

Imperfect CSI: A Key Factor of Uncertainty to Over-the-Air Federated Learning

Jiacheng Yao¹, Graduate Student Member, IEEE, Zhaohui Yang², Member, IEEE,
Wei Xu³, Senior Member, IEEE, Dusit Niyato⁴, Fellow, IEEE, and Xiaohu You, Fellow, IEEE

Abstract—Over-the-air computation (AirComp) has recently been identified as a prominent technique to enhance communication efficiency of wireless federated learning (FL). This letter investigates the impact of channel state information (CSI) uncertainty at the transmitter on an AirComp enabled FL (AirFL) system with the truncated channel inversion strategy. To characterize the performance of the AirFL system, the weight divergence with respect to the ideal aggregation is analytically derived to evaluate learning performance loss. We explicitly reveal that the weight divergence deteriorates as $\mathcal{O}(1/\rho^2)$ as the level of channel estimation accuracy ρ vanishes, and also has a decay rate of $\mathcal{O}(1/K^2)$ with the increasing number of participating devices, K . Building upon our analytical results, we formulate the channel truncation threshold optimization problem to adapt to different ρ , which can be solved optimally. Numerical results verify the analytical results and show that a lower truncation threshold is preferred with more accurate CSI.

Index Terms—Federated learning (FL), over-the-air computation (AirComp), imperfect channel state information (CSI).

I. INTRODUCTION

FEDERATED learning (FL), a distributed machine learning paradigm, has been regarded as a promising technique to support ubiquitous intelligence in the beyond fifth-generation (B5G) wireless networks [1], [2]. In a wireless FL system, the distributed devices, orchestrated by a parameter server (PS), iteratively train a shared learning model through the exchange of model parameters rather than the raw data, thereby protecting data privacy [3], [4]. However, due to the frequent uplink transmissions of model parameters from a large number of devices, the communication overhead and latency of FL become excessively high, which hinders its deployment in resource-constrained wireless networks.

Manuscript received 1 September 2023; accepted 20 September 2023. Date of publication 25 September 2023; date of current version 11 December 2023. This work was supported in part by NSFC under Grant 62022026 and Grant 62211530108; in part by the Fundamental Research Funds for the Central Universities under Grant 2242022k60002 and Grant 2242023K5003; and in part by the Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2023A03. The work of Zhaohui Yang was supported in part by the Zhejiang Lab Program under Grant K2023QA0AL02, and in part by the Zhejiang Science and Technology Program under Grant 2023C01021. The associate editor coordinating the review of this article and approving it for publication was S. Kafaie. (Corresponding author: Wei Xu.)

Jiacheng Yao, Wei Xu, and Xiaohu You are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jcyao@seu.edu.cn; wxu@seu.edu.cn; xhyu@seu.edu.cn).

Zhaohui Yang is with the Research Institute of Intelligent Networks, Zhejiang Lab, Hangzhou 311121, China, and also with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, China (e-mail: yang_zhaohui@zju.edu.cn).

Dusit Niyato is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 308232 (e-mail: dniyato@ntu.edu.sg).

Digital Object Identifier 10.1109/LWC.2023.3318555

To facilitate communication-efficient FL design, over-the-air computation (AirComp) has been greatly adopted for effective uplink model transmission [5], [6], [7]. By exploiting the waveform superposition nature of multiple access (MAC) channels, simultaneous model transmission and over-the-air model aggregation can be achieved, which can reduce the communication latency and save the uplink communication bandwidth substantially. In [5], a truncated channel inversion scheme was proposed to combat deep fadings in an AirComp-aided FL (AirFL), and the fundamental trade-offs between communication and learning was discussed. Then in [6], the power control strategy was further optimized to alleviate the impacts brought by AirComp errors. Considering the constraint of limited wireless communication resources, device selection and power control were jointly optimized to minimize the accuracy loss for AirFL in [7].

However, most of the existing AirFL scheme design and resource allocation optimization optimistically assumed the availability of perfect channel state information (CSI) at the transmitter, which is hardly to acquire in practice especially in a wireless network. More importantly, unlike traditional communication systems, the CSI imperfection in the AirFL system brings a severe impact. To be concrete, considering transmit power constraints, the users in deep fading should be truncated and therefore not participate in AirComp. Moreover, to achieve the uniform model aggregation, channel inversion should be performed at the transmitter. Considering the CSI uncertainties, the model aggregation is perturbed due to the imperfect truncation decision and channel inversion, resulting in the deterioration in learning performance. In [8], the authors considered a bounded CSI error and analyzed the impact of imperfect CSI on the convergence rate of FL. However, few effort has been endeavored to explicitly analyze in theory the aggregation distortion and accuracy loss brought by CSI uncertainty. To the best of our knowledge, there is no theoretical guidance on channel truncation strategy of imperfect CSI.

