy accumulation on the subcarriers. Then, we propose a dynamic power control algorithm to further offset the biased aggregation of the MV aggregation values. We show that the whole scheme can mitigate the impact of the time synchronization error, channel fading and noise. The theoretical convergence proof of the scheme is re-derived.

1. Introduction

With the substantial increase in computation ability and storage capacity of modern intelligent terminals, distributed FL is utilized the most widely, which provides a promising learning paradigm for the current privacy computation. By pushing model training locally (McMahan et al., 2023), FL is able to build global models without directly sharing data. This mechanism largely protects the privacy and security (Chen et al., 2021a) of users and addresses the potential for data leakage in the context of large amounts of data.

However, there are still numerous key challenges in deploying practical FL applications in the real world due to resource constraints (Chen et al., 2021b) and privacy concerns in wireless networks. Several works (Melis et al., 2019) show that if the model parameters or gradients exchanged between EDs and ES are attacked, sensitive information about local data is still exposed. In addition, a large number of model parameters need to be repeatedly transmitted over wireless channels (Li et al., 2020), thus requiring huge communication resources, which is a significant bottleneck.

Table 1. The communication cost of different gradient compression schemes, when training a D-dimensional model with M EDs.

ALGORITHM	BITS PER ITERATION
SGD	64MD
QSGD	(2 + LOG(2M+ 1))MD
TERNGRAD	(2 + LOG(2M+ 1))MD
SIGNSGD WITH MV	2MD

1.1. Related Work

Gradient compression: To overcome the above problems, a prospective solution is gradient quantization, such as SignSGD, QSGD (Alistarh et al., 2017), and cpSGD (Agarwal et al., 2018). But transmitting the gradients can also have some privacy leakage (Zhu et al., 2019), so deeper compression methods should be considered more. In (Bernstein et al., 2018), it is shown that actually gradients are really useful in terms of the direction rather than size. Thus, SignSGD considers to quantize gradients, achieving a 32 times data compression. Also, based on the FL scheme, it is difficult to recover the data information used in the model by hijacking the gradient direction due to reducing the transmission of information (Akoun & Meyer, 2022), thus ensuring the privacy and security of the data information.

Over-the-air Computation: The second option to consider is to adopt the weight superposition over the air, i.e., AirComp (Liu et al., 2020). This combination of communication and computation reduces the latency and bandwidth requirements (Goldenbaum et al., 2013). However, the usual AirComp scheme requires channel state information (CSI) at the EDs or the ES. To alleviate the channel estimation burden, (Zhu et al., 2021) considered that EDs use truncated-channel inversion (TCI) to transmit orthogonal phase shift keying (QPSK) symbols on the orthogonal frequency division multiplexing (OFDM) subcarrier instead of transmit-

[1]School of Control Science and Technology, Shandong University, Jinan, China. [2]Shandong Key Laboratory of Wireless Communication Technologies, Jinan, China.. Correspondence to: Shuaishuai Guo <shuaishuai_guo@sdu.edu.cn>.

Workshop of Federated Learning and Analytics in Practice, colocated with 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA.

ting the gradient or the direction of the gradient directly, which corresponds to a double layer of privacy encryption protection and guarantees absolute security of model weight privacy during communication.

Dynamic Power Control: In the context of the above approach, the privacy security as well as communication overhead issues are effectively addressed, but the multi-user parameter aggregation produces discrepancies on training process of the global model, which has become the focus of the research. Therefore, we consider to adjust transmitting power (Li et al., 2022) with EDs in order to offset the impact of the user parameters on the whole model in the opposite direction of the model convergence during the parameter aggregation, thus achieving better model accuracy.

1.2. Contributions

All the contributions can be listed as follows:

(1) By considering FL training based on SignSGD with Majority Vote (MV), an Over-the-Air Computation scheme is utilized in which symbols (i.e., directions) of random gradients are transmitted by using OFDM symbols. The MV is obtained by energy detection on the ES, so that CSI is not required at the EDs and ES.

(2) We design a dynamic power control scheme that trades off MV and the transmitting signs of each ED to offset the effect of the EDs parameters on the whole model in the opposite direction of the model convergence when the parameters are aggregated.

(3) Then, we re-derive the theoretical convergence proof of the proposed scheme, when MVs are obtained by employing the FL scheme. The experiment results can prove that the proposed scheme is robust to time synchronization errors because it does not encode signs of local stochastic gradients into the phase of the transmitted symbols.

A figure demonstrating the training process is given in Appendix A.

2. Overview of Our Approach

2.1. Federated Learning With Majority Vote

We consider an FL system comprising a single edge server coordinating the learning process across *M* EDs. The aim of FL can be represented as finding an optimal model parameter vector $\mathbf{w}^*$ that minimizes $F(\mathbf{w})$, i.e.,

$$\mathbf{w}^* = \min_{\mathbf{w}\in\mathbb{R}^q} F(\mathbf{w}) = \min_{\mathbf{w}\in\mathbb{R}^q} \frac{1}{|D|} \sum_{\forall(x,y)\in D} f(\mathbf{w};\mathbf{x},y), \quad (1)$$

where $\mathcal{D} = \bigcup_{m=1}^{M}\{\mathcal{D}_m\}$ is the global dataset set and $f(\mathbf{w},\mathbf{x},y)$ is the sample-wise loss function indicating the prediction error, for example, $(\mathbf{x},y)$ with the FL model parameters $\mathbf{w} = [w_1,\ldots,w_q]^{\mathrm{T}} \in \mathbb{R}^q$, and q is the number of model parameters.

Algorithm 1 signSGD-MV based on AirComp

Input: learning rate η, current received global model $\mathbf{w}^{(n)}$, *M* EDs each with $\bar{\mathbf{g}}_m^{(n)}$, initialize $\mathbf{w}^{(0)}$.
repeat
  On Each Edge Device:
  calculate the sign $\bar{\mathbf{g}}_m^{(n)}$ of stochastic gradients.
  update $\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} - \eta\mathbf{v}^{(n)}$.
  On Edge Server:
  pull the sign $\bar{\mathbf{g}}_m^{(n)}$ from *m* ED with non-coherent energy detection via AirComp.
  broadcast $\mathbf{v}^{(n)} = \mathrm{sign}(\sum_{m=1}^{M}\bar{\mathbf{g}}_m^{(n)})$ to all the EDs.
until reach convergence

