Robust Over-the-Air Federated Learning

Hwanjin Kim, Hongjae Nam, and David J. Love

Elmore Family School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN, 47907, USA
Email: {*kim4466, nam86, djlove*}@purdue.edu

2024 58th Annual Conference on Information Sciences and Systems (CISS) | 979-8-3503-6929-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISS59072.2024.10480195

***Abstract*—Interest continues to grow in using federated learning (FL) for a variety of signal processing and communications applications. This paper focuses on a robust design for FL to mitigate the effects of noise and fading channels. To enhance the efficiency of FL in bandwidth-limited environments, over-the-air (OTA) computation has been proposed based on the superposition property of a wireless multiple-access channel (MAC). However, OTA FL inherently faces challenges with channel noise and wireless channel fading in the wireless MAC, which could degrade optimization procedure and significantly reduce the accuracy of the trained model. To tackle this challenge, we introduce a novel approach using a Kalman filter (KF)-based OTA FL algorithm in this paper.**

***Index Terms*—Federated learning, over-the-air computation, Kalman filter**

I. Introduction

Conventional machine learning (ML) techniques typically involve a centralized process, where a data center conducts the training of ML models by utilizing data obtained from edge devices. Federated learning (FL) is a promising ML framework for distributed edge learning applications where multiple clients collaborate under the coordination of a central server (CS) [1], [2]. In FL systems, the local ML models are trained on each device using their local data; then, the local model parameters are uploaded to the CS for global aggregation. However, the aggregation of local updates on edge devices becomes a pivotal role, and the uplink rate limitations remain a notable bottleneck when employing orthogonal transmission [3].

Over-the-air (OTA) computation is a fast model aggregation framework, accomplished through the utilization of the signal superposition characteristic of a wireless multiple-access channel (MAC) [4]–[7]. In contrast to traditional multiple-access methods that view other transmitted signals as interference, OTA computation utilizes co-channel interference as a contributor in the computational process, leading to improved spectral efficiency and reduced communication latency [8]. OTA computation harnesses the ability to combine multiple signals on the same channel simultaneously, which can maximize the overall data throughput and minimize communication latency. Without decoding each device's individual data, the CS attempts to calculate the desired aggregation signal using the data from all edge devices. To enable a model aggregation scheme based on OTA computation, the computation distortion needs to be minimized [9].

Most prior work on OTA computation in FL systems assumes perfect channel state information (CSI), which is impractical [10]–[13]. Also, the effect of channel fading and noise at the CS should be compensated at the transmitter to achieve a reliable and high-performing OTA FL system [14]–[16]. By assuming channel reciprocity, the CSI can be acquired both at the CS and the edge devices through channel estimation. However, this approach inherently poses the risk of channel estimation errors. Moreover, this problem becomes more challenging in time-varying channels [17].

Recognizing the limitations of existing OTA FL systems, our research shifts focus towards a more realistic scenario where CSI is modeled as time-varying and must be imperfectly tracked. Addressing these challenges, we propose a robust OTA FL scheme based on channel estimation using a Kalman filtering to accommodates the imperfect nature of CSI in the time-varying channel [18]–[21]. Then, we develop a robust OTA FL scheme by reducing the mean square error (MSE) of distortion of desired aggregation signal through minimization based on the estimated channels. The numerical results reveal that the proposed robust OTA FL scheme exhibits only a slight performance difference compared to the scenario with perfect CSI.

The remainder of the paper is structured as outlined below. In Section II, we explain a system model and an optimization problem of OTA FL system. In Section III, we propose the Kalman filter-based channel estimator and explain the MSE optimization problem with the estimated channel. In Section VI, we proposed the robust OTA FL scheme based on the worst-case CSI error. Section V presents numerical results to validate our algorithm, followed by conclusions in Section VI.