Against the above backgrounds, we focus on an AirFL system adopting the truncated channel inversion scheme, where only partial CSI is available at the PS. We theoretically characterize the aggregation distortion due to imperfect CSI and the corresponding channel truncation, and derive an upper bound of the weight divergence of the aggregated gradient to evaluate the degradation of learning performance. Our results unravel that as the level of channel estimation accuracy ρ decreases, the weight divergence enlarges at the order of $1/\rho^2$. The analytical result further suggests that increasing the number of participating devices, K , help decrease the weight divergence as $\mathcal{O}(1/K^2)$ and can completely eliminate the impact of CSI imperfection. Moreover, based on the derived analytical results, we derive the optimal truncation threshold as a function of channel estimation uncertainty and system SNR. Numerical tests are conducted to verify the effectiveness of performance analysis and truncation threshold optimization.

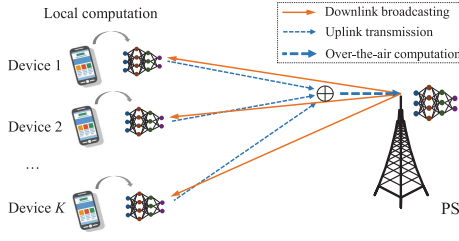


Fig. 1. An architecture of AirFL with one PS and K devices.

II. SYSTEM MODEL OF AIRFL

A. Federated Learning Model

We consider a typical FL algorithm, where a shared machine learning model is trained via the collaboration between a central PS and K distributed devices. Let \mathcal{D}_k denote the local dataset owned by the k th device. The local loss function of model parameters, \mathbf{w} , at the device k is defined as $F_k(\mathbf{w}, \mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|} \sum_{\mathbf{u} \in \mathcal{D}_k} \mathcal{L}(\mathbf{w}, \mathbf{u})$, where \mathbf{u} is a data sample and $\mathcal{L}(\mathbf{w}, \mathbf{u})$ represents the sample-wise loss function. Without loss of generality, we assume that the size of all local datasets is the same, i.e., $|\mathcal{D}_k| = D, \forall k$. Then, the global loss function over all the datasets is given by

$$F(\mathbf{w}) = \frac{1}{K} \sum_{k=1}^K F_k(\mathbf{w}, \mathcal{D}_k). \quad (1)$$

The goal of the FL is to find the optimal model parameters, denoted by \mathbf{w}^* , to minimize the global loss function in (1).

To effectively handle this problem, we apply the widely used FL algorithm in [8]. Specifically, in the m th round of the FL algorithm, the PS firstly broadcasts the up-to-date global parameter \mathbf{w}_m to all devices. With the received global model \mathbf{w}_m and their local datasets, each device runs a stochastic gradient descent (SGD) algorithm on a local mini-batch to compute the local gradient, which follows

$$\mathbf{g}_m^k \triangleq \nabla F_k(\mathbf{w}_m, \mathcal{D}_{k,m}) = \frac{1}{|\mathcal{D}_{k,m}|} \sum_{\mathbf{u} \in \mathcal{D}_{k,m}} \mathcal{L}(\mathbf{w}_m, \mathbf{u}), \quad (2)$$

where $\mathcal{D}_{k,m}$ is the mini-batch selected from \mathcal{D}_k . Next, all devices report the local gradients in (2) to the PS. Upon receiving all the local gradients, PS performs the update as

$$\mathbf{w}_{m+1} = \mathbf{w}_m - \eta \mathbf{g}_m, \quad (3)$$

where η denotes the learning rate and

$$\mathbf{g}_m \triangleq \frac{1}{K} \sum_{k=1}^K \mathbf{g}_m^k. \quad (4)$$

The FL algorithm iterates (2) and (3) until convergence.

B. Over-the-Air Computation for FL

In practice, we adopt the AirComp method for model aggregation in wireless networks, shown in Fig. 1. We express the channel between the k th device and the PS as $d_k^{-\frac{\alpha}{2}} h_k$, where d_k denotes the distance between the PS and device k , α is the large scale path loss exponent, and h_k represents the small-scale fading of the channel. Assume that the channels are independent Rayleigh fading channels, i.e., $h_k \sim \mathcal{CN}(0, 1)$. In general, the small-scale fading of the channel cannot be perfectly estimated at devices. Denote the channel estimate

