Client Sampling for Communication-Efficient Distributed Minimax Optimization

Wen Xu*, Ben Liang*, Gary Boudreau†, and Hamza Sokun†
*Department of Electrical and Computer Engineering, University of Toronto, Canada
†Ericsson Canada Inc., Canada

***Abstract*—Distributed minimax optimization is essential for robust federated learning, offering resiliency against the variability in data distribution. Most previous works focus only on learning guarantees and convergence analysis, without explicit consideration of the communication delay, which can be crucial in practical systems. In this work, we consider the problem of communication-efficient distributed minimax optimization via judicious client sampling, proposing an algorithm termed CE-MINIMAX, which takes into consideration both the training convergence performance and the communication time per training round. We derive convergence bounds for CE-MINIMAX under both convex and non-convex loss functions, which we then use to design the client sampling probabilities in joint consideration of the communication time. We conduct numerical experiments with canonical classification datasets to demonstrate that CE-MINIMAX can achieve higher worst-case test accuracy under substantially reduced communication time, compared with state-of-the-art client sampling schemes for distributed minimax optimization.**

I. INTRODUCTION

With the advent of ubiquitous computing devices and vast amount of available data, distributed machine learning, especially federated learning (FL), has received increasing attention in research [1]–[4]. In FL, the goal is to train a model utilizing the locally stored data from all clients in a distributed manner, with a central server serving as the coordinator. Since each client maintains its local dataset, each local loss function $f_n(\cdot)$ can only be accessed at the corresponding client n.

In standard FL, the optimization objective is to minimize a global loss, defined as a fixed $\boldsymbol{p}$-weighted sum of the local losses, i.e.,

$$\min_{\boldsymbol{w}\in\mathcal{W}} f(\boldsymbol{w}) := \sum_{n=1}^{N} p_n f_n(\boldsymbol{w}), \tag{1}$$

where N is the number of clients and $\boldsymbol{w} \in \mathcal{W}$ contains the model parameters. The weight vector $\boldsymbol{p}$ here is a given constant vector, whose nth element is typically chosen to be $\frac{1}{N}$ or a number proportional to the local data size at client n. Numerous first-order methods such as federated stochastic gradient descent (FEDSGD) [4], federated averaging (FEDAVG) [4], local stochastic gradient descent (LOCAL SGD) [5], and many variants have been proposed to solve (1). However, the minimization objective in (1) only guarantees model performance on a specific $\boldsymbol{p}$-weighted mixture of the local data distributions. It has been observed that the performance of the learned model on some clients can be abysmal in some applications, which are highly undesirable [6], [7].

To address this issue, the so called agnostic federated learning (AFL) was first proposed in [8], which considered the following minimax optimization problem,

$$\min_{\boldsymbol{w}\in\mathcal{W}} \max_{\boldsymbol{p}\in\mathcal{P}} F(\boldsymbol{w},\boldsymbol{p}) := \sum_{n=1}^{N} p_n f_n(\boldsymbol{w}), \tag{2}$$

where $\boldsymbol{p} \in \mathcal{P}$ is a general weight vector to be optimized. The solution of (2) promotes uniform performance on the learned model over the data distributions of all clients, i.e., *distributionally robust* learning [7]. Various distributed minimax optimization methods have been studied in [8]–[17].

In FL, due to the large number of model or gradient parameters that need to be sent from the clients to the server, communication efficiency is of crucial importance [1]–[4]. In particular, in a large network with many clients, only a subset of the clients should be selected to participate, in order to reduce the communication overhead. This is commonly termed *client sampling*. However, none of [8]–[17] above has considered the communication delay. For example, in [8], the clients are selected uniformly at random or only based on the value of the weight vector $\boldsymbol{p}$. Clearly, these client selection schemes are suboptimal under realistic communication conditions or general allocation of resources that are heterogeneous over different clients. However, no prior work has studied the impact of client sampling on the communication overhead in distributed minimax optimization.