In the training process, EDs exploit a mini-batch stochastic gradient descent method to calculate local gradient $\tilde{\mathbf{g}}_m^{(n)} \triangleq [\tilde{g}_{m,1}^{(n)},\ldots,\tilde{g}_{m,q}^{(n)}]^{\mathrm{T}}$ with respect to the current received global model $\mathbf{w}^{(n)}$ as

$$\tilde{\mathbf{g}}_m^{(n)} = \nabla F_m\left(\mathbf{w}^{(n)}\right) = \frac{1}{d_{\mathrm{b}}} \sum_{\forall(\mathbf{x}_\ell,y_\ell)\in\tilde{\mathcal{D}}_m} \nabla f\left(\mathbf{w}^{(n)},\mathbf{x}_\ell,y_\ell\right), \quad (2)$$

where $\tilde{\mathcal{D}}_m \subset \mathcal{D}_m$ is selected data batch from local data set and $d_{\mathrm{b}} = |\tilde{\mathcal{D}}_m|$ as the batch size. In the context of FL processing, SignSGD with the majority vote approach (Bernstein et al., 2018) is investigated to solve above problems. The trained real stochastic gradients are converted into sign values by one-bit quantization scheme, which are denoted as $\bar{\mathbf{g}}_{m,i}^{(n)} \triangleq \mathrm{sign}(\tilde{\mathbf{g}}_{m,i}^{(n)})$. Thus, the parameter MV of the *i*th global gradient estimate at the ES would be enforced as follows:

$$v_i^{(n)} = \mathrm{sign}\left(\sum_{m=1}^{M}\bar{\mathbf{g}}_{m,i}^{(n)}\right). \quad (3)$$

Afterwards, the ES pushes $\mathbf{v}^{(n)} = [v_1^{(n)},\ldots,v_q^{(n)}]^{\mathrm{T}}$ to the EDs, and the models at the EDs are updated as:

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} - \eta\mathbf{v}^{(n)}. \quad (4)$$

This procedure is repeated consecutively until a predetermined convergence criterion is achieved. Combining above process, the corresponding scheme is Algorithm 1.

2.2. Transmitter Design - Dynamic Power Control on FSK-MV

We concentrate on uplink communication process based on the MV with AirComp (i.e., FSK-MV (Sahin et al., 2021)). At the *n*th communication round, the superposed symbol on

the *l*th subcarrier of *m*th OFDM symbol can be given by:

$$\mathbf{y}_{l,s}^{(n)} = \sum_{m=1}^{M} \sqrt{P_m^{(n)}} \mathbf{H}_{m,l,s}^{(n)} t_{m,l,s}^{(n)} + \mathbf{n}_{l,s}^{(n)}, \tag{5}$$

where $\mathbf{H}_{m,l,s}^{(n)} \in \mathbb{C}$ is the channel coefficient with identical Rayleigh distribution, and $t_{m,l,s}^{(n)} \in \mathbb{C}$ is the transmitted symbol, and $\mathbf{n}_{l,s}^{(n)}$ is an additive white Gaussian noise vector on the *l*th subcarrier for $l \in \{0, 1, \ldots, A-1\}$ and $s \in \{0, 1, \ldots, S-1\}$. The EDs perform a low-complexity operation that each ED activates one of the two adjacent subcarriers determined by the time-frequency index pairs to transmit the signs of the gradients. To express this encoding operation rigorously, let *f* be a bijective function that maps $i \in \{1, 2, \ldots, q\}$ to the distinct pairs (s^+, l^+) and (s^-, l^-). Thus, the *m*th ED determines the following bins of modulation symbol $\mathbf{t}_{m,l^+,s^+}^{(n)}$ and $\mathbf{t}_{m,l^-,s^-}^{(n)}$, $\forall i$, as

$$t_{m,l^+,s^+}^{(n)} = \sqrt{E_0} s_{m,i}^{(n)} \mathbb{I}\left[\bar{\mathbf{g}}_{m,i}^{(n)} \triangleq \operatorname{sign}(\tilde{\mathbf{g}}_{m,i}^{(n)}) = 1\right], \tag{6}$$

and

$$t_{m,l^-,s^-}^{(n)} = \sqrt{E_0} s_{m,i}^{(n)} \mathbb{I}\left[\bar{\mathbf{g}}_{m,i}^{(n)} \triangleq \operatorname{sign}(\tilde{\mathbf{g}}_{m,i}^{(n)}) = -1\right], \tag{7}$$

where $E_0 = 2$ is a factor to normalize the OFDM symbol energy, $s_{m,i}^{(n)}$ is a randomization symbol on the unit circle, and $\mathbb{I}$ is the indicate function. As a special case of the mapping function *f*, if $s^- = s^+$ and $l^- = l^+ + 1$, which holds for all *i*, then the adjacent subcarriers of m^+th OFDM symbol is used for the voting scenario, which corresponds to frequency-shift keying (FSK) on the OFDM subcarriers.

Dynamic Power Control: Based on the above analysis of the communication process, we consider the provision of a larger proportion of biased gradient symbols and larger EDs with Gaussian white noise. Therefore, we formulate the dynamic power control design problem in the direction of convergence of the balanced global model in the following form as:

$$P_m^{(n)} = P_m^{(n-1)} + \left| \frac{1}{q} \sum_{i=1}^{q} \left[\mathbb{I}_{\bar{g}_{m,i}^{(n)} = v_i^{(n)}} - \mathbb{I}_{\bar{g}_{m,i}^{(n)} \neq v_i^{(n)}} \right] \right|, \tag{8}$$

where $v_i^{(n)}$ and $\bar{g}_{m,i}^{(n)}$ is form of model parameter aggregation when channel transmission is not considered and the *i*th parameter of the vector shared by *m*th ED, respectively. Specially, the initialized transmission power is $P_m^{(0)} = 1$. The above power control scheme is adopted to be able to achieve the convergence direction convergence with the global model. See the proof in Appendix F.

2.3. Receiver Design - Non-coherent Energy Detection

Based on above theory, we choose to use a non-coherent energy detection (Adeli & Şahin, 2022) approach to obtain the MV. At the receiver ES, we first identify the pairs (s^+, l^+) and (s^-, l^-), and observe the superposed symbols, which are expressed as:

$$r_{l^+,s^+}^{(n)} = \sqrt{E_0} \sum_{\forall m, \bar{g}_{m,i}^{(n)} = 1} \sqrt{P_m^{(n)}} h_{m,l^+,s^+}^{(n)} s_{m,i}^{(n)} + n_{l^+,s^+}^{(n)}, \tag{9}$$

and

$$r_{l^-,s^-}^{(n)} = \sqrt{E_0} \sum_{\forall m, \bar{g}_{m,i}^{(n)} = -1} \sqrt{P_m^{(n)}} h_{m,l^-,s^-}^{(n)} s_{m,i}^{(n)} + n_{l^-,s^-}^{(n)}. \tag{10}$$

Subsequently, we exploit an energy detector to obtain the MV for the *i*th gradient as

$$v_i^{(n)} = \operatorname{sign}\left(\Delta_i^{(n)}\right), \tag{11}$$

where $\Delta_i^{(n)}$ is the sum energy to detect the votes and $\Delta_i^{(n)} = \mathbf{e}_i^+ - \mathbf{e}_i^-$, whose derivation process is shown in Appendix B. Also, e_i^+ and e_i^- are the energies of the superposed symbols on adjacent subcarriers.