at device k by \hat{h}_k and a relationship between h_k and \hat{h}_k can be modelled as $h_k = \rho \hat{h}_k + \sqrt{1 - \rho^2} v_k$, where $\rho \in (0, 1]$ represents the correlation coefficient between h_k and \hat{h}_k , and $v_k \sim \mathcal{CN}(0, 1)$ is the error independent of \hat{h}_k . Note that ρ directly corresponds to the level of channel estimation accuracy and $\rho = 1$ implies the availability of perfect CSI. To overcome the negative impact of deep fading, a must truncated channel inversion scheme is headed for the uplink transmission [5]. To be concrete, only when $|\hat{h}_k|^2$ exceeds a predetermined threshold, γ_{th} , the device is activated to transmit its gradient to PS. Accordingly, the received signal at the PS follows

$$\mathbf{y} = \sum_{k \in \mathcal{S}_m} d_k^{-\frac{\alpha}{2}} h_k \beta_k \mathbf{g}_m^k + \mathbf{z}_m, \quad (5)$$

where \mathcal{S}_m represents the set of activated devices, β_k is the pre-processing factor for device k , and $\mathbf{z}_m \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I})$ is the additive Gaussian noise with power σ^2 . Given the imperfect CSI, the pre-processing factor for device k is chosen as $\beta_k = \frac{\zeta \lambda d_k^{\alpha/2} \hat{h}_k^*}{K |\hat{h}_k|^2}$ [7], where ζ is a scaling factor for ensuring the transmit power constraint and λ is a compensation constant for ensuring unbiasedness of the gradient estimation. For simplicity, we consider the uniform transmit power budget P_{max} at each device and choose the factor ζ to guarantee

$$\mathbb{E} \left[\left\| \beta_k \mathbf{g}_m^k \right\|^2 \right] \leq P_{max}. \quad (6)$$

At the receiver, by scaling \mathbf{y} with $\frac{1}{\zeta}$ and taking the real part, an estimate of the actual gradient in (4) is given by

$$\hat{\mathbf{g}}_m = \frac{1}{K} \sum_{k=1}^K \xi_k \mathbf{g}_m^k + \bar{\mathbf{z}}_m, \quad (7)$$

where $\bar{\mathbf{z}}_m \triangleq \frac{\Re\{\mathbf{z}_m\}}{\zeta}$ is the equivalent noise, and ξ_k is given by

$$\xi_k = \begin{cases} \lambda \frac{\Re\{h_k^* \hat{h}_k\}}{|\hat{h}_k|^2} & |\hat{h}_k|^2 \geq \gamma_{th}, \\ 0 & |\hat{h}_k|^2 < \gamma_{th}. \end{cases} \quad (8)$$

By comparing (7) and (4), the distortion in the gradient estimation comes from two aspects, i.e., the coefficient distortion ξ_k caused by the imperfect CSI, and the scaled additive Gaussian noise $\bar{\mathbf{z}}_m$. Also, we notice that the expectation of ξ_k determines whether the gradient estimation is unbiased, and the variance of ξ_k and $\bar{\mathbf{z}}_m$ measure the gradient estimation distortion, which brings notable deterioration in convergence performance [4].

III. PERFORMANCE ANALYSIS AND OPTIMIZATION

In this section, we theoretically capture the impact of the imperfect CSI on the performance of AirFL. Based on the analytical results, we further optimize the truncation threshold γ_{th} with respect to ρ and system SNR.

A. Performance Analysis for AirFL

Firstly, we need to determine the value of the compensation constant λ for achieving an unbiased gradient estimation.

Lemma 1: In order to ensure the unbiasedness of gradient transmission, the compensation constant truncation and imperfect CSI is chosen by $\lambda = \frac{e^{\gamma_{th}}}{\rho}$.

Proof: Please refer to Appendix A. ■

Then, to facilitate the performance analysis for AirFL, the following Lemma 2 derives the variance of AirComp parameters ξ_k , which directly reflects the mean squared error (MSE) of AirComp [9].

Lemma 2: The variance of ξ_k with unit mean is given by

$$\mathbb{E}[(\xi_k - 1)^2] = e^{\gamma_{\text{th}}} - \frac{1 - \rho^2}{2\rho^2} \text{Ei}(-\gamma_{\text{th}}) e^{2\gamma_{\text{th}}} - 1, \quad (9)$$

where $\text{Ei}(\cdot)$ denotes the exponential integral function.

Proof: Please refer to Appendix B. ■

Remark 1: Note that $\lim_{\gamma_{\text{th}} \rightarrow 0} \text{Ei}(\gamma_{\text{th}}) = -\infty$. Hence, from (9), regardless of the level of channel estimation accuracy, the truncation is necessary to avoid unbounded variance.