In this work, we are motivated to design an algorithm to distributedly train a robust model by solving the minimax optimization problem in (2), with a principled client sampling strategy to also reduce the communication delay. Toward this goal, we need to overcome the following challenges: i) The algorithm for solving the minimax optimization problem should enable flexible client sampling strategies and have convergence guarantees. ii) The client sampling strategy should take into account not only the learning performance but also the communication overhead. iii) The optimization of client sampling should be solved efficiently. In this context, the contributions of this paper are as follows:

- We propose a distributed Communication-Efficient Minimax (CE-MINIMAX) algorithm to solve problem (2) under the FL framework with reduced communication delay. CE-MINIMAX allows random client sampling with any probability distribution, while guaranteeing the con-

This work was supported in part by Ericsson, the Natural Sciences and Engineering Research Council of Canada, and Mitacs.

vergence of FL training. To the best of our knowledge, no existing work in the literature provides explicit consideration of flexible client sampling for distributed minimax optimization.
- We derive convergence bounds for both convex and non-convex loss functions. In the convex case, we bound the expected duality gap, while in the non-convex case, we bound the stationarity of a Moreau envelope of the worst-case loss function. The convergence bounds are then used to optimize the client sampling probabilities in CE-MINIMAX, in joint consideration of the communication time per training round.
- We conduct numerical studies for classification tasks on canonical image datasets over heterogeneous communication environments. Our simulation results show that CE-MINIMAX increases FL robustness for the worst-case test data distribution, and it can substantially reduce the communication time, compared with state-of-the-art client sampling methods for distributed minimax optimization.

The remainder of the paper is organized as follows. We provide a literature review on distributed minimax optimization and client sampling in Section II. The distributed minimax optimization framework is discussed in Section III, followed by the design and analysis details of CE-MINIMAX in Section IV. The simulation results are presented in Section V. We give conclusion remarks in Section VI.

II. RELATED WORK

A. Centralized and Distributed Minimax Optimization

Minimax problems, also sometimes under the name of saddle-point problems, have been extensively studied in the *centralize* setting. They include the seminal work of von Neumann's minimax theorem [18] to solve bilinear minimax problems, as well as more general forms in game theory [19] and optimization [20]–[23]. Numerous first-order methods have been proposed to solve minimax problems, including gradient descent ascent (GDA) and stochastic gradient descent ascent (SGDA) [21], [24], [25] with acceleration [26], [27], extra-gradient (EG) [28]–[31], and optimistic gradient descent ascent (OGDA) [30]–[32].

The authors of [8] were the first to propose *distributed* solutions to minimax optimization. The resultant STOCHASTIC-AFL algorithm uses SGDA and provides learning performance guarantee for convex loss functions. Subsequently, leveraging gradient tracking, FEDGDA-GT was proposed to achieve linear convergence for minimax optimization of Lipschitz smooth and strongly-convex-strongly-concave functions in [9]. Allowing multi-step local updates, the distributionally robust federated averaging (DRFA) [10] and the local stochastic gradient descent ascent (LOCAL-SGDA) [11] were proposed to achieve communication-convergence trade-off for distributed minimax problems. Utilizing different momentum techniques, SGDAM-PEF and SGDAM-REF were proposed and analyzed in [12] while momentum local SGDA combined momentum and local updates in [13]. Applying variance reduction techniques, DGDA-VR was proposed to solve distributed nonconvex strongly-concave minimax problems [14]. It was proposed to normalize the client updates to deal with client heterogeneity in general minimax problems [15]. Furthermore, lower bounds for distributed and decentralized minimax problems were analyzed in [16], [17].

None of these methods explicitly consider per-round communication time with system constraints. Our work fills this gap by introducing communication-aware client sampling.