3. Error Probability Analysis and Convergence Rate Performance

To facilitate subsequent analysis of convergence as well as error probability, several standard assumptions (Shi et al., 2022) are shown as Appendix C.

3.1. Error Probability Analysis

Received Signal Power of MV: The above scheme determines the correct MV by comparing e_i^+ and e_i^- directly. Also, let M_i^+ and M_i^- be the number of EDs that vote for $\bar{g}_{m,i}^{(n)} = 1$ and $\bar{g}_{m,i}^{(n)} = -1$, respectively. Then, we obtain the expressions of the average received signal power as μ_i^+ and μ_i^- with the following lemma:

Lemma 3.1. *For the given* M_i^+ *and* M_i^-, μ_i^+ *and* μ_i^- *can be calculated as*

$$\mu_i^+ \triangleq \mathbb{E}\left[e_i^+\right] = E_0 M_i^+ \vartheta + \sigma_{\mathrm{n}}^2, \tag{12}$$

and

$$\mu_i^- \triangleq \mathbb{E}\left[e_i^-\right] = E_0 M_i^- \vartheta + \sigma_{\mathrm{n}}^2, \tag{13}$$

respectively. The parameter ϑ is the average value of transmission power for all *i*, which is equivalent to a constant. The proof is given in Appendix B.

Bit Error Probability Analysis: Based on several assumptions provided in Appendix C, we proceed to analyze the error probability in the sign aggregation process. According to the Appendix F, the performance of our scheme is

bounded by probability of misidentifying the correct sign for the ith gradient, which is obtained by:

$$P_i^{\mathrm{err}} \triangleq \Pr\left(\mathrm{sign}\left(\Delta_i^{(n)}\right) \neq \mathrm{sign}\left(g_i^{(n)}\right)\right), \quad (14)$$

which is determined by the level of noise introduced by the data-stochasticity and wireless channel. For a more accurate formalization, P_i^{err} for MV can be obtained as follows:

Lemma 3.2. *(Error probability for MV). Under the dynamic power control method, we derive that stochasticity-induced error in the wireless fading channel for all* i *is bounded as:*

$$P_i^{err} \leq \frac{\frac{K}{2}\cdot\sqrt{2}/(3R_i)}{K+2/\beta} + \frac{1/\beta}{K+2/\beta}, \quad (15)$$

where $R_i = \sqrt{d_b}\frac{|g_i^{(n)}|}{\sigma_i}$ is defined as the gradient-signal-to-data-noise ratio (Zhu et al., 2021). Also, the resultant gradient variance reduces from σ_i^2 to σ_i^2/d_b according to Assumption C.5 and Equation (2). We provide the proof in Appendix D. And Lemma 3.2 implies the following results:

Corollary 3.3. *(Legitimate EDs). For $q_i < p_i$, X must be larger than $K/2$, meanwhile must satisfying $P_i^{err} < 1/2$.*

3.2. Convergence Rate over Fading Channel

It is obvious that the proposed scheme maintains the convergence of the original MV. Based on this, we re-derive the theoretical convergence performance of the proposed scheme as follows:

Theorem 3.4. *Consider a FL system based on the proposed scheme, for the mini-batch size $d_{\mathrm{b}} = N/\gamma$ and the learning rate $\eta = 1/\sqrt{\|\mathbf{L}\|_1 d_{\mathrm{b}}}$, the convergence rate in fading channel is given by*

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=0}^{N-1}\left\|\mathbf{g}^{(n)}\right\|_1\right] \leq \frac{1}{\sqrt{N}}\Big(\tau\sqrt{\|\mathbf{L}\|_1}\left(F(\mathbf{w}^{(0)}) - F^* + \frac{\gamma}{2}\right) + \frac{2\sqrt{2}}{6}\sqrt{\gamma}\|\boldsymbol{\sigma}\|_1\Big), \quad (16)$$

where γ is a positive integer, $\tau = \left(1+\frac{2}{\beta K}\right)\frac{1}{\sqrt{\gamma}}$, and $\beta \triangleq \frac{E_0\vartheta}{\sigma_{\mathrm{n}}^2}$. More details of the convergence analysis are shown in Appendix F. The detailed convergence rate analysis can be found in Appendix G.

Acknowledgements

The work is supported in part by the National Natural Science Foundation of China under Grant 62171262; in part by Shandong Provincial Natural Science Foundation under Grant ZR2021YQ47; in part by the Taishan Young Scholar under Grant tsqn201909043; in part by Major Scientific and Technological Innovation Project of Shandong Province under Grant 2020CXGC010109.

4. Experiments

For experiments, we investigate benchmark image dataset: MNIST. We run our experiments with 31 normal EDs, and partition the training dataset according to the labels. For fair comparison, we set the same hyper-parameters (batch size as 128, local epoch as 1, and learning rate as 0.004). We compare our proposed scheme with two baseline algorithms: SignSGD, FedAvg and FSK-MV to obtain performance results.

In FSK-MV and the proposed algorithm, we configure the arrival time of the EDs signal to be different. In the experiments on the MNIST dataset, we compare in independent identically distributed (iid) as well as non-independent identically distributed (non-iid) data to demonstrate more precisely the performance advantages and disadvantages of the algorithm compared to existing designs.

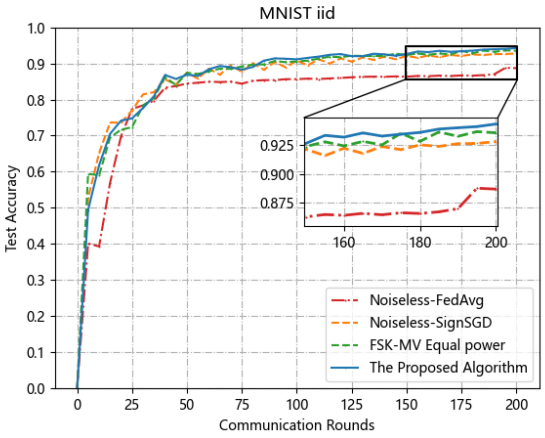

Figure 1. Test accuracy versus communication round under iid setting on the MNIST dataset.

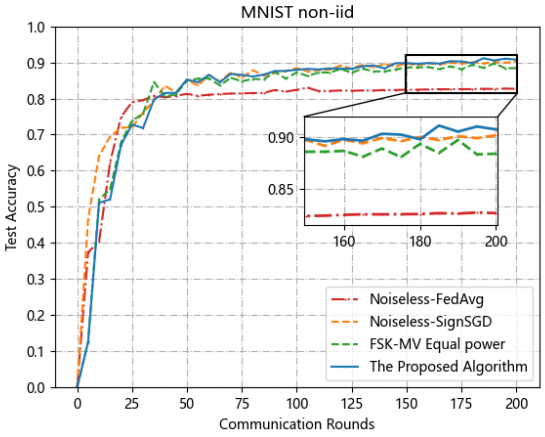

Figure 2. Test accuracy versus communication round under non-iid setting on the MNIST dataset.