Remark 2: It is worth noting that for other power control schemes of AirComp, e.g., that in [6], the method of theoretical analysis still applies through treating them as special truncated channel inversion schemes. Without loss of generality, we therefore consider the most commonly adopted truncated channel inversion scheme as a general analysis.

Next, to evaluate the accuracy loss caused by the imperfect model aggregation, we choose a popular performance metric as the expected weight divergence with respect to $\hat{\mathbf{g}}_m$ and \mathbf{g}_m [10], defined by $\Delta^2 = \mathbb{E}[\|\hat{\mathbf{g}}_m - \mathbf{g}_m\|^2]$. It is worth noting that Δ^2 corresponds to the MSE of the gradient estimation at the PS and directly reflects the accuracy of gradient estimation via the AirComp, which determines the convergence performance. To pave the way for performance analysis, we make the following widely used assumption [6].

Assumption: The stochastic gradients on random batches are uniformly bounded, i.e., $\mathbb{E}[\|\mathbf{g}_m^k\|^2] \leq G^2$. And the obtained global gradient, \mathbf{g}_m , is unbiased and variance bounded, i.e.,

$$\mathbb{E}[\mathbf{g}_m] = \nabla F(\mathbf{w}_m), \quad \mathbb{E}[\|\mathbf{g}_m - \nabla F(\mathbf{w}_m)\|] \leq \delta^2. \quad (10)$$

Then, the weight divergence can be accurately characterized under this general assumption.

Theorem 1: The weight divergence, Δ^2 , is bounded by

$$\Delta^2 \leq \frac{G^2}{K^2} \left(e^{\gamma_{\text{th}}} - \frac{1 - \rho^2}{2\rho^2} \text{Ei}(-\gamma_{\text{th}}) e^{2\gamma_{\text{th}}} - 1 + \frac{\sigma^2 \max_k \{d_k^\alpha\} e^{2\gamma_{\text{th}}}}{2P_{\max} \rho^2 \gamma_{\text{th}}} \right). \quad (11)$$

Proof: Please refer to Appendix C. ■

Remark 3: According to (11), the imperfect channel estimation deteriorates the learning performance with the order of $\frac{1}{\rho^2}$. It validates our statement that the accurate channel estimation is a key to the AirFL system.

Remark 4: Especially for high SNR regime, i.e., $\frac{P_{\max}}{\sigma^2} \rightarrow \infty$, the weight divergence Δ^2 is dominated by the impact of imperfect CSI rather than noise. It implies that, for a given level of channel estimation accuracy, the accuracy loss caused by imperfect CSI can no longer be compensated by increasing the transmit power while only weakens the impact of the noise. This is the key observation that is different from the impact of CSI error in pure communication systems for data recovery.

Remark 5: By direct inspection of (11), as the increase of the number of devices, K , the weight divergence decreases as $\mathcal{O}(1/K^2)$ and eventually tends towards zero, i.e., the impact of imperfect CSI is completely eliminated. This phenomenon can be qualitatively explained by the law of Large Numbers,

that is, the randomness of aggregation distortion is eliminated when the participating devices tend to be infinite many.

Then, starting from (11), the convergence performance of FL is characterized in the following theorem.

Theorem 2: Suppose the loss function F is L -Lipschitz with respect to \mathbf{w} and the learning rate satisfies $\eta < \frac{2}{L}$. The convergence of FL algorithm is bounded by

$$\begin{aligned} & \frac{1}{M} \sum_{m=0}^{M-1} \mathbb{E}[\|\nabla F(\mathbf{w}_m)\|^2] \\ & \leq \frac{1}{M} \left(\frac{F(\mathbf{w}_0) - \mathbb{E}[F(\mathbf{w}_M)]}{\eta - \frac{L\eta^2}{2}} + \frac{ML\eta(\Delta^2 + \delta^2)}{2 - L\eta} \right). \end{aligned} \quad (12)$$

Proof: Please refer to the online version [11, Appendix D]. ■

This theorem implies that the convergence is guaranteed with sufficiently large M and the expected gradient norm converges to $\frac{L\eta(\Delta^2 + \delta^2)}{2 - L\eta}$, which linearly increases with respect to Δ^2 .