B. Client Sampling in Standard Federated Learning

It is well-known that client selection can reduce the communication overhead in distributed learning [33]–[42]. An early work to perform client selection in the FL setting was [33], where the FEDCS algorithm was proposed to choose as many clients as possible with resource constraints on computation and communication. A meta algorithm FLANP was proposed in [34] to make FL straggler-resilient under system heterogeneity by geometrically increasing client participation. For wireless FL, a joint learning, wireless resource allocation, and client selection problem was proposed and solved in [35], and the FEDL algorithm was proposed to optimize CPU-cycle control, uplink power control, accuracy level and learning rate in [36]. All of [33]–[36] considered deterministic client selection.

It is possible to allow probabilistic client sampling where each client is chosen with some probability to avoid combinatorial optimization of client selection and to achieve more flexibility in performance improvement. Using the norms of the gradients as the *importance* of the data at different clients, client sampling for FL was proposed and analyzed in [37] without any explicit system consideration. Importance and channel-aware probabilistic client sampling was proposed in [38], sampling one client per round. Optimization to minimize convergence time was solved in [39], assuming that one client has to always be connected to the central server. Further improvements to optimize the client sampling probabilities for communication efficiency have been proposed in [40]–[42].

However, all of the aforementioned methods only consider client sampling for the standard FL problem of minimizing the weighted sum loss as defined in (1). They are not applicable to the minimax optimization problem, which is the focus of our work.

III. SYSTEM MODEL AND PROBLEM STATEMENT

A. Federated Learning with Minimax Optimization

We consider the conventional client-server system architecture in FL, where a central server coordinates the training of a machine learning model with N clients indexed by $\mathcal{N} = [N] := \{1, \dots, N\}$ over a star network topology. All clients can directly communicate with the central server but no pair of clients can communicate with each other.

For any $n \in [N]$, the local dataset $\mathcal{D}_n$ at client n is drawn independently from data distribution $\mathcal{P}_n$ and has size D_n. We refer to the j-th data point at client n as $z_{n,j}$ for all $n \in [N]$ and $j \in [D_n]$. We denote the loss incurred on one data

point $z_{n,j}$ under model $\boldsymbol{w}$ by $\ell_{n,j}(\boldsymbol{w}) = \ell(\boldsymbol{w}, z_{n,j})$ for all $n \in [N]$ and $j \in [D_n]$. Then, the average loss incurred on a mini-batch of data points J under model $\boldsymbol{w}$ is $\ell_{n,J}(\boldsymbol{w}) = \frac{1}{|J|}\sum_{j\in J}\ell_{n,j}(\boldsymbol{w})$ for all $n \in [N]$ and $J \subseteq \mathcal{D}_n$, and the local empirical loss for the entire dataset at client n under model $\boldsymbol{w}$ is $f_n(\boldsymbol{w}) = \frac{1}{D_n}\sum_{j\in[D_n]}\ell_{n,j}(\boldsymbol{w})$ for all $n \in [N]$.

The objective of our FL system is distributionally robust learning [7], [8], i.e., to minimize the loss over the worst mixture of the N clients' local data distributions, formulated as a minimax optimization problem in (2). Note that $\boldsymbol{p} \in \mathcal{P}$ in (2). If $\boldsymbol{p}$ is unconstrained, then problem (2) is equivalent to minimizing the local loss of a client with the worst-case local data distribution. However, for general $\mathcal{P}$, we are interested in the worst mixture of data distributions.

B. Client Sampling and Communication Overhead

We assume T discrete time slots, i.e., FL training rounds, to solve problem (2), whose index set is $\mathcal{T} = \{0, 1, \ldots, T-1\}$. In each round t, each client $n \in [N]$ is selected independently with probability $q_n^{(t)}$ to participate in the training. We denote the client sampling vector in round t by $\boldsymbol{q}^{(t)} = [q_1^{(t)}, \ldots, q_n^{(t)}, \ldots, q_N^{(t)}]$. We further assume the expected number of scheduled clients per round is set to m, i.e., $\sum_{n=1}^{N} q_n^{(t)} = m$ for all $t \in \mathcal{T}$. This can be interpreted as a limitation due to the per-round resource budget given to the system.

