IEEE Control Systems Letters paper presented at
2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Boosting Fairness and Robustness in Over-the-Air Federated Learning*

Halil Yigit Oksuz[1,2], Fabio Molinari[1], Henning Sprekeler[2,3], Jörg Raisch[1,2]

***Abstract*— Over-the-Air Computation is a beyond-5G communication strategy that has recently been shown to be useful for the decentralized training of machine learning models due to its efficiency. In this letter, we propose an Over-the-Air federated learning algorithm that aims to provide fairness and robustness through minmax optimization. By using the epigraph form of the problem at hand, we show that the proposed algorithm converges to the optimal solution of the minmax problem. Moreover, the proposed approach does not require reconstructing channel coefficients by complex encoding-decoding schemes as opposed to state-of-the-art approaches. This improves both efficiency and privacy.**

I. INTRODUCTION

In a traditional federated learning setting (as in [1]–[4], and some references therein), we consider a system of N agents connected to a central unit, and their objective is to accomplish a machine learning task in a decentralized manner. In a supervised learning setting, each agent i has its own local dataset represented by $D_i = \{d_i^n\}_{n=1}^{|D_i|}$, where $|D_i|$ is the number of data points and $i = 1, 2, \cdots, N$. The dataset D_i consists of pairs of inputs u_i^n and labels z_i^n, i.e., $d_i^n = (u_i^n, z_i^n)$, and the objective is to build a global parametric model that is able to predict the correct labels of the given data points. To this end, each agent i uses the private local cost function

$$g_i(\theta) = \frac{1}{|D_i|} \sum_{n=1}^{|D_i|} \mathcal{L}_i(d_i^n, \theta), \tag{1}$$

where $\mathcal{L}_i(d_i^n, \theta)$ is the error function representing the difference between the predicted output of the model with parameter θ and the actual label of the given data point d_i^n. If we define the average of all local cost functions as the global cost function $g(\theta)$, the objective of the overall system is to solve the constrained optimization problem

$$\theta^* = \arg\min_{\theta \in \Theta} g(\theta) = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} g_i(\theta), \tag{2}$$

where $\Theta \subset \mathbb{R}^m$ is a nonempty constraint set. If the central unit had access to the datasets of all agents, a centralized gradient descent-based optimization could be employed to address the global learning task at hand [5]. However, in a federated learning setting, e.g., [3], [6], [7], each agent has access only to its own local (possibly private) dataset. Having carried out the local optimization steps, they transmit the updated versions of the local parameter estimates to the central unit. Subsequently, the central unit aggregates these local parameter estimates and transmits the aggregated version to the agents for the next optimization steps.

For large-scale systems, where information exchange and cooperation are vital, one critical challenge for the averaging-based federated learning algorithms is heterogeneity [8]–[10]. When the data is heterogeneous, i.e., the fact that agents observe data from different distributions, the parameter vectors minimizing the local cost functions will in general vary significantly between different agents, and minimizing the cost function (2) may not be desirable. Instead, using a worst-case optimization problem may reflect practical requirements more accurately. Another problem that we encounter in large-scale networks is that the communication load on the overall system increases with the number of agents [11]–[14]. When multiple agents transmit information at the same time and in the same frequency band, signals are affected by the physical phenomenon of interference. Standard communication protocols[1] prevent interference by transmitting signals orthogonally. However, these techniques are not resource-efficient in the sense that they increase the need for bandwidth or the number of communication rounds, which in general leads to a decrease in total throughput and efficiency [15], [16].

In this letter, we present a federated learning algorithm that aims to improve the performance of the worst-performing agent in the system, thus providing fairness and robustness against heterogeneity. We leverage a beyond-5G communication strategy, called Over-the-Air Computation, which is more efficient as the number of agents grows [16]. Unlike the existing literature on Over-the-Air computation, e.g., [8], [17], the proposed algorithm can operate despite the inherently unknown nature of channel coefficients. We do not assume any knowledge of (nor the capability to reconstruct) channel coefficients, and therefore we will not need extra pre-processing efforts to reconstruct the channel, which makes the proposed scheme more time and resource-efficient. Moreover, since the channel coefficients are completely unknown, privacy is inherently guaranteed, as discussed in [18].

The remainder of this letter is organized as follows: we

* This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

[1] H.Y. Oksuz, F. Molinari, and J. Raisch are with the Control Systems Group at Technische Universität Berlin, Germany. {oksuz@tu-berlin.de}, {molinari,raisch}@control.tu-berlin.de.

[2] H.Y. Oksuz, H. Sprekeler, and J. Raisch are with Exzellenzcluster Science of Intelligence, Technische Universität Berlin, Marchstr. 23, 10587, Berlin, Germany.

[3] H. Sprekeler is with the Modelling Cognitive Processes Group at Technische Universität Berlin, Germany.

[1] In TDMA (Time Division Multiple Access), agents are assigned different time slots when they can transmit, whereas in FDMA (Frequency Division Multiple Access), different frequency bands are allocated to different users.

present the problem setup in Section II. In Section III, we introduce the proposed algorithm, whose convergence analysis is presented in Section IV. A numerical example is presented in Section V. Concluding remarks are given in Section VI.