B. Optimization of the Truncation Threshold

Under the guarantee of unbiased estimation in Lemma 1 and the general assumption of gradients, we note that the impacts of learning algorithms and the wireless transmission are decoupled. Hence, unlike some existing works [6], [12], the truncation threshold optimization can be isolated from the specific learning parameters, thus being defined by

$$\max_{\gamma_{\text{th}} > 0} h(\gamma_{\text{th}}) \triangleq e^{\gamma_{\text{th}}} - k_1 \text{Ei}(-\gamma_{\text{th}}) e^{2\gamma_{\text{th}}} + k_2 \frac{e^{2\gamma_{\text{th}}}}{\gamma_{\text{th}}}, \quad (13)$$

where $k_1 \triangleq \frac{1 - \rho^2}{2\rho^2}$, and $k_2 \triangleq \frac{\sigma^2 \max_k \{d_k^\alpha\}}{2P_{\max} \rho^2}$ are positive constants.

Theorem 3: The objective function in (13) is convex.

Proof: Please refer to Appendix D. ■

Based on convexity of $h(\cdot)$, the optimal value of γ_{th} can be easily obtained from a bisection method with low complexity. Specifically, the derivative of $h(x)$ is

$$h'(x) = e^x - k_1 \frac{e^x}{x} - 2k_1 \text{Ei}(-x) e^{2x} + k_2 e^{2x} \frac{2x - 1}{x^2}. \quad (14)$$

Since $\lim_{x \rightarrow 0} h'(x) < 0$ and $\lim_{x \rightarrow \infty} h'(x) > 0$, the unique zero point of $h'(x)$, i.e., the optimal solution of γ_{th} , can be found through the bisection search.

IV. SIMULATION RESULTS

In this section, we provide simulation results to verify the performance analysis and truncation threshold optimization. We train a multi-layer perceptron (MLP) on the popular MNIST dataset via the AirFL algorithm. The distance d_k is uniformly distributed over (0, 500) m. Unless otherwise specified, the other parameters are set as: $K = 10$, $\alpha = 2.2$, $P_{\max} = 0.1$ W, $\sigma^2 = -40$ dBm, and $\eta = 0.005$.

Fig. 2 depicts the numerical variance of ξ_k obtained from Monte-Carlo simulations, compared with the theoretical result in (9). It shows that the numerical results matches well with the theoretical results, which verifies our analysis. Moreover, the channel estimation accuracy level, ρ , imposes more significant impacts on the variance than the truncation threshold.

In Fig. 3, we evaluate the impact of truncation threshold optimization on learning performance and compare it with

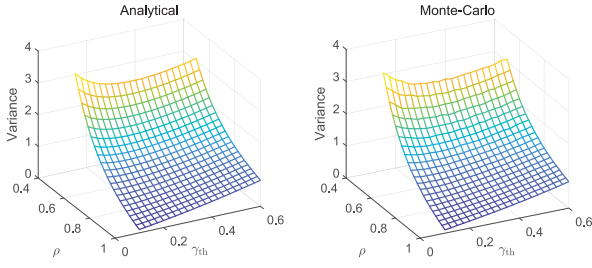
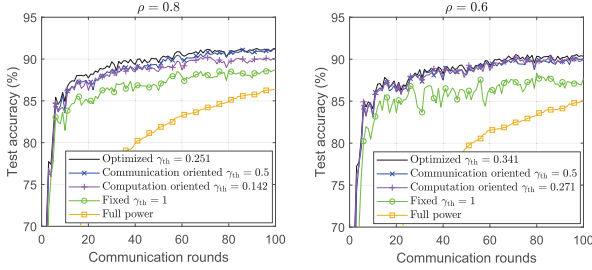
Fig. 2. Variance versus ρ and γ_{th} .

Fig. 3. Test accuracy versus different truncation thresholds.

other schemes. The four benchmark schemes are described as: “Communication oriented” and “Computation oriented” schemes represent γ_{th} is optimized to minimize the noise related and computation related term in (11), respectively; “Fixed γ_{th} ” represents the truncation threshold is set as a constant [5]; The “Full power” scheme represents the transmitter does not perform power control and only compensates channel phase offset [13]. For all the tested setups, the proposed optimization method outperforms all the benchmarks due to the joint consideration of communication and computation. It is observed that the test accuracy first improves and then deteriorates with γ_{th} . This is because with the increase of γ_{th} , the performance is first limited by noise and then limited by less participating devices. Also, for larger ρ , a lower truncation threshold is preferred. Moreover, compared with the full power scheme, the proposed power control strategy successfully alleviates the impact of data heterogeneity, leading to a much prominent performance gain.

V. CONCLUSION

In this letter, we theoretically analyzed the performance of AirFL with imperfect CSI and optimized a channel truncation strategy. The analytical results revealed the importance of accurate channel estimation for AirFL. Our results can also be extended to performance analysis for other power control schemes of AirFL and the joint optimization of communication and computation remains an challenging research direction.