Notation: Upper-case and lower-case bold letters are used to represent matrices and column vectors, respectively. $\mathbf{A}^{\mathrm{H}}$ stands for the conjugate transpose of the matrix $\mathbf{A}$, and a^* represents the conjugate of the scalar variable a. $\mathbb{E}[\cdot]$ represents the expectation, and $\mathrm{diag}(\cdot)$ denotes the diagonal matrix. $|\cdot|$ denotes the amplitude of the scalar, and $\|\cdot\|$ is the norm of the vector. $\mathbb{C}^{m\times n}$ indicates the set of all complex matrices of size $m \times n$. $\mathbf{0}$ is the all zero vector, and $\mathbf{I}_m$ is the identity matrix of size $m \times m$. $\mathcal{CN}(m, \sigma^2)$ represents the complex normal distribution having the average value of m and the variance of σ^2.

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00245280); in part by the National Science Foundation (NSF) under Grant EEC1941529, Grant CNS2225578, and Grant CNS2212565; in part by the Office of Naval Research (ONR) under Grant N000142112472.

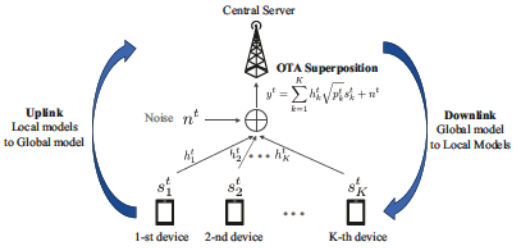

Fig. 1. An illustration of OTA FL system. The local devices train each local ML model, then transmit the model parameter to the CS. The CS receives the aggregated signal through the OTA superposition. Finally, the global model is transmitted to each local device from the CS.

II. System Model

The use of ML algorithms has significantly increased. However, training them on a single machine demands an excessive amount of computations, substantial memory requirements, and considerably delayed training time. Therefore, recent studies have focused on distributed and FL algorithms to address the challenges posed by massive datasets at the mobile edge, with the goals of enhancing privacy, managing limited bandwidth, and reducing computational costs. We regard the FL scenario with a set of distributed devices $\mathcal{K} = \{1, 2, .., K\}$ and a central server connected over a wireless MAC as illustrated in Fig. 1. Here, $\mathcal{D} = \{\mathbf{x}(i) : i = 1, \cdots, N\}$ represents the dataset used in the training phase. The local dataset at k-th device is denoted as $\mathcal{D}_k \subseteq \mathcal{D}$. The local loss function in the k-th device is given as

$$F_k(\boldsymbol{w}^t) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}_k} f(\boldsymbol{w}^t; \mathbf{x}_k^t(i)), \tag{1}$$

where $\boldsymbol{w}^t$ denotes the model parameters at the t-th time slot and f is the loss function. Accordingly, the global model is formed by aggregating the local models, and the global loss function is minimized by

$$\min_{\boldsymbol{w}^t} F(\boldsymbol{w}^t) = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} F_k(\boldsymbol{w}^t). \tag{2}$$

The OTA computation, an analog approach, enables rapid model aggregation through the exploitation of the superposition characteristic inherent in wireless channels via computing a nomographic function [4]. The OTA computation allows local edge devices to upload their model updates simultaneously over fading channels. We assume that both the CS and local devices are equipped with only one antenna. The received superposition signal at the CS is

$$y^t = \sum_{k=1}^{K} h_k^t \sqrt{p_k^t} s_k^t + n^t, \tag{3}$$

where h_k^t is the channel gain from the k-th device to the CS at t-th time slot, p_k^t is the transmit power, s_k^t is transmitted symbol, and $n^t \sim \mathcal{CN}(0, \sigma_n^2)$ is the additive complex Gaussian noise. We assume that the transmitted symbols s_k^t are independent for different devices with zero mean and unit variance.