II. PROBLEM SETUP

The set of real numbers is denoted by $\mathbb{R}$, and $\mathbb{R}^m$ represents m-dimensional Euclidean space. $\mathbb{N}$ and $\mathbb{N}_0$ respectively denote the set of natural numbers and the set of nonnegative integers. Given a finite set T, its cardinality is denoted by $|T|$. For a vector $x \in \mathbb{R}^m$, x^T denotes its transpose. Euclidean norm of the vector $x \in \mathbb{R}^m$ is denoted by $||x||$. The projection of $x \in \mathbb{R}^m$ onto a nonempty closed convex set $S \subset \mathbb{R}^m$ is denoted by $\mathbf{P}_S(x)$, i.e., $\mathbf{P}_S(x) = \arg\min_{s\in S} ||s-x||$. Projection is nonexpansive, i.e., $||\mathbf{P}_\Theta(x) - \mathbf{P}_\Theta(y)|| \leq ||x-y||$ holds for all $x, y \in \mathbb{R}^m$ if Θ is a nonempty closed convex set (see [19]). For given $a, b \in \mathbb{R}$, the function $\max\{a,b\}$ takes value a if $a > b$, and b otherwise. The expected value of a random variable p is denoted by $\mathbb{E}[p]$. Given a logical argument $g_i(x)$, $\mathbf{1}_{\{g_i(x)\}}$ denotes the function that takes value 1 when $g_i(x)$ is true and 0 when $g_i(x)$ is false. For a function $f : \mathbb{R}^m \to (-\infty, \infty]$, we define $D_f = \{x \in \mathbb{R}^m \big| f(x) < \infty\}$. Then, for a subgradient of a convex function f with respect to x at $\tilde{x} \in D_f$, denoted by $\partial_x f(\tilde{x})$, the inequality $f(\tilde{\mathbf{x}}) + \partial f^T(\tilde{\mathbf{x}})(\mathbf{x} - \tilde{\mathbf{x}}) \leq f(\mathbf{x})$ holds for all $\mathbf{x} \in D_f$.

A. Minmax Reformulation

In a federated learning setting with N agents, where $V = \{1, 2, \cdots, N\}$ denotes the index set, we are interested in improving the performance of the worst-performing agent by solving the following optimization problem:

$$\min_{\theta\in\Theta} \max_{i\in V} g_i(\theta). \tag{3}$$

We aim to compute a parameter vector estimate minimizing the worst-case loss observed among all agents, thus providing some form of fairness [20], [21]. However, it is difficult and inefficient to use (3) since it cannot be naturally split into independent subproblems over agents, requiring coordination for joint decision-making. This coordination overhead leads to scalability and efficiency challenges. Instead, we can consider an alternative (epigraph) form:

$$\begin{aligned} \min_{\alpha\in\mathbb{R}} \quad & \alpha \\ \text{subject to} \quad & g_i(\theta) \leq \alpha \qquad \forall \theta \in \Theta,\, i \in V; \end{aligned} \tag{4}$$

where the optimal value α^* is assumed to be finite. Moreover, $g_i(\theta^*) \leq \alpha^*$ holds if $\theta^* \in \Theta$ is optimal for (4). Let $\bar{g}_i\big((\theta,\alpha)\big) = \max\{g_i(\theta) - \alpha, 0\}$ and $p_i > 0$ respectively denote a penalty for violating the constraint in (4) for $i \in V$ and a weight of this penalty. Then, by following [22] and [23], one can rewrite (4) as

$$\min_{\theta\in\Theta, \alpha\in\mathbb{R}} \quad \alpha + \sum_{i=1}^{N} p_i \bar{g}_i\big((\theta,\alpha)\big). \tag{5}$$

It has been shown in [22] that any solution for (5) is also a solution of (4) if $p_i > 1$ for all $i \in V$. Hence, (5) can be considered as the global loss function for the federated learning setting.

B. Over-the-Air Communication Model

In wireless communication systems, the wireless multiple access channel (WMAC) model has been extensively used to characterize communication between multiple transmitters and a single receiver over fading channels, e.g., [24], [25]. Throughout this letter, we employ the WMAC-based communication model described by [15], [26]. Let each agent $i \in V$ simultaneously transmit information $\mathbf{s}_i(k) \in \mathbb{R}^m$ to the central unit in the same frequency band at each time step $k \in \mathbb{N}_0$. However, this information is corrupted by the channel and superimposed by the receiver, i.e., the received information by the central unit is given by

$$\mathbf{s}^{rec}(k) = \sum_{i=1}^{N} \lambda_i(k) \mathbf{s}_i(k), \tag{6}$$

where the $\lambda_i(k)$ are unknown time-varying positive channel coefficients, i.e., $\lambda_i(k) > 0$ for all $i = 1, 2, \cdots, N$.

Note that employing Over-the-Air computation has two main advantages: (i) the channel coefficients $\lambda_i(k)$ in (6) are unknown, and it is impossible to reconstruct $s_i(k)$ from $\mathbf{s}^{rec}(k)$, which inherently provides *privacy*; (ii) our approach does not require any knowledge of the channel coefficients, nor do they need to be reconstructed. This makes our algorithm highly efficient, in particular for large N. This is demonstrated in numerical experiments in Section V.