In Figures 1 and 2, we provide the test accuracy results for

iid/non-iid data by taking time-synchronization errors. For two baseline algorithms, simple noise-free aggregation is not otherwise optimized, and the aggregation of parameters leads to slightly worse error accuracy than the proposed algorithm. Also, the results about the proposed scheme and FSK-MV indicate that both have a high level of test accuracy with the time synchronization error.

Due to the dynamic power control scheme, our test results are still superior to the FSK-MV even under the terrible environmental conditions. Also the usage of non-coherent detection causes high test accuracy without the utilization of CSI at the ED. The ultimate experimental results indicate that the whole scheme can mitigate the impact of time synchronization error, channel fading and noise.

References

Adeli, M. H. and Şahin, A. Multi-cell Non-coherent Over-the-air Computation for Federated Edge Learning. In *ICC 2022 - IEEE International Conference on Communications*, pp. 4944–4949, 2022.

Agarwal, N., Suresh, A. T., Yu, F., Kumar, S., and Mcmahan, H. B. cpSGD: Communication-efficient and differentially-private distributed SGD, 2018.

Akoun, J. and Meyer, S. Signsgd: Fault-tolerance to blind and byzantine adversaries, 2022.

Alistarh, D., Grubic, D., Li, J. Z., Tomioka, R., and Vojnovic, M. QSGD: Communication-efficient sgd via gradient quantization and encoding. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1707–1718, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Allen-Zhu, Z. Natasha 2: Faster Non-convex Optimization than SGD, 2018.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signSGD: Compressed optimisation for non-convex problems, 2018.

Chen, M., Gündüz, D., Huang, K., Saad, W., Bennis, M., Feljan, A. V., and Poor, H. V. Distributed learning in wireless networks: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 39 (12):3579–3605, 2021a.

Chen, M., Yang, Z., Saad, W., Yin, C., Poor, H. V., and Cui, S. A joint learning and communications framework for Federated Learning over wireless networks. *IEEE Transactions on Wireless Communications*, 20(1):269–283, 2021b.

Goldenbaum, M., Boche, H., and Stańczak, S. Harnessing interference for analog function computation in wireless sensor networks. *IEEE Transactions on Signal Processing*, 61(20):4893–4906, 2013.

Li, P., Erdol, H., Briggs, K., Wang, X., Piechocki, R., Ahmad, A., Inacio, R., Kapoor, S., Doufexi, A., and Parekh, A. Transmit power control for indoor small cells: A method based on Federated reinforcement Learning, 2022.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated Learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Liu, W., Zang, X., Li, Y., and Vucetic, B. Over-the-air Computation systems: Optimization, analysis and scaling laws. *IEEE Transactions on Wireless Communications*, 19(8):5488–5502, 2020.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2023.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706, 2019.

Sahin, A., Everette, B., and Hoque, S. S. M. Distributed Learning over a wireless network with FSK-based Majority Vote, 2021.

Shi, G., Guo, S., Ye, J., Saeed, N., and Dang, S. Multiple parallel Federated Learning via Over-the-Air Computation. *IEEE Open Journal of the Communications Society*, 3:1252–1264, 2022.

Zhu, G., Wang, Y., and Huang, K. Broadband analog aggregation for low-latency Federated Edge Learning. *IEEE Transactions on Wireless Communications*, 19(1):491–506, 2020.

Zhu, G., Du, Y., Gündüz, D., and Huang, K. One-bit Over-the-air Aggregation for communication-efficient Federated Edge Learning: Design and convergence analysis. *IEEE Transactions on Wireless Communications*, 20(3): 2120–2135, 2021.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients, 2019.

Şahin, A., Everette, B., and Hoque, S. S. M. Over-the-Air Computation with DFT-spread OFDM for Federated Edge Learning. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1886–1891, 2022.

A. Training Process of Federated Learning via Non-Coherent Over-the-air Computation

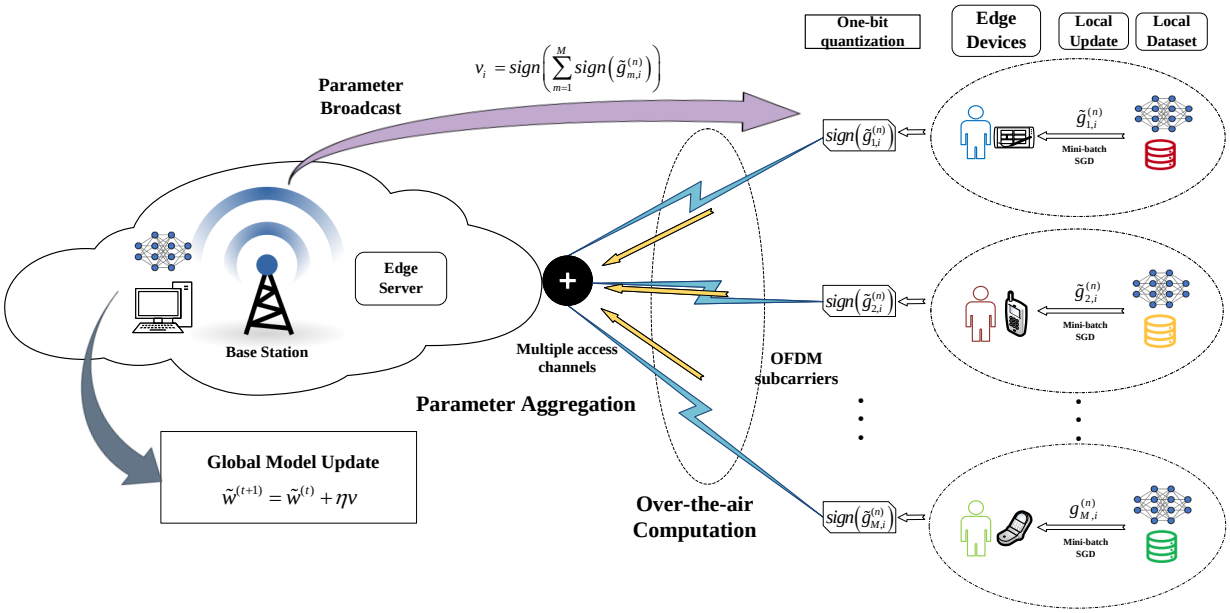

Figure 3. The FL system of the proposed scheme comprising of M EDs coordinated by an ES.

B. Details of Non-Coherent Energy Detection

After exploiting the modulation and power control scheme at the transmitter side, we complete the symbol transmission through the Aircomp scheme. Then, we can obtain the model parameters for uplink aggregation at the receiver ES side and we assume that CSI is not available at this time.