Besides the accuracy of solution to problem (2), we are also interested in the communication overhead over the T training rounds. In particular, we note that in practical network systems, the clients can have diverse communication capabilities and channel conditions. For example, a client with weak channel will require increased duration to transmit its parameter updates to the server. Let $T_n^{(t)}$ be the uplink communication time of any client n in round t, which can be computed using channel state estimates from classical communication techniques such as pilot signaling. Then, the expected total communication time for round t is given by

$$\Gamma(\boldsymbol{q}^{(t)}) = \mathbb{E}_{\{a_n^{(t)}\}_{n=1}^N}\left[\sum_{n\in[N]} a_n^{(t)} T_n^{(t)}\right] = \sum_{n\in[N]} q_n^{(t)} T_n^{(t)}, \quad (3)$$

where $a_n^{(t)}$ is the indicator function for whether client n is scheduled in round t. A typical example of this is time division multiple access (TDMA), where the total communication time is determined by the sum of all client-server transmission times.

For downlink communication, since the server can broadcast model $\boldsymbol{w}$ to all clients with constant time requirement, we do not include it in our communication overhead optimization. Therefore, our goal in this work is to design a distributed optimization scheme to solve the minimax problem (2) while minimizing the total uplink communication time, i.e., the sum of (3) over all T training rounds.

IV. Client Sampling for Communication-Efficient Distributed Minimax Optimization

We propose a new algorithm called Communication-Efficient Minimax (CE-Minimax) with flexible client sampling to improve communication efficiency. We first provide an outline on the two phases of the proposed algorithm, we then derive a convergence bound for its minimax optimization performance, which is followed by optimization design of the client sampling probabilities based on the convergence bound and communication overhead.

A. Outline of CE-Minimax

CE-Minimax adopts the general SGDA framework [8], [21], [24], [25], but we add special treatment to accommodate the client sampling probability vector $\boldsymbol{q}^{(t)}$. The server first initializes a global model $\boldsymbol{w}^{(0)}$ and a uniform weight vector $\boldsymbol{p}^{(0)} = [1/N, \ldots, 1/N]$. In each training round t, there are two phases. In Phase 1 the model parameter vector $\boldsymbol{w}$ is updated, while in Phase 2 the weight vector $\boldsymbol{p}$ is updated. The algorithm terminates after T training rounds. The pseudocode of CE-Minimax is given in Algorithm 1.

In Phase 1, the server samples clients to obtain a subset $\mathcal{S}^{(t)} \subseteq [N]$ of all clients, with sampling probabilities given by the vector $\boldsymbol{q}^{(t)}$ from the solution to optimization problem **P1**, which is later detailed in Section IV-C. Specifically, the server samples each client n via an independent Bernoulli trial with success probability given by $q_n^{(t)}$. The server then broadcasts the current model $\boldsymbol{w}^{(t)}$ to all clients $n \in \mathcal{S}^{(t)}$. After receiving the model $\boldsymbol{w}^{(t)}$, each client $n \in \mathcal{S}^{(t)}$ samples a mini-batch of data points $J_n^{(t)} \subseteq \mathcal{D}_n$ uniformly at random and calculates the stochastic gradient of the local loss $\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})$ in parallel. Each client $n \in \mathcal{S}^{(t)}$ then sends $\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})$ to the server.