III. FEDERATED FAIR OVER-THE-AIR LEARNING (FEDFAIR) ALGORITHM

The FedFAir algorithm is summarized in Algorithm 1. At the beginning, the central unit selects $\theta(0) \in \Theta$ and $\alpha(0) \in \mathbb{R}$. Then, through iterations, the central unit computes (7) and broadcasts $v(k)$ and $\theta(k)$. Subsequently, each agent computes its own $\alpha_i(k)$ and $\theta_i(k)$ by using the local update rules (8) and (9) with the step size $\eta(k)$, respectively. Afterward, in the first round of agent-to-central unit communications, all agents transmit simultaneously (and in the same frequency band) their local values $\alpha_i(k)$ to the central unit. In the second round, each agent transmits simultaneously (and in the same frequency band) their local parameter vectors $\theta_i(k)$. Finally, in the third round, the constant $\rho_i(k) = 1$ is transmitted by all agents again simultaneously and in the same frequency band. We assume that the delays between the three rounds are sufficiently small such that the channel coefficients can be considered constant over the three rounds. According to the WMAC model, the central unit receives the vector (11) and two scalars given in (12) and (13) at each time step $k \in \mathbb{N}_0$. Finally, the central unit computes (14) and (15), which can be rewritten in the following form:

$$\theta(k+1) = \mathbf{P}_\Theta\left(\sum_{i=1}^{N} h_i(k)\theta_i(k)\right), \tag{16}$$

$$\alpha(k+1) = \sum_{i=1}^{N} h_i(k)\alpha_i(k), \tag{17}$$

Algorithm 1: FedFAir

Initialization: $\theta(0) \in \Theta$, $\alpha(0) \in \mathbb{R}$, and $p_i \in \mathbb{R}$

Loop:

for each time step $k \in \mathbb{N}_0$ **do**

The central unit computes:

$$v(k) = \alpha(k) - \frac{\eta(k)}{N} \tag{7}$$

The central unit broadcasts $\theta(k)$ and $v(k)$

Each agent i updates its local variables:

$$\theta_i(k) = \theta(k) - \eta(k) p_i \partial_\theta \bar{g}_i\big((\theta(k), v(k))\big) \tag{8}$$

$$\alpha_i(k) = v(k) - \eta(k) p_i \partial_\alpha \bar{g}_i\big((\theta(k), v(k))\big) \tag{9}$$

$$\rho_i(k) = 1 \tag{10}$$

Each agent i transmits $\theta_i(k)$, $\alpha_i(k)$, and $\rho_i(k)$

The central unit receives:

$$\theta^{\text{rec}}(k) = \sum_{i=1}^{N} \lambda_i(k) \theta_i(k) \tag{11}$$

$$\alpha^{\text{rec}}(k) = \sum_{i=1}^{N} \lambda_i(k) \alpha_i(k) \tag{12}$$

$$\rho^{\text{rec}}(k) = \sum_{i=1}^{N} \lambda_i(k) \rho_i(k) \tag{13}$$

The central unit updates:

$$\theta(k+1) = \mathbf{P}_\Theta\left(\frac{\theta^{\text{rec}}(k)}{\rho^{\text{rec}}(k)}\right) \tag{14}$$

$$\alpha(k+1) = \frac{\alpha^{\text{rec}}(k)}{\rho^{\text{rec}}(k)} \tag{15}$$

end for

where $\mathbf{P}_\Theta$ is the projection operator onto the set Θ, $h_i(k) = \frac{\lambda_i(k)}{\sum_{i=1}^{N} \lambda_i(k)}$ are the normalized channel coefficients, and the $\lambda_i(k)$ are the unknown time-varying positive real channel coefficients. By construction, the $h_i(k)$ are also positive, and, for all $k \geq 0$,

$$\sum_{i=1}^{N} h_i(k) = 1. \tag{18}$$

Next, we state our assumptions on individual objective functions and the step size as follows:

Assumption 1: The constraint set $\Theta \subset \mathbb{R}^m$ is convex and compact. As a consequence (see [27, Theorem 2.41, p.40]), Θ is then also closed and bounded.

Assumption 2: The individual objective functions $g_i(\theta)$ are convex over $\mathbb{R}^m$. Consequently, the minimax problem (3) is also convex since the point-wise maximum of convex functions preserves convexity [19, Proposition 1.2.4]. In addition, the local cost functions $\bar{g}_i\big((\theta, \alpha)\big)$ are also convex but nondifferentiable, and their sets of subdifferentials are nonempty for all $x \in \mathbb{R}$, and $i \in V$ [19, Proposition 4.2.1].

Assumption 3: The step size $\eta(k)$ in the FedFAir algorithm is chosen to satisfy $\eta(k) > 0$, $\sum_{k=0}^{\infty} \eta(k) = \infty$, and $\sum_{k=0}^{\infty} \eta^2(k) < \infty$.

Assumption 4: The unknown time-varying positive real channel coefficients $\lambda_i(k)$ $(i = 1, 2, \cdots, N)$ are assumed to be random variables, independent across time and agents.