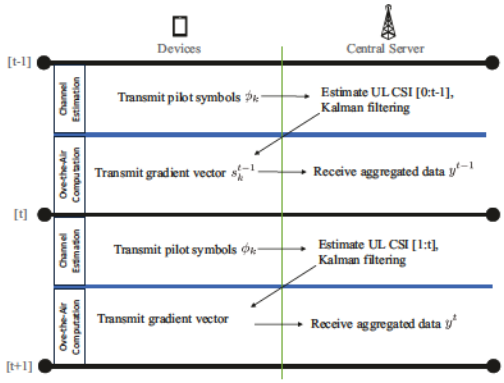

Fig. 2. The overall procedures of robust OTA FL with the Kalman filter-based channel estimation. In the channel estimation phase, the CS estimates the channel using the Kalman filtering based on the pilot transmission. Then, the each local device transmits their local model parameter to the CS in the OTA computation phase.

For the channels between the CS and the devices, we assume a block-fading model where the channel remains time-invariant within each coherence interval.

To model the time-varying channel, h_k^t is assumed to follow the Gauss-Markov process

$$h_k^t = \alpha_k h_k^{t-1} + \sqrt{1 - \alpha_k^2} z_k^t, \tag{4}$$

where α_k is the temporal correlation coefficient and $z_k^t \sim \mathcal{CN}(0, \sigma_z^2)$ represents the innovation process.

The estimated aggregation signal at the CS is

$$\hat{s}^t = \frac{r^t}{K} y^t = \frac{r^t}{K} \sum_{k=1}^{K} h_k^t \sqrt{p_k^t} s_k^t + \frac{r^t}{K} n^t, \tag{5}$$

where r^t is the receive scaling factor. When we have perfect CSI in OTA FL systems, the aim is to minimize the MSE between the desired aggregation signal $s^t = \frac{1}{K} \sum_{k=1}^{K} s_k^t$ and estimated aggregation signal $\hat{s}^t$

$$\text{MSE} = \mathbb{E}(|\hat{s}^t - s^t|^2) = \frac{1}{K^2} \sum_{k=1}^{K} \left| r^t h_k^t \sqrt{p_k^t} - 1 \right|^2 + \frac{|r^t|^2 \sigma_n^2}{K^2}. \tag{6}$$

Therefore, the optimization problem with perfect CSI subject to the sum-power constraint can be represented as

$$\begin{aligned} \min_{\{p_k^t\}, r^t} \quad & \sum_{k=1}^{K} \left| r^t h_k^t \sqrt{p_k^t} - 1 \right|^2 + |r^t|^2 \sigma_n^2 \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k^t \leq P. \end{aligned} \tag{7}$$

Most of the previous OTA FL work assumed perfect CSI, which does not hold in practice [11]–[13]. Therefore, we proposed a robust OTA FL system with a Kalman filter-based channel estimation [22]. The robust OTA FL systems consist of two phases, which are channel estimation and OTA computation phases. In the channel estimation phase,

the CS estimates the time-varying channel based on Kalman filtering using a pilot transmission from each device. Then, the local device transmits its model parameter update and the CS receives the superposition signal in the OTA computation phase. In each time slot t, the transmitted symbol s_k^t can be the gradient or model parameter of k-th device. The overall robust OTA FL system is summarized in Fig. 2.

III. Channel Estimation

In this section, we develop the channel estimation based on Kalman filtering in OTA FL systems. Using pilot transmission from each device, the CS estimates the channels from each devices. Each devices transmit orthogonal pilots, and the received signal at CS can be represented as

$$\mathbf{y}_p = \sum_{k=1}^{K} \sqrt{\rho} h_k^t \phi_k^{\mathrm{H}} + \mathbf{n}_p, \tag{8}$$

where $\mathbf{y}_p \in \mathbb{C}^{1\times\tau}$, ρ is the signal-to-noise ratio (SNR) in the channel estimation phase, $\phi_k \in \mathbb{C}^{\tau\times 1}$ is the pilot vector for device k, $\phi_k^{\mathrm{H}}\phi_{k'} = 0$ for $k \neq k'$, $\|\phi_k\| = 1$, and $\mathbf{n}_p$ is the Gaussian noise, where each element of $\mathbf{n}_p$ follows $\mathcal{CN}(0,1)$. To estimate the channel of device k, the pilot vector ϕ_k is used as

$$y_{p,k}^t = \mathbf{y}_p \phi_k = \sqrt{\rho} h_k^t + n_{p,k}^t, \tag{9}$$

where $n_{p,k}^t = \mathbf{n}_p \phi_k$.