It is known that $\Delta_i^{(n)} = \mathbf{e}_i^+ - \mathbf{e}_i^-$ is the sum of the energy values for which the detection vote is 1 or -1, where $e_i^+ \triangleq |r_{l^+,m^+}^{(n)}|_2^2$ and $e_i^- \triangleq |r_{l^+,m^+}^{(n)}|_2^2$, $\forall i$. At the receiver side, it is not expected to recover the sign of the local stochastic gradient, since we need the vote summation, and do not exploit any method to resolve the interference introduced by the channel. Instead, the choice is to exploit the interference for aggregation and to compare the energy magnitude on two different subcarriers to detect the MV in (11).

Here the interference usage for aggregation means that we incorporate the transmitting power, noise power and other channel factors together into the subcarriers energy magnitude. This operation is equivalent to using the interference term as the standard content of the detection, and then performing an one-bit quantization scheme.

The Proof of Lemma 3.1: According to the above description, the specific proof details are obtained as follows. We assume that the multipath channels between the EDs and the ES are independent. To simplify the notation, we omit the index n. Since (9) is a weighted summation of independent complex Gaussian random variables with zero mean and unit variance (i.e., channel coefficients), $t_{l^+,m^+}^{(n)}$ is a zero mean random variable, where its variance is

$$\begin{aligned}\mu_i^+ = \mathbb{E}\left[e_i^+\right] = \mathbb{E}\left[\left|r_{l^+,m^+}^{(n)}\right|_2^2\right] = \mathbb{E}\left[E_0 \sum_{\bar{s}_{\forall m,i}^{(n)}=1} P_m + \sigma_{\mathrm{n}}^2\right] \\ = E_0 M_i^+ \mathbb{E}\left[P_m\right] + \sigma_{\mathrm{n}}^2 = E_0 M_i^+ \vartheta + \sigma_{\mathrm{n}}^2.\end{aligned} \tag{17}$$

The same analysis can be done for μ_i^- .

As given in Lemma 3.1, μ_i^+ and μ_i^- are linear functions of M_i^+ and M_i^-. Also, we obtain the correct MV because the symbols energy may not coherently add up. However, the detection performance depends on the parameter ϑ that has an efficient trade-off about the direction of convergence of the global model on e_i^+ and e_i^-.

C. Assumptions for The Subsequent Analysis

In this work, we consider the well-known Lipschitz continuity utilized to explain some of the assumptions. Moreover, in order to make the developed theory applicable to neural networks rather than assuming a convex loss function, we require a lower bound. Specially, it is worth noting that the minimum assumption required to guarantee convergence to the stabilization point.

Definition C.1. A function *f* is *L*-Lipschitz over a set *s* with respect to a norm $\|\cdot\|$ if there exist a real constant $L > 0$ such that $\|f(\mathbf{y}) - f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|, \forall \mathbf{x}, \mathbf{y} \in S$.

Lemma C.2. *(Lemma 1.2.3 (Allen-Zhu, 2018)). For a differentiable function $f : \mathbb{R}^Q \to \mathbb{R}$, let ∇f be* L*-Lipschitz on $\mathbb{R}^Q$ with respect to norm $\|\cdot\|_2$. Then, for any $\mathbf{y}, \mathbf{y}$ from $\mathbb{R}^Q$,*

$$\left|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^{\mathrm{T}}(\mathbf{y} - \mathbf{x})\right| \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2. \tag{18}$$

Assumption C.3. (Bounded Loss Function). For all parameter vectors $\mathbf{w}$, the lower bound of the associated loss function is some value F^*, $F(\mathbf{w}) \geq F^*, \forall \mathbf{w}$.

Assumptions C.3 and C.4 as follow, on the Lipschitz smoothness and bounded variance, respectively, are standard in the stochastic optimization literature.

Assumption C.4. (Smoothness). Let $\mathbf{g}$ denote the gradient of the loss function $F(\mathbf{w})$ evaluated at $\mathbf{w}$. For all $\mathbf{w}$ and $\mathbf{w}'$, the expression from (21) is given by

$$|F(\mathbf{w}') - (F(\mathbf{w}) - \mathbf{g}^{\mathrm{T}}(\mathbf{w}' - \mathbf{w}))| \leq \frac{1}{2}\sum_{i=1}^{q} L_i \left(w_i' - w_i\right)^2, \tag{19}$$

where we can assume that there exists a vector of non-negative constants $\mathbf{L} = [L_1, \ldots, L_q]^{\mathrm{T}}$.

Assumption C.5. (Variance bound). The stochastic gradient estimates $\tilde{\mathbf{g}}_k = [\tilde{g}_{k,1}, \ldots, \tilde{g}_{k,q}]^{\mathrm{T}} = \nabla F_k\left(\mathbf{w}^{(n)}\right), \forall k$ are independent and unbiased estimates of $\mathbf{g} = [\mathbf{g}_1, \ldots, \mathbf{g}_q]^{\mathrm{T}} = \nabla F(\mathbf{w})$ (the true gradient) with a coordinate bounded variance, i.e.,

$$\mathbb{E}\left[\tilde{\mathbf{g}}_k\right] = \mathbf{g}, \forall k, \tag{20}$$

$$\mathbb{E}\left[\left(\tilde{g}_{k,i} - g_i\right)^2\right] \leq \sigma_i^2/d_{\mathrm{b}}, \forall k, i, \tag{21}$$

where $\sigma = [\sigma_1, \ldots, \sigma_q]^{\mathrm{T}}$ is a non-negative constant vector, $\tilde{g}_{k,i}$ and g_i denote the *i*th element of $\tilde{\mathbf{g}}_k$ and $\mathbf{g}$.

After these analysis above, another significant assumption is that the data-stochasticity induced gradient noise. Meanwhile, this assumption causes the discrepancy between $\tilde{\mathbf{g}}_k$ and $\mathbf{g}$, which is unimodal and symmetric as verified by experiments in (Bernstein et al., 2018) and formally described as follow.

Assumption C.6. (Unimodal, symmetric gradient noise). For any given $\mathbf{w}$, each elements of the vector $\tilde{\mathbf{g}}_k, \forall k$, has a unimodal distribution that is also symmetric around its mean.

At this time, it can be obviously appreciated that Gaussian noise is a special case. Noting that even for small batches of other magnitudes, we expect the central limit theorem to be in effect and to bring the typical gradient noise distribution close to a Gaussian distribution.

Assumption C.7. (Independent, identical, and unbiased gradients). The local stochastic gradient estimates are independent and unbiased, i.e., $\mathbb{E}_{\tilde{\mathcal{D}}_m}\left[\tilde{g}_{m,i}^{(t)}\right] = g_i^{(n)}, \forall m, i$.

Assumption C.8. (Exponential probability distribution). For given M_i^+ and M_i^-, e_i^+ and e_i^- are exponential random variables, where their means are μ_i^+ and μ_i^-, respectively.

Assumption C.7 does not claim that the local stochastic gradients are unbiased estimates of the global gradients. Therefore, they are accommodated to heterogeneous data distribution scenarios where the sum of local stochastic gradients is unbiased.