Remark 1: Assumption 4 is standard in the modeling of WMACs (see [28, Ch 2.3, Ch 2.4], [29, Ch 5.4], and [30]). We do not assume any specific probability distribution for the channel coefficients.

IV. CONVERGENCE PROPERTIES

We start with the following lemmas that are essential for the proofs presented in the letter.

Lemma 1 ([31, Thm. 3.4.2]): If a sequence $\{a(k)\}$ of real numbers converges to a real number x, then any subsequence $\{a(k_t)\}$ of $\{a(k)\}$ also converges to x.

Lemma 2 ([32, Lem. 11, p.50]): Let $\{v(k)\}$, $\{b(k)\}$, $\{u(k)\}$, and $\{c(k)\}$ be nonnegative sequences of random variables. Suppose that

(i) $\sum_{k=0}^{\infty} b(k) < \infty$ and $\sum_{k=0}^{\infty} c(k) < \infty$ hold almost surely,

(ii) For each $k \in \mathbb{N}_0$, the following holds almost surely

$$\mathbb{E}\big[v(k+1)|F_k\big] \leq \big(1 + b(k)\big)v(k) - u(k) + c(k), \tag{19}$$

where $\mathbb{E}\big[v(k+1)|F_k\big]$ denotes the conditional expectation for the given $F_k = \{v(t), u(t), c(t), t = 0, 1, \cdots, k\}$.

Then, $\{v(k)\}$ converges to some $v \geq 0$ and $\sum_{k=0}^{\infty} u(k) < \infty$ almost surely.

Lemma 3 ([33, Prop. 2]): Let $\{a(k)\}$ be a nonnegative sequence and $\{b(k)\}$ an eventually nonnegative sequence, i.e., $\exists \tilde{k} \geq 0$ such that $b(k) \geq 0$ $\forall k \geq \tilde{k}$. Let $\sum_{k=0}^{\infty} a(k) = \infty$ and $\sum_{k=0}^{\infty} a(k) b(k) < \infty$. Then, there exists a subsequence $\{b(k_t)\}$ of $\{b(k)\}$ such that $\lim_{t\to\infty} b(k_t) = 0$.

Lemma 4: [23] Let $\alpha^* = \min_{\theta\in\Theta} \max_{i\in V} g_i(\theta)$, $\theta^* = \arg\min_{\theta\in\Theta} \max_{i\in V} g_i(\theta)$, and $p_i > 1$ for all $i \in V$. Then, for $\bar{g}_i\big((\theta, \alpha)\big) = \max\{g_i(\theta) - \alpha, 0\}$, we have $\sum_{i=1}^{N} p_i \bar{g}_i\big((\theta, \alpha)\big) + \alpha \geq \alpha^*$ for all $\theta \in \Theta$ and $\alpha \in \mathbb{R}$, where equality holds if and only if $\theta = \theta^*$ and $\alpha = \alpha^*$.

We are now ready to present the main result.

Theorem 1: Suppose that Assumptions 1, 2, 3, and 4 hold. Let $\theta^* \in \Theta$ and $\alpha^* \in \mathbb{R}$ respectively be an optimal solution and the optimal value for the problem (3). If $p_i > \max\left(1, \frac{1}{N\mathbb{E}[h_i(k)]}\right)$ for all $i \in V$, then $\lim_{k\to\infty} \theta(k) = \theta^*$ and $\lim_{k\to\infty} \alpha(k) = \alpha^*$ with probability 1.

Proof: We start by introducing $\beta(k) = \big[\theta(k)^T \;\; \alpha(k)\big]^T$, $y(k) = \big[\theta(k)^T \;\; v(k)\big]^T$. Then, for any $\beta^* = \big[\theta^{*T} \;\; \alpha^*\big]^T \in \Theta \times \mathbb{R}$, by using (7)-(17), the non-expansive property of the projection $\mathbf{P}_\Theta$, and the fact that $\theta^* = \mathbf{P}_\Theta(\theta^*)$, we have

$$\begin{aligned}
||\beta(k+1) - \beta^*||^2 &= \left|\left| \begin{bmatrix} \mathbf{P}_\Theta\Big(\sum_{i=1}^{N} h_i(k)\theta_i(k)\Big) - \mathbf{P}_\Theta(\theta^*) \\ \sum_{i=1}^{N} h_i(k)\alpha_i(k) - \alpha^* \end{bmatrix} \right|\right|^2 \\
&\leq \left|\left| \begin{bmatrix} \theta(k) - \theta^* - \eta(k)\sum_{i=1}^{N} h_i(k) p_i \partial_\theta \bar{g}_i\big((\theta(k), v(k))\big) \\ v(k) - \alpha^* - \eta(k)\sum_{i=1}^{N} h_i(k) p_i \partial_\alpha \bar{g}_i\big((\theta(k), v(k))\big) \end{bmatrix} \right|\right|^2 \\
&= \left|\left| y(k) - \beta^* - \eta(k) \sum_{i=1}^{N} h_i(k) p_i \partial \bar{g}_i\big(y(k)\big) \right|\right|^2,
\end{aligned} \tag{20}$$

where $\partial \bar{g}_i\big(y(k)\big) = \big[\partial_\theta g_i\big(\theta(k)\big) \quad -1\big]^T \mathbf{1}_{\{g_i(\theta) \geq \alpha\}}$.