APPENDIX A PROOF OF LEMMA 1

To determine λ , we start with the expectation of ξ_k , expressed as

$$\mathbb{E}[\xi_k] = \lambda \mathbb{E} \left[\frac{\Re\{h_k^* \hat{h}_k\}}{|\hat{h}_k|^2} \middle| |\hat{h}_k|^2 \geq \gamma_{th} \right] \Pr\{|\hat{h}_k|^2 \geq \gamma_{th}\}, \quad (15)$$

which should be equal to 1 to guarantee an unbiased gradient estimation in (7). Considering that h_k and its estimate \hat{h}_k are correlated, we first introduce a new random variable to tackle

with this difficulty, which follows

$$x \triangleq \frac{1}{\sqrt{1-\rho^2}} \left(\frac{\Re\{h_k^* \hat{h}_k\}}{|\hat{h}_k|^2} - \rho \right) = \frac{\Re\{v_k^* \hat{h}_k\}}{|\hat{h}_k|^2}, \quad (16)$$

where v_k and \hat{h}_k are uncorrelated Gaussian variables with zero mean and unit variance. Then, we have

$$\mathbb{E} \left[x \middle| |\hat{h}_k|^2 \geq \gamma_{th} \right] = \mathbb{E} \left[\frac{v_k^* \hat{h}_k + v_k \hat{h}_k^*}{2|\hat{h}_k|^2} \middle| |\hat{h}_k|^2 \geq \gamma_{th} \right] = 0. \quad (17)$$

Then, by comparing (15) and (16), through some linear transformations, we arrive at $\mathbb{E}[\xi_k] = \lambda e^{-\gamma_{th}} \rho = 1$, which implies that $\lambda = e^{\gamma_{th}} / \rho$ and the proof completes.

APPENDIX B PROOF OF LEMMA 2

We derive the variance of ξ_k by using the form of conditional expectation as

$$\begin{aligned} \mathbb{E}[(\xi_k - 1)]^2 &= \mathbb{E} \left[\left(\frac{\Re\{h_k^* \hat{h}_k\}}{|\hat{h}_k|^2 \rho} - 1 \right)^2 \middle| |\hat{h}_k|^2 \geq \gamma_{th} \right] \\ &\quad \times \Pr\{|\hat{h}_k|^2 \geq \gamma_{th}\} + \Pr\{|\hat{h}_k|^2 < \gamma_{th}\} \\ &= \frac{e^{\gamma_{th}}(1-\rho^2)}{\rho^2} \mathbb{E}[(x-c)^2 | y \leq -\gamma_{th}] + 1 - e^{-\gamma_{th}}, \end{aligned} \quad (18)$$

where $y \triangleq -|\hat{h}_k|^2$ and $c \triangleq \frac{\rho(1-e^{\gamma_{th}})}{e^{\gamma_{th}}\sqrt{1-\rho^2}}$. To calculate the conditional expectation, we first need to find the joint distribution of x and y . The joint cumulative distribution function (CDF) of x and y equals

$$\begin{aligned} F_{xy}(t, \gamma) &= \Pr \left\{ \frac{\Re\{v_k^* \hat{h}_k\}}{|\hat{h}_k|^2} < t, -|\hat{h}_k|^2 < \gamma \right\} \\ &= \Pr \left\{ v_k^* \hat{h}_k + v_k \hat{h}_k^* - 2t|\hat{h}_k|^2 < 0, -|\hat{h}_k|^2 < \gamma \right\} \\ &= \Pr \left\{ \mathbf{z}^H \mathbf{A}_1 \mathbf{z} < 0, \mathbf{z}^H \mathbf{A}_2 \mathbf{z} < \gamma \right\}, \end{aligned} \quad (19)$$

where $\mathbf{z} \triangleq [\hat{h}_k, v_k]^H$, and

$$\mathbf{A}_1 \triangleq \begin{bmatrix} -2t & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{A}_2 \triangleq \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}. \quad (20)$$

According to [14, eq. (3.2c.5)], the joint moment generating function (MGF) of the two quadratic forms, $z_1 \triangleq \mathbf{z}^H \mathbf{A}_1 \mathbf{z}$ and $z_2 \triangleq \mathbf{z}^H \mathbf{A}_2 \mathbf{z}$, follows

$$\mathcal{M}_{z_1, z_2}(s_1, s_2) = \det(\mathbf{I} - s_1 \mathbf{A}_1 - s_2 \mathbf{A}_2)^{-1}. \quad (21)$$