D. The proof of Lemma 3.2

We consider the establishment of equivalent mathematical events $\operatorname{sign}(\Delta_i^{(n)}) = \operatorname{sign}(g_i^{(n)})$ described by a well-defined random variable with known distribution. To address this, assume that $\operatorname{sign}(g_i^{(n)}) = 1$ as a standard. Let *X* denote the number of edge devices with correct sign at the *i*th element of the gradient vector, namely, with $\operatorname{sign}(\tilde{g}_{m,i}^{(n)}) = \operatorname{sign}(g_i^{(n)})$, i.e., $\operatorname{sign}(g_i^{(n)}) = 1$. For the scenario, the random variable *X* can then be model as the sum of *K* independent Bernoulli trials, and binomial with success probability and failure probability denoted by:

$$p_i \triangleq \mathbb{P}\left[\operatorname{sign}\left(\tilde{g}_{m,i}^{(n)}\right) = \operatorname{sign}\left(g_i^{(n)}\right)\right], \tag{22}$$

and

$$q_i \triangleq \mathbb{P}\left[\operatorname{sign}\left(\tilde{g}_{m,i}^{(n)}\right) \neq \operatorname{sign}\left(g_i^{(n)}\right)\right], \tag{23}$$

where $p_i + q_i = 1$. They are intuitively determined by the randomness of the data.

The Proof of Error Probability: Regarding the bounds on stochasticity-induced error, we mainly deal with the term of P_i^{err}. Through the previous descriptions, the following treatment is available for (14), for all m. This implies that

$$P_i^{\text{err}} = \sum_{M_i^+=0}^{M} \mathbb{P}\left[\operatorname{sign}\left(\Delta_i^{(n)}\right) \neq 1 \mid X = M_i^+\right] \mathbb{P}\left[X = M_i^+\right]. \tag{24}$$

According to the properties of Bernoulli distribution, the second term on the right side of the equation can be expressed as follow:

$$\mathbb{P}\left[X = K_i^+\right] = \begin{pmatrix} K \\ K_i^+ \end{pmatrix} p_i^{K_i^+} q_i^{K-K_i^+}. \tag{25}$$

To calculate $\mathbb{P}\left[\operatorname{sign}(\Delta_i^{(n)}) \neq 1 \mid X = M_i^+\right]$, we need to derive the probability problem in dynamic power control, which can be shown as

$$\begin{aligned} &P\left(\operatorname{sign}(\Delta_i) \neq \operatorname{sign}(g_i^{(n)})\right) = p_i^{\text{err}}, \\ &P\left(\operatorname{sign}(\Delta_i) = \operatorname{sign}(g_i^{(n)})\right) = 1 - p_i^{\text{err}}, \\ &P\left(\operatorname{sign}(\tilde{g}_{k,i}^{(n)}) \neq \operatorname{sign}(q_i^{(n)})\right) = q_i, \\ &P\left(\operatorname{sign}(\tilde{g}_{k,i}^{(n)}) = \operatorname{sign}(g_i^{(n)})\right) = 1 - q_i. \end{aligned} \tag{26}$$

Also, we can obtain the probability of the dynamic power control term as $[(1-p_i^{er})(1-q_i)] - (p_i^{err} \cdot q_i) = (1 - p_i^{err} - q_i)$. Based on the similarity analysis in (Şahin et al., 2022), it is simple to prove that e_i^+ and e_i^- are exponential random variables. And we derive the dynamic power control term in terms of energy detection (i.e., probabilistic representation) as

$$\mathbb{P}\left[\operatorname{sign}(\Delta_i^{(n)}) \neq 1 \mid X = M_i^+\right] = \frac{(1 - P_i^{err} - q_i) K_i^- E_0 \vartheta + \sigma_n^2}{K E_0 \vartheta + 2\sigma_n^2}. \tag{27}$$

Then, further combining the terms in (27) to simplify, an upper bound on P_i^{err} can be completed as

$$\begin{aligned} P_i^{err} &\leq \sum_{K_i^+=0}^{K} \frac{K_i^-(1-q_i)E_0\vartheta + \sigma_n^2}{KE_0\vartheta + 2\sigma_n^2} \cdot \begin{pmatrix} K \\ K_i^+ \end{pmatrix} p_i^{K_i^+} q_i^{K-K_i^+} \\ &= \sum_{K_i^+=0}^{K} \frac{\left(K - K_i^+\right) \cdot p_i E_0\vartheta + \sigma_n^2}{KE_0\vartheta + 2\sigma_n^2} \cdot \begin{pmatrix} K \\ K_i^+ \end{pmatrix} p_i^{K_i^+} q_i^{K-K_i^+}. \end{aligned} \tag{28}$$

Analysis of (28) shows that the right-hand side of the inequality can be split into two parts and solved for separately by utilizing the properties of binomial coefficients.

(1) The-first-part can be obtained as

$$\begin{aligned}
&\sum_{K_i^+=0}^{K} \frac{\left(K-K_i^+\right) p_i E_0 \vartheta}{K E_0 \vartheta+2\sigma_n^2}\left(\begin{array}{c} K \\ K_i^+ \end{array}\right) p_i^{K_i^+} q_i^{K-K_i^+} \\
=&\frac{K \cdot p_i}{K+2/\beta}\left(p_i+q_i\right)^K-\sum_{K_i^+=0}^{K} \frac{p_i \cdot K_i^+}{k+2/\beta}\left(\begin{array}{c} K \\ K_i^+ \end{array}\right) p_i^{K_i^+} q_i^{K-K_i^+} \\
=&\frac{K p_i}{K+2/\beta}-\frac{p_i E\left(K_i^+\right)}{K+2/\beta}=\frac{K p_i\left(1-p_i\right)}{K+2/\beta}=\frac{K\left(1-q_i\right) q_i}{K+2/\beta}.
\end{aligned} \tag{29}$$

(2) The-second-part can be obtained as

$$\begin{aligned}
&\sum_{K_i^+}^{K} \frac{\sigma_n^2}{K E_0 \vartheta+2\sigma_n^2}\left(\begin{array}{c} K \\ K_i^+ \end{array}\right) p_i^{K_i^+} q_i^{K-K_i^+} \\
&\quad=\frac{1/\beta}{K+2/\beta}\left(\begin{array}{c} K \\ K_i^+ \end{array}\right) p_i^{K_i^+} q_i^{k-k_i^+} \\
&\quad=\frac{1/\beta}{K+2/\beta}\left(p_i+q_i\right)^K=\frac{1/\beta}{K+2/\beta}.
\end{aligned} \tag{30}$$

Thus, this upper bound on P_i^{err} can be obtained after combining (29) with (30) as

$$P_i^{err} \leq \frac{K q_i\left(1-q_i\right)+1/\beta}{K+2\beta}. \tag{31}$$

Expecting to obtain a more accurate upper bound, we further scale for $q_i\left(1-q_i\right)$, and an established fact is $q_i<1/2$.

$$y(a)=a(1-a)-b a \quad \left(0<a<\tfrac{1}{2}, 0<b<1\right), \tag{32}$$

where a is substituted with q_i, and b represents the slope of the function. Then we need to process (32) with $y'(a)=(1-b)-2a=0$, and get $a=(1-b)/2 \in \left(0, \frac{1}{2}\right)$. Under the analysis of function monotonicity, it is necessary to ensure that $y(a)>0$ holds, which gives $0<b<1/2$. Ultimately, we can rewrite (31) in the form as follow:

$$P_i^{err} \leq \frac{K q_i/2}{K+2/\beta}+\frac{1/\beta}{K+2/\beta}. \tag{33}$$

To proceed with, we need a bound on q_i that can be associated with the signal-to-noise ratio of a component of the stochastic gradient as S_i, which is defined in Lemma 3.2.