To estimate the channel based on the Kalman filtering, we reformulate the channel model using vectorized notations as

$$\mathbf{h}^t = \boldsymbol{\alpha}\mathbf{h}^{t-1} + \boldsymbol{\beta}\mathbf{z}^t, \tag{10}$$

where $\mathbf{h}^t = [h_1^t, ..., h_K^t]^{\mathrm{T}}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_1, ..., \alpha_K)$, $\boldsymbol{\beta} = \mathrm{diag}(\beta_1, ..., \beta_K)$, where $\beta_k = \sqrt{1-\alpha_k^2}$, and $\mathbf{z}^t = [z_1^t, ..., z_K^t]^{\mathrm{T}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{V}_z)$. The measurement equation in vectorized form can be expressed as

$$\mathbf{y}_p^t = \mathbf{S}\mathbf{h}^t + \mathbf{n}_p^t, \tag{11}$$

where $\mathbf{y}_p^t = [y_{p,1}^t, ..., y_{p,K}^t]^{\mathrm{T}}$, $\mathbf{S} = \sqrt{\rho}\mathbf{I}_K$, and $\mathbf{n}_p^t = [n_{p,1}^t, ..., n_{p,K}^t]^{\mathrm{T}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{V}_n)$. The Kalman filter-based channel estimation is summarized in Algorithm 1. After applying the Kalman filter, the channel estimate of k-th device $\hat{h}_k^t$ is the k-th element of $\hat{\mathbf{h}}^{t|t}$.

The channel estimation error is defined as

$$\Delta h_k^t = h_k^t - \hat{h}_k^t. \tag{12}$$

With the estimated CSI, the optimization problem in (7) can be reformulated as

$$\begin{aligned} \min_{\{p_k^t\}, r^t} \quad & \sum_{k=1}^{K} \left| r^t (\hat{h}_k^t + \Delta h_k^t)\sqrt{p_k^t} - 1 \right|^2 + |r^t|^2 \sigma_n^2 \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k^t \leq P. \end{aligned} \tag{13}$$

Algorithm 1 Kalman Filter-Based Channel Estimation

1: Initialization:

$$\hat{\mathbf{h}}^{0|-1} = \mathbf{0}, \mathbf{M}^{0|-1} = \mathbf{I}_K$$

2: Prediction:

$$\hat{\mathbf{h}}^{t|t-1} = \boldsymbol{\alpha}\hat{\mathbf{h}}^{t-1|t-1}$$

3: Minimum prediction MSE:

$$\mathbf{M}^{t|t-1} = \boldsymbol{\alpha}\mathbf{M}^{t-1|t-1}\boldsymbol{\alpha}^{\mathrm{H}} + \boldsymbol{\beta}\mathbf{V}_z\boldsymbol{\beta}^{\mathrm{H}}$$

4: Kalman gain:

$$\mathbf{K}^t = \mathbf{M}^{t|t-1}\mathbf{S}^{\mathrm{H}}\left(\mathbf{S}\mathbf{M}^{t|t-1}\mathbf{S}^{\mathrm{H}} + \mathbf{V}_n\right)^{-1}$$

5: Correction:

$$\hat{\mathbf{h}}^{t|t} = \hat{\mathbf{h}}^{t|t-1} + \mathbf{K}^t\left(\mathbf{y}_p^t - \mathbf{S}\hat{\mathbf{h}}^{t|t-1}\right)$$

6: Minimum MSE:

$$\mathbf{M}^{t|t} = \left(\mathbf{I}_K - \mathbf{K}^t\mathbf{S}\right)\mathbf{M}^{t|t-1}$$