Applying the inverse Laplace transformation, we express the probability in (19) as

$$\begin{aligned} &\Pr \left\{ \mathbf{z}^H \mathbf{A}_1 \mathbf{z} < 0, \mathbf{z}^H \mathbf{A}_2 \mathbf{z} < \gamma \right\} \\ &= \frac{1}{(2\pi i)^2} \int_{\epsilon_1 - i\infty}^{\epsilon_1 + i\infty} \int_{\epsilon_2 - i\infty}^{\epsilon_2 + i\infty} \frac{e^{\gamma s_2}}{s_1 s_2} \mathcal{M}_{z_1, z_2}(s_1, s_2) ds_2 ds_1 \\ &\stackrel{(a)}{=} \frac{1}{2\pi i} \int_{\epsilon_1 - i\infty}^{\epsilon_1 + i\infty} \frac{1}{s_1(1+2ts_1-s_1^2)} ds_1 \\ &\quad - \frac{1}{2\pi i} \int_{\epsilon_1 - i\infty}^{\epsilon_1 + i\infty} \frac{1}{s_1(1+2ts_1-s_1^2)} e^{-(1+2ts_1-s_1^2)\gamma} ds_1 \end{aligned}$$

$$\begin{aligned}
& \stackrel{(b)}{=} \frac{t + \sqrt{1+t^2}}{2\sqrt{1+t^2}} + \frac{1}{2\pi i} \int_{\epsilon-i\infty}^{\epsilon+i\infty} \frac{(t + \sqrt{t^2+s+1})e^{\gamma s}}{2s(s+1)\sqrt{t^2+s+1}} ds \\
& \stackrel{(c)}{=} \frac{t}{2\sqrt{1+t^2}} \left(1 - \text{Erf} \left(\sqrt{-\gamma(1+t^2)} \right) U(-\gamma) \right) \\
& \quad + \frac{e^{\gamma}}{2} \text{Erfc}(-\sqrt{-\gamma t}) U(-\gamma), \tag{22}
\end{aligned}$$

where (a) comes from Eq. (5.2.4) in supplements of [15], (b) exploits the Cauchy's residue theorem and $s \triangleq s_1^2 - 2ts_1 - 1$, (c) is due to Eq. (5.3.7) in supplements of [15], $\text{Erf}(\cdot)$, $\text{Erfc}(\cdot)$, and $U(\cdot)$ represent the error function, the complementary error function and the Heaviside function, respectively. From (19) and (22) and by taking the derivative of joint CDF, the joint probability density function (PDF) of x and y equals

$$f_{xy}(t, \gamma) = \sqrt{-\frac{\gamma}{\pi}} e^{\gamma(1+t^2)} U(-\gamma). \tag{23}$$

From (23), we calculate the conditional expectation in (18) as

$$\begin{aligned}
\mathbb{E}[(x-c)^2 | y \leq -\gamma_{\text{th}}] &= \int_{-\infty}^{\infty} \frac{\int_{-\infty}^{-\gamma_{\text{th}}} f_{xy}(t, \gamma) dt}{\Pr\{y \leq -\gamma_{\text{th}}\}} d\gamma \\
&= e^{\gamma_{\text{th}}} \int_{-\infty}^{-\gamma_{\text{th}}} \int_{-\infty}^{\infty} (t-c)^2 \sqrt{-\frac{\gamma}{\pi}} e^{\gamma(1+t^2)} dt d\gamma \\
&\stackrel{(a)}{=} e^{\gamma_{\text{th}}} \int_{-\infty}^{-\gamma_{\text{th}}} \frac{2c^2\gamma - 1}{2\gamma} e^{\gamma} d\gamma \\
&\stackrel{(b)}{=} c^2 - \frac{1}{2} \text{Ei}(-\gamma_{\text{th}}) e^{\gamma_{\text{th}}}, \tag{24}
\end{aligned}$$

where (a) exploits [16, eqs. (3.462.8) and (3.321.1)] and the fact that $\int_{-\infty}^{\infty} te^{\gamma t^2} dt = 0$. The equality in (b) comes from the definition of the exponential integral function, $\text{Ei}(\cdot)$. Applying (24) into (18), we complete the proof.