Under the unimodal symmetric gradient noise assumption mentioned in Assumption C.6, we can obtain the following bound on q_i. The failure probability for the sign bit of a single ED is exploited by the following Lemma D.1.

Lemma D.1. *(Failure probability under conditions of unimodal symmetric gradient noise). Based on several previous assumptions, the failure probability satisfies:*

$$\begin{aligned}
q_i &= \mathbb{P}\left[\operatorname{sign}\left(\tilde{g}_{k,i}^{(n)}\right) \neq \operatorname{sign}\left(g_i^{(n)}\right)\right] \\
&\leq \begin{cases} \frac{2}{9} \frac{1}{R_i^2}, & \text{if } R_i>\frac{2}{\sqrt{3}} \\ \frac{1}{2}-\frac{R_i}{2 \sqrt{3}}, & \text{otherwise,} \end{cases}
\end{aligned} \tag{34}$$

which is less than $1/2$ *for all cases.*

The proof process relies on the properties of certain probability distributions, which is captured in Appendix E. Under the symmetry assumption, by exploiting the derivations in Lemma D.1, $q_i \leq \frac{\sqrt{2}\sigma_i}{3|g_i^{(n)}|\sqrt{n_{\mathrm{b}}}}$ still holds true. Then, we combine the upper bound on q_i with (33) to complete the proof of Lemma 3.2.

E. The Proof of Lemma D.1

The Proof of Failure Probability: Under the background of Assumption C.5 and Assumption C.6, for a unimodal symmetric random variable Y with mean μ and variance σ^2, the following Gauss' inequality holds:

$$\mathbb{P}[|Y-\mu|>y] \leq \begin{cases} \frac{4}{9}\frac{\sigma^2}{y^2}, & \text{if } \frac{y}{\sigma} > \frac{2}{\sqrt{3}} \\ 1-\frac{y}{\sqrt{3}\sigma}, & \text{otherwise.} \end{cases} \tag{35}$$

Then applying symmetry followed by Gauss' inequality, the failure probability can be obtained by

$$\begin{aligned} &\mathbb{P}\left[\operatorname{sign}(\tilde{g}_{k,i}^{(n)}) \neq \operatorname{sign}(g_i)\right] = \mathbb{P}\left[\tilde{g}_{k,i}^{(n)} - g_i \geq |g_i|\right] \\ &= \frac{1}{2}\mathbb{P}\left[|\tilde{g}_{k,i}^{(n)} - g_i| \geq |g_i|\right] \\ &\leq \begin{cases} \frac{2}{9}\frac{\sigma_i^2}{d_b|g_i^{(n)}|^2}, & \text{if } \frac{|g_i^{(n)}|}{\sigma_i/\sqrt{n_b}} > \frac{2}{\sqrt{3}} \\ \frac{1}{2}-\frac{|g_i^{(n)}|}{2\sqrt{3}\sigma_i/\sqrt{d_b}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{36}$$

Eventually, we complete the proof, which is utilised to infer Lemma 3.2.

F. The Proof of Theorem 3.4

Note that the OFDM symbols may non-coherently add up and their amplitudes may not be aligned in fading channel. Hence, the MV calculated in (11) is different from the original MV given in (3). Then, we provide a complete overview of the convergence performance. Firstly, we define the convergence rate (Zhu et al., 2021) as the rate at which the expected value of average norm of the gradient of $F(\mathbf{w})$ diminishes as the number of total communication rounds *N* and *M*, when the training is done in the presence of the proposed scheme.

The Proof of Convergence Rate Analysis: The proof is carried out following widely-adopted strategy of the norm of gradient with respect to the expected improvement made in a single algorithm step. And we compare this with total possible improvement under Assumption C.3.

To begin with, we firstly bound the improvement of the objective for the data-stochasticity induced noise based on Assumption C.4. For processing, we exploit the contents of (4) in our inference process as a way to decompose (the data and the channel) stochasticity-induced error that we need to analyze. Thus, by utilising Assumption C.4 and (4), we can write as

$$\begin{aligned} F(\mathbf{w}^{(n+1)}) - F(\mathbf{w}^{(n)}) \leq {\mathbf{g}^{(n)}}^{\mathrm{T}} \left(\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}\right) \\ + \frac{1}{2}\sum_{i=1}^{q} L_i \left(w_i^{(n+1)} - w_i^{(n)}\right)^2. \end{aligned} \tag{37}$$

Then, we exploit (4) to make a substitution with $(\tilde{v}_i^{(n)})^2 = 1$ whether it is $+1$ or -1. Therefore, we have

$$\begin{aligned} F(\mathbf{w}^{(n+1)}) - F(\mathbf{w}^{(n)}) &\leq -\eta {\mathbf{g}^{(n)}}^{\mathrm{T}} \tilde{\mathbf{v}}^{(n)} + \frac{1}{2}\sum_{i=1}^{q} L_i (\eta \tilde{v}_i^{(n)})^2 \\ &= -\eta \left\|\mathbf{g}^{(n)}\right\|_1 \operatorname{sign}\left(\Delta_i^{(n)}\right) + \frac{\eta^2}{2}\|\mathbf{L}\|_1. \end{aligned} \tag{38}$$

The first term on the right side of the equation can be analyzed to know that the value of $\operatorname{sign}(\cdot)$ cannot be determined, so

the term will have a randomness error. Thus, we then proceed to obtain as

$$\begin{aligned} F(\mathbf{w}^{(n+1)}) - F(\mathbf{w}^{(n)}) &\leq -\eta \mathbf{g}^{(n)^{\mathrm{T}}} \mathbf{v}^{(n)} + \frac{\eta^2}{2}\|\mathbf{L}\|_1 \\ &= -\eta \|\mathbf{g}^{(n)}\|_1 + \frac{\eta^2}{2}\|\mathbf{L}\|_1 \\ &\quad + 2\eta \sum_{i=1}^{q} |\mathbf{g}_i^{(n)}| \mathbb{I}\left[\operatorname{sign}\left(\Delta_i^{(n)}\right) \neq \operatorname{sign}\left(g_i^{(n)}\right)\right]. \end{aligned} \tag{39}$$

Thus, the expected improvement of the left term can be written as an inequality as follows.

$$\begin{aligned} \mathbb{E}\left[F(\mathbf{w}^{(n+1)}) - F(\mathbf{w}^{(n)}) \mid \mathbf{w}^{(n)}\right] &\leq -\eta \|\mathbf{g}^{(n)}\|_1 + \frac{\eta^2}{2}\|\mathbf{L}\|_1 \\ &\quad + \underbrace{2\eta \sum_{i=1}^{q} |\mathbf{g}_i^{(n)}| \underbrace{\mathbb{P}\left[\operatorname{sign}\left(\Delta_i^{(n)}\right) \neq \operatorname{sign}\left(g_i^{(n)}\right)\right]}_{\triangleq P_i^{err}}}_{\text{Stochasticity-induced error}}. \end{aligned} \tag{40}$$

For the analysis of the above equation, the main challenge is to obtain an upper bound on the error term in (40). The bound is a function of the stochasticity of the local gradients and the detection performance of the proposed scheme.