We can further expand (20) as

$$||\beta(k+1)-\beta^*||^2 \leq \Big|\Big|y(k)-\eta(k)\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i\big(y(k)\big)-\beta^*\Big|\Big|^2 = \big||y(k)-\beta^*\big||^2 - 2\eta(k)\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i^T\big(y(k)\big)\big(y(k)-\beta^*\big) + \Big|\Big|\eta(k)\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i\big(y(k)\big)\Big|\Big|^2. \tag{21}$$

For the first term on the right-hand side of (21), we have

$$\big||y(k)-\beta^*\big||^2 = \left|\left|\begin{bmatrix}\theta(k)-\theta^*\\ \alpha(k)-\alpha^*\end{bmatrix}-\begin{bmatrix}0\\ \frac{\eta(k)}{N}\end{bmatrix}\right|\right|^2 = \big||\beta(k)-\beta^*\big||^2-\frac{2\eta(k)}{N}\big(\alpha(k)-\alpha^*\big)+\frac{\eta^2(k)}{N^2}. \tag{22}$$

which follows from (7). Note that the boundedness of the constraint set Θ implies $\exists L_\Theta>0$ such that $\big||\partial\bar{g}_i\big(y(k)\big)\big||\leq L_\Theta$, which together with (18), the convexity of the function $||\cdot||^2$, and the fact that $1<p_i\leq p_{max}$ with $p_{max}=\max_i p_i$ ($i\in V$) can be used to find an upper-bound for the last term on the right-hand side of (21) as

$$\Big|\Big|\eta(k)\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i\big(y(k)\big)\Big|\Big|^2 = \eta^2(k)\Big|\Big|\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i\big(y(k)\big)\Big|\Big|^2 \leq \eta^2(k)M_1, \tag{23}$$

where $M_1=p_{max}^2L_\Theta^2$. Let F_k represent the past iterates of $\alpha(k)$ and $\theta(k)$, i.e., $F_k=\{\alpha(t),\,\theta(t),t=0,1,\cdots,k\}$ for $k\in\mathbb{N}_0$. Subsequently, by using (22), (23), and taking the expectation conditioned on F_k of both sides of (21) together with the linearity of the expectation, we obtain

$$\mathbb{E}\big[||\beta(k+1)-\beta^*||^2|F_k\big] \leq \big||\beta(k)-\beta^*\big||^2-\frac{2\eta(k)}{N}(\alpha(k)-\alpha^*) - 2\eta(k)\mathbb{E}\big[\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i^T\big(y(k)\big)\big(y(k)-\beta^*\big)|F_k\big] + \eta^2(k)M_2, \tag{24}$$

where $M_2=M_1+\frac{1}{N^2}$. Note that due to Assumption 4, the statistics of $h_i(k)$ ($i=1,2,\cdots,N$) are independent of $h_i(t)$ for $t<k$, which also implies that $y(k)$ and $h_i(k)$ are statistically independent at time k since the statistics of $y(k)$ are dependent only on $h_i(t)$ for $t<k$ and $i=1,2,\cdots,N$. Hence, $\mathbb{E}\big[h_i(k)|F_k\big]=\mathbb{E}\big[h_i(k)\big]$ holds for all $k\in\mathbb{N}_0$. Then, by using the linearity of the expectation again, we can write the third term on the right-hand side of (24) as

$$\begin{aligned}&-2\eta(k)\mathbb{E}\big[\sum_{i=1}^{N}h_i(k)p_i\partial\bar{g}_i^T\big(y(k)\big)\big(y(k)-\beta^*\big)|F_k\big]\\ &=-2\eta(k)\sum_{i=1}^{N}\mathbb{E}\big[h_i(k)p_i\partial\bar{g}_i^T\big(y(k)\big)\big(y(k)-\beta^*\big)|F_k\big]\\ &=-2\eta(k)\sum_{i=1}^{N}\mathbb{E}\big[h_i(k)\big]p_i\partial\bar{g}_i^T\big(y(k)\big)\big(y(k)-\beta^*\big).\end{aligned} \tag{25}$$

Moreover, by Assumption 2 (convexity of $\bar{g}_i(\cdot)$), we have

$$\sum_{i=1}^{N}p_i\partial_y\bar{g}_i^T\big(y(k)\big)\big(y(k)-\beta^*\big) \geq \sum_{i=1}^{N}p_i\Big(\bar{g}_i\big(y(k)\big)-\bar{g}_i\big(\beta^*\big)\Big) = \sum_{i=1}^{N}p_i\bar{g}_i\big(y(k)\big), \tag{26}$$

which follows from the fact that for any $\beta^*=\begin{bmatrix}\theta^{*T} & \alpha^*\end{bmatrix}^T\in\Theta\times\mathbb{R}$, we have $g_i(\theta^*)-\alpha^*\leq 0$, and therefore $\bar{g}_i(v^*)=\max\{g_i(\theta^*)-\alpha^*,0\}=0$ for all $i\in V$ and $k\in\mathbb{N}_0$. Thus, by using (26), (25) can be written as