APPENDIX C PROOF OF THEOREM 1

The weight divergence is reformulated as

$$\begin{aligned}
\mathbb{E}[\|\hat{\mathbf{g}}_m - \mathbf{g}_m\|^2] &= \mathbb{E} \left[\left\| \frac{1}{K} \sum_{k=1}^K (\xi_k - 1) \mathbf{g}_m^k + \bar{\mathbf{z}}_m \right\|^2 \right] \\
&\stackrel{(a)}{=} \frac{1}{K^2} \sum_{k=1}^K \mathbb{E}[(\xi_k - 1)^2] \mathbb{E}[\|\mathbf{g}_m^k\|^2] + \mathbb{E}[\|\bar{\mathbf{z}}_m\|^2], \tag{25}
\end{aligned}$$

where (a) is due to the zero mean and independence between $\xi_k - 1$, $\forall k$. As for the noise term, recall that $\bar{\mathbf{z}}_m = \frac{\Re\{\mathbf{z}_m\}}{\zeta}$ and we have $\mathbb{E}[\|\bar{\mathbf{z}}_m\|^2] = \frac{\sigma^2}{2\zeta^2}$. According to the transmit power constraint in (6), the scaling factor ζ must satisfy

$$\max_{k \in \mathcal{S}_m} \left\{ \frac{\zeta^2 \lambda^2 d_k^\alpha}{K^2 |\hat{h}_k|^2} \mathbb{E}[\|\mathbf{g}_m^k\|^2] \right\} \leq P_{\max}. \tag{26}$$

Note that for all $k \in \mathcal{S}_m$, we have $|\hat{h}_k|^2 \geq q\gamma_{\text{th}}$. Combining (26) with the value of λ and the bound assumption of $\mathbb{E}[\|\mathbf{g}_m^k\|^2] \leq qG^2$, ζ is set as $\zeta = \frac{K\rho\sqrt{P_{\max}\gamma_{\text{th}}}}{G \max_k \{d_k^{\alpha/2}\} e^{\gamma_{\text{th}}}}$. Then, combining all the derived results, we obtain (11) and complete the proof.

APPENDIX D PROOF OF THEOREM 3

We check the second derivative of $h(x)$ as

$$\begin{aligned}
h''(x) &= e^x + \frac{k_1 e^x}{x^2} (-4x^2 e^x \text{Ei}(-x) - 3x + 1) \\
&\quad + \frac{2k_2 e^x (2x^2 - 2x + 1)}{x^3}. \tag{27}
\end{aligned}$$

It is obvious that the first and the third terms in (27) are positive for positive x . To prove the nonnegativity of the second term, we use the inequalities [17, eq. (5.1.20)]

$$-\text{Ei}(-x) > \frac{1}{2} e^{-x} \ln \left(1 + \frac{2}{x} \right) > e^{-x} \frac{1}{x+1}, \tag{28}$$

which yields to $-4x^2 e^x \text{Ei}(-x) - 3x + 1 > \frac{(x-1)^2}{x+1} \geq 0$. Then, we conclude that $h''(x) > 0$ for $x > 0$ and hence $h(x)$ is convex.

REFERENCES

- [1] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 1, pp. 9–39, Jan. 2023.
- [2] G. Zhu et al., "Pushing AI to wireless network edge: An overview on integrated sensing, communication, and computation towards 6G," *Sci. China Inf. Sci.*, vol. 66, pp. 1–19, Feb. 2023.
- [3] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, Mar. 2021.
- [4] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020.
- [5] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.
- [6] X. Cao, G. Zhu, J. Xu, and S. Cui, "Transmission power control for over-the-air federated averaging at network edge," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1571–1586, May 2022.
- [7] W. Guo, R. Li, C. Huang, X. Qin, K. Shen, and W. Zhang, "Joint device selection and power control for wireless federated learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2395–2410, Aug. 2022.
- [8] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Mar. 2021.
- [9] W. Zhang, J. Xu, W. Xu, X. You, and W. Fu, "Worst-case design for RIS-aided over-the-air computation with imperfect CSI," *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 2136–2140, Sep. 2022.
- [10] J. Yao, Z. Yang, W. Xu, M. Chen, and D. Niyato, "GoMORE: Global model reuse for resource-constrained wireless federated learning," *IEEE Wireless Commun. Lett.*, vol. 12, no. 9, pp. 1543–1547, Sep. 2023.
- [11] J. Yao, Z. Yang, W. Xu, D. Niyato, and X. You, "Imperfect CSI: A key factor of uncertainty to over-the-air federated learning," 2023, *arXiv:2307.12793*.
- [12] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, Aug. 2021.
- [13] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "BEV-SGD: Best effort voting SGD against Byzantine attacks for analog-aggregation-based federated learning over the air," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18946–18959, Oct. 2022.
- [14] A. Mathai and S. B. Provost, *Quadratic Forms in Random Variables: Theory and Applications*. New York, NY, USA: Marcel Dekker, 1992.
- [15] A. Polyanin and A. Manzhirov, *Handbook of Integral Equations*, 2nd ed., New York, NY, USA: Taylor & Francis, 2008.
- [16] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. New York, NY, USA: Academic Press, 2000.
- [17] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. New York, NY, USA: Academic, 1972.