IV. Proposed Robust OTA FL Scheme

In this section, we propose a robust OTA FL approach based on the worst-case CSI error. We assume that the CSI error Δh_k^t is bounded[1]

$$|\Delta h_k^t| \leq \epsilon, \tag{14}$$

where ϵ is the CSI error bound. In the presence of worst-case CSI error, the optimization problem subject to sum-power constraint can be formulated as

$$\begin{aligned} \min_{\{p_k^t\}, r^t} \max_{\Delta h_k^t} \quad & \sum_{k=1}^{K} \left| r^t (\hat{h}_k^t + \Delta h_k^t)\sqrt{p_k^t} - 1 \right|^2 + |r^t|^2 \sigma_n^2 \\ \text{s.t.} \quad & \sum_{k=1}^{K} p_k^t \leq P, \\ & |\Delta h_k^t| \leq \epsilon \ \forall k. \end{aligned} \tag{15}$$

The objective of the optimization problem is non-convex, which renders global optimization more difficult. To tackle this issue, we propose an alternative optimization methodology where we fix certain parameters while optimizing the rest. When we assume that $\{p_k^t\}$ and r^t are fixed, the problem in (15) can be separated into K individual sub-problems

$$\begin{aligned} \min_{\Delta h_k^t} \quad & -\left| r^t (\hat{h}_k^t + \Delta h_k^t)\sqrt{p_k^t} - 1 \right|^2 \\ \text{s.t.} \quad & |\Delta h_k^t|^2 - \epsilon^2 \leq 0. \end{aligned} \tag{16}$$

[1]This assumption does not hold in the Kalman filtering, but we assume that the error is bounded to handle the CSI error.

We apply the Karush-Kuhn-Tucker (KKT) conditions [23] to (16)

$$\begin{aligned}\min_{\Delta h_k^t,\lambda_k}\quad & \mathcal{L}(\Delta h_k^t,\lambda_k) \qquad (17)\\ &= -\left|r^t(\hat{h}_k^t+\Delta h_k^t)\sqrt{p_k^t}-1\right|^2+\lambda_k(|\Delta h_k^t|^2-\epsilon^2)\\ \text{s.t.}\quad & \frac{\partial\mathcal{L}(\Delta h_k^t,\lambda_k)}{\partial\Delta h_k^{t\,*}}=0,\\ & |\Delta h_k^t|^2-\epsilon^2\le 0,\\ & \lambda_k(|\Delta h_k^t|^2-\epsilon^2)=0,\\ & \lambda_k\ge 0,\end{aligned}$$

where λ_k is the KKT multiplier for k-th device. The derivative of KKT condition is

$$\frac{\partial\mathcal{L}(\Delta h_k^t,\lambda_k)}{\partial\Delta h_k^{t\,*}}=-|m_k^t|^2\hat{h}_k^t+m_k^{t\,*}-|m_k^t|^2\Delta h_k^t+\lambda_k\Delta h_k^t, \qquad (18)$$

where $m_k^t=r^t\sqrt{p_k^t}$. By setting (18) to zero, the optimal solution of Δh_k^t is

$$\Delta h_k^t=\frac{|m_k^t|^2\hat{h}_k^t-m_k^{t\,*}}{\lambda_k-|m_k^t|^2}. \qquad (19)$$

Then, the objective is given by

$$\sum_{k=1}^{K}\left|\frac{m_k^t\hat{h}_k^t-1}{1-\lambda_k^{-1}|m_k^t|^2}\right|^2+|r^t|^2\sigma_n^2. \qquad (20)$$

The optimization problem in (15) can be reformulated as

$$\begin{aligned}\min_{\{m_k^t\}}\quad & \sum_{k=1}^{K}\left|\frac{m_k^t\hat{h}_k^t-1}{1-\lambda_k^{-1}|m_k^t|^2}\right|^2+|r^t|^2\sigma_n^2 \qquad (21)\\ \text{s.t.}\quad & \sum_{k=1}^{K}p_k^t\le P.\end{aligned}$$