$$-2\eta(k)\sum_{i=1}^{N}\mathbb{E}\big[h_i(k)\big]p_i\partial\bar{g}_i\big(y(k)\big)\big(y(k)-\beta^*\big) \leq -2\eta(k)\sum_{i=1}^{N}\mathbb{E}\big[h_i(k)\big]p_i\bar{g}_i\big(y(k)\big), \tag{27}$$

which can then be used to rewrite (24) as

$$\mathbb{E}\big[||\beta(k+1)-\beta^*||^2|F_k\big] \leq \big||\beta(k)-\beta^*\big||^2+\eta^2(k)M_2 - \frac{2\eta(k)}{N}\Big(\alpha(k)-\alpha^*+N\sum_{i=1}^{N}\bar{p}_i\bar{g}_i\big(y(k)\big)\Big) \tag{28}$$

where $\bar{p}_i=\mathbb{E}\big[h_i(k)\big]p_i$. We can add and subtract the term $2\eta(k)\big(\sum_{i=1}^{N}\bar{p}_i\bar{g}_i\big(\beta(k)\big)\big)$ from the right hand side of (28) to get

$$\begin{aligned}\mathbb{E}\big[||\beta(k+1)-\beta^*||^2|F_k\big] &\leq \big||\beta(k)-\beta^*\big||^2+\eta^2(k)M_2\\ &\quad-\frac{2\eta(k)}{N}\Big(\alpha(k)-\alpha^*+N\sum_{i=1}^{N}\bar{p}_i\bar{g}_i\big(\beta(k)\big)\Big)\\ &\quad+2\eta(k)\bar{p}_{max}\sum_{i=1}^{N}\big|\bar{g}_i\big(\beta(k)\big)-\bar{g}_i\big(y(k)\big)\big|,\end{aligned} \tag{29}$$

which follows from the fact that

$$\sum_{i=1}^{N}\bar{p}_i\bar{g}_i\big(\beta(k)\big)-\sum_{i=1}^{N}\bar{p}_i\bar{g}_i\big(y(k)\big) \leq \bar{p}_{max}\sum_{i=1}^{N}\big|\bar{g}_i\big(\beta(k)\big)-\bar{g}_i\big(y(k)\big)\big|$$

where $\bar{p}_{max}=\max_i\bar{p}_i$ ($i\in V$). By using (7) and the relation $|\max\{x_1,0\}-\max\{x_2,0\}|\leq|x_1-x_2|$ for scalars x_1 and x_2, one can obtain

$$|\bar{g}_i\big(\beta(k)\big)-\bar{g}_i\big(y(k)\big)| \leq |g_i\big(\theta(k)\big)-\alpha(k)-g_i\big(\theta(k)\big)+v(k)| = \frac{\eta(k)}{N}, \tag{30}$$

which can then be used to obtain

$$\mathbb{E}\big[||\beta(k+1)-\beta^*||^2|F_k\big] \leq \big||\beta(k)-\beta^*\big||^2+\eta^2(k)M_3 - \frac{2\eta(k)}{N}\Big(\alpha(k)-\alpha^*+\sum_{i=1}^{N}w_i\bar{g}_i\big(\beta(k)\big)\Big) \tag{31}$$

where $w_i=N\bar{p}_i$ and $M_3=M_2+2p_{max}$. Suppose that $p_i>\max\Big(1,\frac{1}{N\mathbb{E}[h_i(k)]}\Big)$. Then, $w_i=N\bar{p}_i=N\mathbb{E}\big[h_i(k)\big]p_i>1$. Note that since $\alpha^*=\min_{\theta\in\Theta}\max_{i\in V}g_i(\theta)$, by Lemma 4, we have $\alpha(k)-\alpha^*+\sum_{i=1}^{N}w_i\bar{g}_i\big(\beta(k)\big)\geq 0$ for all $k\in\mathbb{N}_0$. Moreover,

by Assumption 3, we have $\sum_{k=0}^{\infty}\eta^2(k)<\infty$, which together with (31) and Lemma 2 give

$$\lim_{k\to\infty}\left|\left|\beta(k)-\beta^*\right|\right|^2=\vartheta\geq 0, \tag{32}$$

$$\sum_{k=0}^{\infty}\eta(k)\Big(\alpha(k)-\alpha^*+\sum_{i=1}^{N}w_i\bar{g}_i\big(\beta(k)\big)\Big)<\infty, \tag{33}$$

for any $\beta^*=\left[\theta^{*T}\ \ \alpha^*\right]^T\in\Theta\times\mathbb{R}$ and $\alpha^*\in\mathbb{R}$ with probability 1. Additionally, $\sum_{k=0}^{\infty}\eta(k)=\infty$ also holds by Assumption 3, which together with Lemma 3 imply that there exists a subsequence such that the following holds with probability 1,

$$\lim_{l\to\infty}\Big(\alpha(k_l)-\alpha^*+\sum_{i=1}^{N}w_i\bar{g}_i\big(\beta(k_l)\big)\Big)=0, \tag{34}$$

which also implies that along a further subsequence, we have limit points $\lim_{l\to\infty}\alpha(k_l)=\bar{\alpha}$ and $\lim_{l\to\infty}\beta(k_l)=\bar{\beta}$ with probability 1, which satisfy

$$\bar{\alpha}-\alpha^*+\sum_{i=1}^{N}w_i\bar{g}_i\big(\bar{\beta}\big)=0. \tag{35}$$