The server aggregates all available $\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})$ to construct the stochastic gradient $\tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})$ via

$$\tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) = \sum_{n\in\mathcal{N}} \frac{p_n^{(t)} a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)}). \quad (4)$$

Note that in the aggregation of local gradients in (4), we compensate each local gradient by dividing its sampling probability such that the aggregated stochastic gradient is always an unbiased estimation of the full gradient, i.e., $\mathbb{E}_{\boldsymbol{q}^{(t)}, J_n^{(t)}\sim\mathcal{D}_n}[\sum_{n\in\mathcal{N}} \frac{p_n^{(t)} a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})] = \sum_{n\in\mathcal{N}} p_n^{(t)} \nabla_{\boldsymbol{w}} f_n(\boldsymbol{w}^{(t)}) = \nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})$. Then, the server updates the model via

$$\boldsymbol{w}^{(t+1)} = \Pi_{\mathcal{W}}(\boldsymbol{w}^{(t)} - \eta_w \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})), \quad (5)$$

where η_w is the learning rate of model updates and $\Pi_{\mathcal{W}}(\cdot)$ is the projection onto the set $\mathcal{W}$.

In Phase 2, the server first samples $\mathcal{U}^{(t)}$ clients of size m uniformly at random and broadcasts the model $\boldsymbol{w}^{(t)}$ to all sampled clients. Each client $n \in \mathcal{U}^{(t)}$ samples a mini-batch $K_n^{(t)}$ of its data points uniformly at random and calculates the sampled loss $\ell_{n,K_n^{(t)}}(\boldsymbol{w}^{(t)})$ in parallel. After all sampled

Algorithm 1: CE-MINIMAX

Input: initial model $\boldsymbol{w}^{(0)}$, initial weight vector $\boldsymbol{p}^{(0)} = [1/N, \ldots, 1/N]$, learning rates η_w and η_p, total number of rounds T, hyperparameter λ, and m.

Output: $\hat{\boldsymbol{w}} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{|\mathcal{S}^{(t)}|} \sum_{n\in\mathcal{S}^{(t)}} \boldsymbol{w}_n^{(t)}$ and $\hat{\boldsymbol{p}} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{p}^{(t)}$.

for each round $t = 0, \ldots, T-1$ **do**
  // Phase 1
  Server calculates $\boldsymbol{q}^{(t)}$ via solving **P1**.
  Server samples clients $\mathcal{S}^{(t)} \subseteq \mathcal{N}$ by $\boldsymbol{q}^{(t)}$.
  Server broadcasts $\boldsymbol{w}^{(t)}$ to $\mathcal{S}^{(t)}$.
  for each client $n \in \mathcal{S}^{(t)}$ **do**
    Client n samples a mini-batch of data points $J_n^{(t)} \subseteq \mathcal{D}_n$ uniformly at random.
    Client n calculates $\nabla_{\boldsymbol{w}} \ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})$.
    Client n sends $\nabla_{\boldsymbol{w}} \ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})$ to the server.
  end for
  Server constructs $\tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})$ via (4).
  Server updates the model parameters via (5).

  // Phase 2
  Server samples clients $\mathcal{U}^{(t)} \subseteq \mathcal{N}$ uniformly at random.
  Server broadcasts $\boldsymbol{w}^{(t)}$ to $\mathcal{U}^{(t)}$.
  for each client $n \in \mathcal{U}^{(t)}$ **do**
    Client n samples a mini-batch of data points $K_n^{(t)} \subseteq \mathcal{D}_n$ uniformly at random.
    Client n calculates $\ell_{n,K_n^{(t)}}(\boldsymbol{w}^{(t)})$.
  end for
  Server constructs $\tilde{\nabla}_{\boldsymbol{p}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})$ via (6).
  Server updates the weight vector via (7).
end for

clients send the estimated loss $\ell_{n,K_n^{(t)}}(\boldsymbol{w}^{(t)})$ to the server, the server constructs the stochastic gradient $\tilde{\nabla}_{\boldsymbol{p}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})$, where its nth element is defined as

$$[\tilde{\nabla}_{\boldsymbol{p}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})]_n = \begin{cases} \frac{N}{m} \ell_{n,K_n^{(t)}}(\boldsymbol{w}^{(t)}), & \text{if } n \in \mathcal{U}^{(t)}, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Finally, the server updates the weight vector via projected gradient ascent

$$\boldsymbol{p}^{(t+1)} = \Pi_{\mathcal{P}}(\boldsymbol{p}^{(t)} + \eta_p \tilde{\nabla}_{\boldsymbol{p}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})), \quad (7)$$

where η_p is the learning rate of weight vector updates and $\Pi_{\mathcal{P}}(\cdot)$ is the projection onto the set $\mathcal{P}$.