Since the larger p_k^t leads to the smaller MSE, the power constraint can be an active, i.e., $\sum_{k=1}^{K}p_k^t=P$. Therefore, the optimization problem in (21) becomes

$$\min_{\{m_k^t\}}\quad \sum_{k=1}^{K}\left|\frac{m_k^t\hat{h}_k^t-1}{1-\lambda_k^{-1}|m_k^t|^2}\right|^2+\frac{\sigma_n^2}{P}\sum_{k=1}^{K}|m_k^t|^2. \qquad (22)$$

Then, we further decouple the above optimization problem into K sub-problems for each $k\in\{1,...,K\}$

$$\min_{m_k^t}\quad \left|\frac{m_k^t\hat{h}_k^t-1}{1-\lambda_k^{-1}|m_k^t|^2}\right|^2+\frac{\sigma_n^2}{P}|m_k^t|^2. \qquad (23)$$

In KKT conditions, we have

$$\lambda_k(|\Delta h_k^t|^2-\epsilon^2)=0. \qquad (24)$$

With $\lambda_k\neq 0$, we have

$$|\Delta h_k^t|^2=\epsilon^2 \iff \left|\frac{|m_k^t|^2\hat{h}_k^t-m_k^{t\,*}}{\lambda_k-|m_k^t|^2}\right|^2=\epsilon^2. \qquad (25)$$

When we use (25), the optimization problem in (23) becomes

$$\min_{m_k^t}\ \epsilon^2\frac{|\lambda_k|^2}{|m_k^{t\,*}|^2}+\frac{\sigma_n^2}{P}|m_k^t|^2. \qquad (26)$$

By plugging $\lambda_k>0$ based on (25), we have

$$\begin{aligned}&\min_{m_k^t}\ \epsilon^2\frac{||m_k^t|^2+|(m_k^{t\,*}-|m_k^t|^2\hat{h}_k^t)|/\epsilon|^2}{|m_k^{t\,*}|^2}+\frac{\sigma_n^2}{P}|m_k^t|^2\\ =&\min_{m_k^t}\ g(m_k^t)=\epsilon^2|m_k^t|^2+2\epsilon|m_k^{t\,*}||m_k^t\hat{h}_k^t-1|\\ &+\left||m_k^t|\hat{h}_k^t-\frac{m_k^{t\,*}}{|m_k^t|}\right|^2+\frac{\sigma_n^2}{P}|m_k^t|^2. \qquad (27)\end{aligned}$$

The optimal value of m_k^t is obtained by minimizing $g(m_k^t)$ based on the truncated Newton-conjugate gradient, an algorithm implements a search along conjugate directions and generally results in more rapid convergence compared to utilizing steepest descent directions [24], [25]. With the optimal value of m_k^t, we can get the optimal value of r^t and p_k^t using $m_k^t=r^t\sqrt{p_k^t}$ and $\sum_{k=1}^{K}p_k^t=P$. With the optimal value of r^t and p_k^t, the estimated aggregation signal $\hat{s}^t$ can be obtained.

V. Numerical Results

In this section, we present the numerical results to access the performance of the proposed robust OTA FL scheme. We assume the OTA FL system consists of a single CS and K local devices, each equipped with a single antenna. We define the SNR as P/σ_n^2. We also define the error bound as $\epsilon=\delta|h_k^t|$, where δ is the error bound coefficient. We compare the proposed scheme to the perfect CSI case. For the perfect CSI case, we employ the proposed scheme with sum-power constraint in [26]. The MSE is considered as the performance metric, using

$$\text{MSE}=\mathbb{E}(|\hat{s}^t-s^t|^2), \qquad (28)$$

where s^t is the desired aggregation signal and $\hat{s}^t$ is the estimated aggregation signal.