Hence, by Lemma 4 and the fact that $w_i>1$ for all $i\in V$, it follows that $\bar{\alpha}=\alpha^*$ and $\bar{\theta}=\theta^*$ are optimal for the minmax problem (3) with probability 1. This implies that $\lim_{l\to\infty}\left|\left|\beta(k_l)-\beta^*\right|\right|^2=0$ with probability 1. Hence, by Lemma 1, we have $\vartheta=0$ in (32) with probability 1, and that concludes the proof. ■

V. NUMERICAL EXAMPLE

We consider a federated logistic regression problem, where each agent has its own private data. The objective is to collaboratively build a global model that can carry out a binary classification task. In this setting, each agent has 2 different classes of data, labeled by 0 or 1. We represent the dataset of the i-th agent by $D_i=\{d_i^n\}_{n=1}^{|D_i|}$, where $d_i^n=(u_i^n,z_i^n)\in\mathbb{R}^m\times\{0,1\}$. The objective is to find a separation rule so that agents can identify the correct classes of some unseen data from different classes. To this end, the following convex loss function is used by each agent:

$$g_i(\theta,d_i)=-\frac{1}{|D_i|}\Bigg(\sum_{n=1}^{|D_i|}z_i^n\log\left(S(\theta^Tu_i^n)\right)+(1-z_i^n)\log\left(1-S(\theta^Tu_i^n)\right)\Bigg), \tag{36}$$

where $S(x)=\frac{1}{1+e^{-x}}$ and $S(\theta^Tu_i^n)$ is the local estimate of the i-th agent. The step size is chosen as $\eta(k)=\frac{0.01}{(k+1)^{0.6}}$.

To demonstrate the fairness and robustness of the FedFAir algorithm, we consider a system of 12 agents, each with an individual imbalanced dataset, and the total number of training samples is different for each agent. Moreover, different types of noise, sampled from different distributions, are injected into the data available to each agent to ensure statistical heterogeneity. The parameter vector is of dimension 4, i.e., $\theta\in\mathbb{R}^4$, and $p_i>1$ is chosen for all agents $i=1,2,\cdots,12$. Additionally, the constraint set is given by $\Theta=\{\theta\in\mathbb{R}^4|\ ||\theta||\leq 10\}$. In order to monitor the performance of the system, we randomly sample some previously unseen test data at every iteration, and then we measure the performance of the global model on this test data by computing its classification accuracy during training. As it can be seen from Fig. 1, the FedFAir algorithm achieves around 90% accuracy after approximately 5000 transmissions.

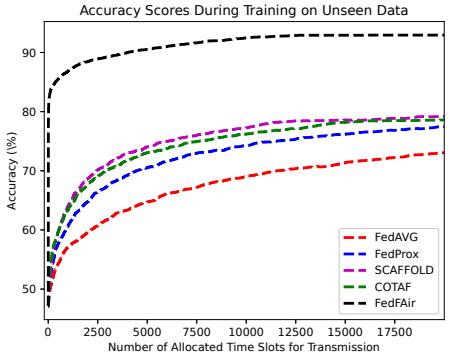
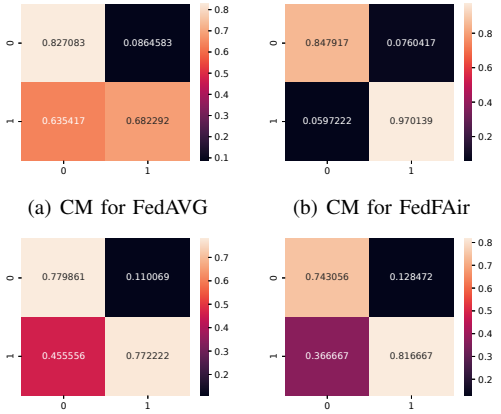


Fig. 1. Comparison of the performance of FedFAir, FedAVG, COTAF, FedProx, and SCAFFOLD algorithms on unseen data during training in terms of their accuracy.



(a) CM for FedAVG (b) CM for FedFAir

(c) CM for FedProx (d) CM for SCAFFOLD

Fig. 2. Comparison of Confusion Matrices (CMs)

We have also compared FedFAir with FedAVG [1] and three other state-of-the-art algorithms, COTAF [8], FedProx [12], and SCAFFOLD [34]. Note that single-step gradient update is carried out on agents' devices for FedFair while the other algorithms require agents to run multiple gradient steps on their devices, which increases the performance but also the computational costs. As can be seen in Fig. 1, COTAF, FedProx, and SCAFFOLD perform better than FedAVG, but the performance of FedFAir is observed to be the best. This is even more remarkable as COTAF, which also uses Over-the-Air computation, assumes all channel coefficients to be known. Simulations for COTAF were therefore performed for known channel coefficients, while they were unknown in the FedFAir simulations. Moreover, due to the heterogeneity in the distribution of the data, we observe from the confusion

matrices (CMs) in Fig. 2 that the FedAVG algorithm cannot properly identify the samples of data with labels 1. Even though we see some improvements for COTAF, FedProx, and SCAFFOLD, the FedFAir algorithm performs much better by recognizing a large percentage of data, both labeled 0 and 1.