Note that for simplicity we choose uniform client sampling for the weight $\boldsymbol{p}$ update in Phase 2. We can allow a more flexible strategy there but that does not change our convergence analysis, and it has limited impact on our communication overhead since each client only transmits a float number in Phase 2 instead of the entire gradient in Phase 1.

B. Convergence Analysis with Client Sampling

In order to optimize $\{\boldsymbol{q}^{(t)}\}_{t=0}^{T-1}$, we first need to analyze its impact on the convergence performance of the CE-MINIMAX algorithm. Since we are solving the minimax optimization problem (2) instead of the conventional minimization problem (1), the existing convergence analysis of first-order methods for distributed minimization such as [5] does not directly apply.

A convergence bound was derived in [8] for minimax optimization, account for how the optimization variables $\boldsymbol{w}$ and $\boldsymbol{p}$ are updated iteratively. However, [8] considered only the limited scenario where a single client is chosen in each training round, so its analysis does not apply to CE-MINIMAX, where multiple clients are sampled in each round based on a general vector $\boldsymbol{q}^{(t)}$. Furthermore, the analysis in [8] is limited to convex loss functions. Here, we consider both convex and non-convex loss functions, which enables more general applicability to the training of machine learning models, such as neural networks, that induce non-convex loss.

We adopt the following commonly used assumptions [8], [10], [43].

Assumption 1 (Bounded Domains)**.** *The diameters of the compact convex set $\mathcal{W}$ and $\mathcal{P}$ are $R_{\mathcal{W}}$ and $R_{\mathcal{P}}$, respectively.*

Assumption 2 (Bounded Gradients)**.** *There exists some positive G_w and G_p such that $\|\nabla_{\boldsymbol{w}} f_n(\boldsymbol{w})\|_2 \leq G_w, \forall \boldsymbol{w} \in \mathcal{W}, \forall n \in \mathcal{N}$ and $\|\nabla_{\boldsymbol{p}} F(\boldsymbol{w}, \boldsymbol{p})\|_2 \leq G_p, \forall \boldsymbol{w} \in \mathcal{W}, \forall \boldsymbol{p} \in \mathcal{P}$.*

Assumption 3 (Bounded Variance of Stochastic Gradients)**.** *There exist some non-negative constants $\sigma_{w,I}$ and σ_p such that $\mathbb{E}[\|\nabla_{\boldsymbol{w}} \ell_{n,J_n^{(t)}}(\boldsymbol{w}) - \nabla_{\boldsymbol{w}} f_n(\boldsymbol{w})\|_2^2] \leq \sigma_{w,I}^2$ holds for all $n \in \mathcal{N}$, $t \in \mathcal{T}$, $\boldsymbol{w} \in \mathcal{W}$ and $J_n^{(t)} \sim \mathcal{D}_n$ and $\mathbb{E}[\|\tilde{\nabla}_{\boldsymbol{p}} F(\boldsymbol{w}, \boldsymbol{p}) - \nabla_{\boldsymbol{p}} F(\boldsymbol{w}, \boldsymbol{p})\|_2^2] \leq \sigma_p^2$ holds for all $\boldsymbol{w} \in \mathcal{W}$, $\boldsymbol{p} \in \mathcal{P}$.*

1) Convex loss: We first analyze the convergence performance when the local loss functions are convex in $\boldsymbol{w}$.