In our OTA FL system, we employ the MNIST dataset, which consists of 28×28 images depicting handwritten digits [27]. The MNIST dataset has 10 classes with 60,000 training and 10,000 test samples. For the classification of MNIST dataset, we use a convolution neural network (CNN). The CNN model consists of a single convolution layer with a 5×5 filter and 64 channels, followed by a max pooling layer with a 2×2 filter and strides of 2. After that, there are two fully connected layers with 128 neurons mapped by the ReLU activation layer and 10 neurons, followed by the softmax layer. In OTA FL scheme, each device has the independent and identically distributed (i.i.d.) MNIST dataset to train its local model. Then, each device transmits their local model updates to the CS. Finally, the test accuracy is evaluated through the global model.

Fig. 3 shows the MSEs of the proposed scheme with various CSI error bound and the perfect CSI case according to SNR. We set $K=8$ and $\delta=0.01,0.05,0.1,0.2$. The figure clearly shows that the proposed scheme performs better with the lower

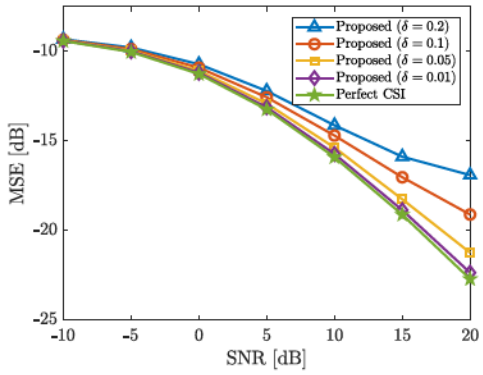

Fig. 3. MSE of proposed scheme with various CSI error bound and perfect CSI case according to SNR with $K = 8$.

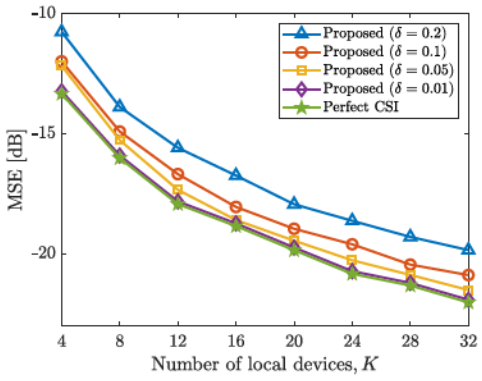

Fig. 4. MSE of proposed scheme with various CSI error bound and perfect CSI case according to number of local devices K with SNR $= 10$ dB.

error bound. With $\delta = 0.01 \approx 0$, we have the nearly the same performance of the perfect CSI case.

Fig. 4 reveals the MSEs of the proposed scheme with various CSI error bound and the perfect CSI case according to number of local devices K with SNR $= 10$ dB. The figure reveals that the MSEs of proposed scheme and perfect CSI case decrease as the number of local devices increase. Furthermore, it can be interpreted that the improvement in MSE performance is due to the noise averaging effect, which is a consequence of the increased number of local devices.

In Fig. 5, we evaluate the test accuracy of the proposed schemes and the perfect CSI case according to communication round. We set $\alpha_k = 0.95$, $K = 4$, and SNR $= 25$ dB. As the number of communication rounds increases, the test accuracy of the proposed schemes improves. The perfect CSI case assumes that the CS in each communication round knows the current CSI. In the proposed robust OTA FL scheme, the CS first estimates the current channel based on Kalman filtering. Then, the aggregation model can be obtained using the estimated channel. The proposed OTA FL schemes give substantial test accuracy even with the CSI error. Moreover, the test accuracy becomes worse when the CSI error bound is set to high.

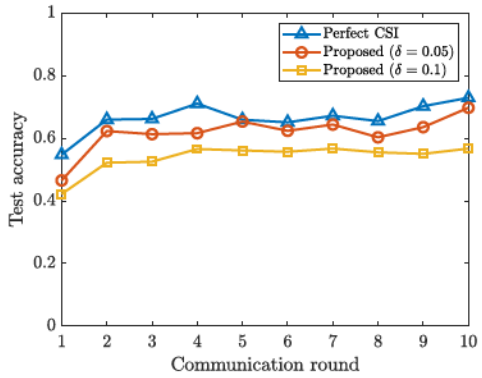

Fig. 5. Test accuracy of proposed scheme with $\delta = 0.05, 0.1$ and perfect CSI case according to communication round with SNR $= 25$ dB.