Additionally, we compare the FedFAir algorithm and the other algorithms in terms of communication efficiency. For the latter, the TDMA scheme is used for communication between agents and the central unit. In this case, an individual time slot is allocated for each agent to transmit its parameter vector at each communication round. After receiving parameter vectors, the central unit computes the average of them and sends it back to the agents. Since we consider a system with $N = 12$ agents, it takes $N = 12$ time slots per communication round for each agent to transmit its updated parameter vector, while only 3 are needed for the FedFAir algorithm (see (13)), independent of the value of N. This makes the execution of one iteration of the FedFAir algorithm $N/3 = 4$ times faster than the execution of an iteration of the other algorithms.

VI. CONCLUSION

In this letter, we have introduced the FedFAir algorithm, which uses Over-the-Air Computation to carry out efficient decentralized learning while providing fairness and improved performance. We have shown that the FedFAir algorithm converges to an optimal solution of the minimax problem. Furthermore, FedFAir ensures fairness by minimizing the maximum loss across all agents, regardless of their individual characteristics, thus promoting robustness against data heterogeneity. We have also illustrated our theoretical findings with a numerical example.

Future research will include the development of resilient federated learning algorithms when there are malicious agents in the system.

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] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[3] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, B. McMahan *et al.*, "Towards federated learning at scale: System design," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[5] A. Nedic, "Distributed gradient methods for convex machine learning problems in networks: Distributed optimization," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 92–101, 2020.

[6] V. P. Chellapandi, A. Upadhyay, A. Hashemi, and S. H. Żak, "On the convergence of decentralized federated learning under imperfect information sharing," *IEEE Control Systems Letters*, 2023.

[7] T. Omori and K. Kashima, "Combinatorial optimization approach to client scheduling for federated learning," *IEEE Control Systems Letters*, 2023.

[8] 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.

[9] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 14–41, 2022.

[10] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–44, 2023.

[11] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[12] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[13] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[14] H. Y. Oksuz, F. Molinari, H. Sprekeler, and J. Raisch, "Federated learning in wireless networks via over-the-air computations," in *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023, pp. 4379–4386.

[15] F. Molinari, N. Agrawal, S. Stańczak, and J. Raisch, "Max-consensus over fading wireless channels," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 2, pp. 791–802, 2021.

[16] M. Frey, I. Bjelaković, and S. Stańczak, "Over-the-air computation in correlated channels," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5739–5755, 2021.

[17] 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.

[18] F. Molinari and J. Raisch, "Exploiting wireless interference for distributively solving linear equations," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 2999–3006, 2020.

[19] D. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex analysis and optimization*. Athena Scientific, 2003, vol. 1.

[20] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[21] Z. Hu, K. Shaloudegi, G. Zhang, and Y. Yu, "Federated learning meets multi-objective optimization," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 4, pp. 2039–2051, 2022.

[22] D. P. Bertsekas, "Necessary and sufficient conditions for a penalty method to be exact," *Mathematical programming*, vol. 9, no. 1, pp. 87–99, 1975.

[23] K. Srivastava, A. Nedić, and D. Stipanović, "Distributed min-max optimization in networks," in *2011 17th International Conference on Digital Signal Processing (DSP)*. IEEE, 2011, pp. 1–8.

[24] R. Ahlswede, "Multi-way communication channels," in *Second International Symposium on Information Theory: Tsahkadsor, Armenia, USSR, Sept. 2-8, 1971*, 1973.

[25] A. Giridhar and P. Kumar, "Toward a theory of in-network computation in wireless sensor networks," *IEEE Communications Magazine*, vol. 44, no. 4, pp. 98–107, apr 2006.

[26] F. Molinari, N. Agrawal, S. Stańczak, and J. Raisch, "Over-the-air max-consensus in clustered networks adopting half-duplex communication technology," *IEEE Transactions on Control of Network Systems*, 2022.

[27] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1976, vol. 3.

[28] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2012.

[29] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012.

[30] B. Sklar, "Rayleigh fading channels in mobile digital communication systems. i. characterization," *IEEE Communications magazine*, vol. 35, no. 7, pp. 90–100, 1997.

[31] R. G. Bartle and D. R. Sherbert, *Introduction to real analysis*. Wiley New York, 2000.

[32] B. T. Polyak, *Introduction to optimization. optimization software*. Inc., Publications Division, New York, 1987.

[33] Y. I. Alber, A. N. Iusem, and M. V. Solodov, "On the projected subgradient method for nonsmooth convex optimization in a hilbert space," *Mathematical Programming*, vol. 81, pp. 23–35, 1998.

[34] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.