Assumption 4 (Convexity)**.** *For all $n \in \mathcal{N}$, the local loss $f_n(\cdot)$ is convex, i.e., $f_n(\boldsymbol{w}_1) \geq f_n(\boldsymbol{w}_2) + \nabla f_n(\boldsymbol{w}_2)^T(\boldsymbol{w}_1 - \boldsymbol{w}_2)$ holds $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$.*

Clearly, in this case, for any given $\boldsymbol{p} \in \mathcal{P}$, $F(\boldsymbol{w}, \boldsymbol{p})$ is convex in $\boldsymbol{w}$ as it is a non-negative weighted sum of convex functions. Furthermore, for any given $\boldsymbol{w} \in \mathcal{W}$, $F(\boldsymbol{w}, \boldsymbol{p})$ is linear in $\boldsymbol{p}$. Hence, (2) is a constrained convex-concave minimax optimization problem.

The optimality of a solution $(\hat{\boldsymbol{w}}, \hat{\boldsymbol{p}})$ for constrained convex-concave optimization can be measured by the duality gap, given by

$$\text{gap}(\hat{\boldsymbol{w}}, \hat{\boldsymbol{p}}) := \max_{\boldsymbol{p}\in\mathcal{P}} F(\hat{\boldsymbol{w}}, \boldsymbol{p}) - \min_{\boldsymbol{w}\in\mathcal{W}} F(\boldsymbol{w}, \hat{\boldsymbol{p}}), \quad (8)$$

where $\hat{\boldsymbol{p}} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{p}^{(t)}$ is the time-averaged weights, $\hat{\boldsymbol{w}} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{|\mathcal{S}^{(t)}|} \sum_{n\in\mathcal{S}^{(t)}} \boldsymbol{w}_n^{(t)}$ contains the time-averaged model parameters, and $\mathcal{S}^{(t)}$ is the set of sampled clients for model updates in round t. A minimax point $(\boldsymbol{w}^*, \boldsymbol{p}^*)$, i.e., Nash equilibrium (NE), always exists in the constrained convex-concave minimax optimization problem. By definition, the duality gap of any minimax point $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ is zero, i.e., $\text{gap}(\boldsymbol{w}^*, \boldsymbol{p}^*) = \max_{\boldsymbol{p}\in\mathcal{P}} F(\boldsymbol{w}^*, \boldsymbol{p}) - \min_{\boldsymbol{w}\in\mathcal{W}} F(\boldsymbol{w}, \boldsymbol{p}^*) = 0$.

Therefore, the smaller gap$(\hat{\boldsymbol{w}}, \hat{\boldsymbol{p}})$ is, the better the solution $(\hat{\boldsymbol{w}}, \hat{\boldsymbol{p}})$ is.

Theorem 1 (Convex-Concave)**.** *Under Assumption 1-4, the expected duality gap of* CE-MINIMAX *with arbitrary client sampling probabilities* $\{\boldsymbol{q}^{(t)}\}_{t=0}^{T-1}$ *has the following upper bound:*

$$\mathbb{E}\left[gap(\hat{\boldsymbol{w}}, \hat{\boldsymbol{p}})\right] \leq \frac{R_{\mathcal{P}}^2}{2\eta_p T} + \frac{\eta_p(G_p^2+\sigma_p^2)}{2} + \frac{R_{\mathcal{P}}\sigma_p}{\sqrt{T}} + \frac{R_{\mathcal{W}}^2}{2\eta_w T} + \frac{\eta_w}{2T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2+\sigma_{w,I}^2) + \frac{R_{\mathcal{W}}}{T}\sqrt{\sum_{t=0}^{T-1}\sum_{n=1}^{N} p_n^{(t)}\Psi(q_n^{(t)})}, \quad (9)$$

where $\Psi(q_n^{(t)}) = (\frac{(1-q_n^{(t)})G_w^2}{q_n^{(t)}} + \frac{\sigma_{w,I}^2}{q_n^{(t)}})$.