VI. Conclusion

In this paper, we proposed a robust OTA FL scheme based on the Kalman filter-based channel estimation. First, we developed the Kalman filter-based channel estimation in OTA FL system. Then, we proposed the robust OTA FL scheme by minimizing the distortion of the aggregation model. In the proposed robust OTA FL scheme, we consider a CSI error bound to handle the imperfect CSI. The numerical results showed that the proposed robust OTA FL scheme achieves comparable performance compared to the perfect CSI case in terms of the MSE performance and the test accuracy.

Potential future work involves deriving the MSE performance in closed form and implementing a multi-antenna framework for both the the CS and the local devices.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[2] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[3] 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.

[4] M. Goldenbaum, H. Boche, and S. Stańczak, "Harnessing interference for analog function computation in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 4893–4906, 2013.

[5] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.

[6] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[7] A. Abdi, Y. M. Saidutta, and F. Fekri, "Analog compression and communication for federated learning over wireless MAC," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[8] Z. Wang, Y. Zhao, Y. Zhou, Y. Shi, C. Jiang, and K. B. Letaief, "Over-the-air computation: Foundations, technologies, and applications," *arXiv:2210.10524*, 2022.

[9] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7498–7513, 2020.
[10] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3796–3811, 2021.
[11] H. Yang, P. Qiu, J. Liu, and A. Yener, "Over-the-air federated learning with joint adaptive computation and power control," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 1259–1264.
[12] E. Becirovic, Z. Chen, and E. G. Larsson, "Optimal MIMO combining for blind federated edge learning with gradient sparsification," in *2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC)*, 2022, pp. 1–5.
[13] M. Kim and D. Park, "Joint beamforming and learning rate optimization for over-the-air federated learning," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 10, pp. 13 706–13 711, 2023.
[14] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning in fading channels," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.
[15] J. Yao, Z. Yang, W. Xu, D. Niyato, and X. You, "Imperfect CSI: A key factor of uncertainty to over-the-air federated learning," *IEEE Wireless Communications Letters*, vol. 12, no. 12, pp. 2273–2277, 2023.
[16] X. Yu, B. Xiao, W. Ni, and X. Wang, "Optimal adaptive power control for over-the-air federated edge learning under fading channels," *IEEE Transactions on Communications*, vol. 71, no. 9, pp. 5199–5213, 2023.
[17] B. Tegin and T. M. Duman, "Federated learning with over-the-air aggregation over time-varying channels," *IEEE Transactions on Wireless Communications*, vol. 22, no. 8, pp. 5671–5684, 2023.
[18] C. Komninakis, C. Fragouli, A. Sayed, and R. Wesel, "Multi-input multi-output fading channel tracking and equalization using Kalman estimation," *IEEE Transactions on Signal Processing*, vol. 50, no. 5, pp. 1065–1076, 2002.
[19] J. Choi, D. J. Love, and P. Bidigare, "Downlink training techniques for FDD massive MIMO systems: Open-loop and closed-loop training with memory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 802–814, 2014.
[20] H. Kim and J. Choi, "Channel estimation for one-bit massive MIMO systems exploiting spatio-temporal correlations," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.
[21] S. G. Larew and D. J. Love, "Adaptive beam tracking with the unscented Kalman filter for millimeter wave communication," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1658–1662, 2019.
[22] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.
[23] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[24] S. G. Nash, "A survey of truncated-newton methods," *Journal of Computational and Applied Mathematics*, vol. 124, no. 1, pp. 45–59, 2000, numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037704270000426X
[25] N. Jorge and J. W. Stephen, *Numerical optimization*. Spinger, 2006.
[26] X. Zang, W. Liu, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimal design with sum-power constraint," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1524–1528, 2020.
[27] L. Deng, "The MNIST database of handwritten digit images for machine learning research [Best of the Web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.