Accordingly, based on Lemma 3.1 and several definitionins Theorem 3.4, an upper bound on the stochasticity-induced error can be represented by the proof related to Appendix D as follows:

$$\begin{aligned} \sum_{i=1}^{q} |g_i^{(n)}| p_i^{err} &\leq \sum_{i=1}^{q} |g_i^{(n)}| \frac{\frac{K}{2} \cdot \sqrt{2}/(3R_i)}{K + 2/\beta} \\ &\quad + \sum_{i=1}^{q} |g_i^{(n)}| \frac{1/\beta}{K + 2/\beta}. \end{aligned} \tag{41}$$

Then, substituting Lemma 3.2 into (41), this specific upper bound is written as

$$\begin{aligned} \sum_{i=1}^{q} |g_i^{(n)}| p_i^{err} &\leq \sum_{i=1}^{q} \frac{\frac{K}{2} \cdot |g_i^{(n)}|}{K + 2/\beta} \cdot \frac{\sqrt{2}|\sigma_i^{(n)}|}{3|g_i^{(n)}|\sqrt{d_b}} \\ &\quad + \sum_{i=1}^{q} |g_i^{(n)}| \frac{1/\beta}{K + 2/\beta}. \end{aligned} \tag{42}$$

Now, after a series of collations, we can obtain as follow:

$$\sum_{i=1}^{q} |g_i^{(n)}| p_i^{err} \leq \frac{K}{K + 2/\beta} \cdot \frac{\sqrt{2}}{6\sqrt{d_b}} \|\sigma\|_1 + \frac{1/\beta}{K + 2/\beta} \|g^{(n)}\|_1. \tag{43}$$

Then, under considering Assumption C.3, we perform a telescoping sum over the iterations and calculate the expectation

over the randomness in the trajectory as

$$\begin{aligned}
&F(\mathbf{w}^{(0)}) - F^* \geq F(\mathbf{w}^{(0)}) - F(\mathbf{w}^{(N)}) \\
=&\mathbb{E}\left[\sum_{n=0}^{N-1} F(\mathbf{w}^{(n)}) - F(\mathbf{w}^{(n+1)})\right] \\
\geq& E\left[\sum_{n=0}^{N-1}\left(\left(\eta - 2\eta\cdot\frac{1/\beta}{K+2/\beta}\right)\|g^{(n)}\|_1 - \frac{\eta^2}{2}\|L\|_1\right.\right. \\
&\left.\left. -\frac{2\eta\cdot K}{K+2/\beta}\cdot\frac{\sqrt{2}}{6\sqrt{d_b}}\cdot\|\sigma\|_1\right)\right] \\
=& E\left[\sum_{n=0}^{N-1}\left(\frac{K\eta\cdot\|g^{(n)}\|_1}{K+2/\beta} - \frac{\eta^2\|L\|_1}{2} - \frac{2\sqrt{2}K\eta\cdot\|\sigma\|_1}{6(K+2/\beta)\sqrt{d_b}}\right)\right].
\end{aligned} \tag{44}$$

In order to derive the term of required convergence rate, we rearrange (44) and use the expressions for d_b and η, while conducting a series of simplifications to obtain as follow:

$$\begin{aligned}
E\left[\frac{1}{N}\sum_{n=0}^{N-1}\|g^{(n)}\|_1\right] \leq& (1+\frac{2}{K\beta})\frac{\sqrt{\gamma}}{2\sqrt{N}}\cdot\sqrt{\|L\|_1} \\
&+(1+\frac{2}{K\beta})\cdot\frac{\sqrt{\|L\|_1}\sqrt{N}}{N\sqrt{\gamma}}(F(w^{(0)}) - F^*(w)) + \frac{2\sqrt{2}\sqrt{\gamma}\|\sigma\|_1}{6\sqrt{N}}).
\end{aligned} \tag{45}$$

Finally, by replacing pertinent term with some expressions in Theorem 3.4, the proof process is completed and (16) is reached.

G. The Relevant Conclusions of Convergence Rate

Based on the above derivation, proof process and the visual representation of Theorem 3.4, we can infer the followings:

- In connection with a larger SNR (i.e., a larger $1/\sigma_n^2$) and a large number of EDs (i.e., a larger *K*), the convergence rate with FSK-MV in fading channel improves since τ decreases.
- The dynamic power control account for a better convergence in this scenario rate since ϑ increases with a state. It is higher equivalence degree of the sign between edge device gradient parameters and aggregation gradient parameters.
- Under ideal power control, the convergence rate becomes similar to SignSGD in an ideal channel [(Zhu et al., 2020), Theorem 1] gradually.

Note that the proposed scheme has no impact in terms of convergence, since the interference of channel fading and noise is used for energy aggregation by the usage of a non-coherent detection scheme at the receiver side.

H. Performance Comparison about The Proposed Scheme

The discussion in the above part of this section has adequately addressed the actual performance performance regarding such problems. And there is a stark contrast related to the problems that can be solved in the uplink and downlink communications of FL. The communication issues solved are mainly focused on the following two areas:

- **Resistance to Time-Varying Fading Channel:** In contrast to the approaches in (Zhu et al., 2020), the proposed scheme does not utilize the CSI for TCI at the EDs. It is therefore compatible with time-varying channels and does not lose the gradient information that we need during transmission owing to TCI. Because of the advantages mentioned above, a shown trade-off is that it quadruples the amount of time-frequency resources required in AirComp compared to OBDA in (Zhu et al., 2021). More importantly, with addition of dynamic power control, interference from fading channels and noise can be addressed effectively.

- **Resistance to Time-Synchronization Errors:** It is worth noting that our holistic comprehensive scheme greatly enhances the resistance capability to time-synchronization errors. The obvious reason is that the time misalignment among the EDs or the uncertainty in receiver synchronization within the CP window lead to phase rotation in the frequency domain. while the comprehensive scheme does not encode information on the amplitude or phase, it does not use any channel-dependent information in the EDs and ES. Therefore, the considered process is more robust to time-synchronization errors compared to OBDA scheme.