Proof. Please see Appendix A. □

Remark 1. *When* $q_n^{(t)} = 1$ *holds for all* $n \in \mathcal{N}$ *and* $t \in \mathcal{T}$*, i.e., full participation of clients, and we let* $\eta_w = \frac{R_w}{\sqrt{T(G_w^2+\sigma_{w,I}^2)}}$ *and* $\eta_p = \frac{R_p}{\sqrt{T(G_p^2+\sigma_p^2)}}$ *in* (9)*, we obtain the same convergence rate of* $\mathcal{O}(\frac{1}{\sqrt{T}})$ *achieved in [8].*

2) Non-convex loss: With non-convex loss functions, an NE may not exist, and the duality gap is no longer a meaningful measure of optimality. To facilitate our analysis, we define $\Phi(\boldsymbol{w}) = \max_{\boldsymbol{p}\in\mathcal{P}} F(\boldsymbol{w}, \boldsymbol{p})$ such that our optimization formulation (2) is equivalent to $\min_{\boldsymbol{w}\in\mathcal{W}} \Phi(\boldsymbol{w})$. Note that $\Phi(\boldsymbol{w})$ can be computed efficiently because $F(\boldsymbol{w}, \cdot)$ is linear in $\boldsymbol{p}$ given any $\boldsymbol{w} \in \mathcal{W}$. However, since $\Phi(\cdot)$ itself is non-convex in $\boldsymbol{w}$, the problem of finding a global minimum of $\Phi(\cdot)$ is in general NP-hard. Instead, we will follow a common approach in the non-convex optimization literature to find a stationary point of Φ. Note that the function $\Phi(\cdot)$ generally is not differentiable, making it impossible to directly use the gradient to find such a stationary point.

Instead, we leverage a Moreau envelope on $\Phi(\cdot)$ for convergence analysis [25], [43]. The μ-Moreau envelope of a function Φ with a positive parameter μ is $\Phi_\mu(\boldsymbol{w}) = \min_{\boldsymbol{v}\in\mathcal{W}}\{\Phi(\boldsymbol{v}) + \frac{1}{2\mu}\|\boldsymbol{v}-\boldsymbol{w}\|^2\}$. To properly choose the hyperparameter μ and facilitate convergence analysis, we further assume our minimax objective F is L-smooth, i.e., has Lipschitz gradients,

Assumption 5 (Smoothness)**.** *There exists a positive* L *such that* $\|\nabla F(\boldsymbol{w}_1, \boldsymbol{p}_1) - \nabla F(\boldsymbol{w}_2, \boldsymbol{p}_2)\| \leq L\|(\boldsymbol{w}_1, \boldsymbol{p}_1) - (\boldsymbol{w}_2, \boldsymbol{p}_2)\|$*, holds* $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$ *and* $\forall \boldsymbol{p}_1, \boldsymbol{p}_2 \in \mathcal{P}$*.*

With this smoothness assumption, Φ can still be non-smooth but it is a L-weakly convex function [25]. We then choose $\mu = \frac{1}{2L}$ and observe that $\Phi_{1/2L}(\cdot)$ is differentiable. Hence, the stationarity of the function Φ can be approximately measured by $\|\nabla\Phi_{1/2L}(\cdot)\|$.

Theorem 2. *Under Assumption 1-3 and Assumption 5, the* $(1/2L)$*-Moreau envelope of* Φ *in* CE-MINIMAX *with arbitrary client sampling probabilities* $\{\boldsymbol{q}^{(t)}\}_{t=0}^{T-1}$ *satisfies the following upper bound:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\Phi_{1/2L}(\boldsymbol{w}^{(t)})\right\|^2\right] &\leq \frac{4\Phi_{1/2L}(\boldsymbol{w}^{(0)})}{\eta_w T} + 8L\left(\frac{R_{\mathcal{P}}^2}{2\eta_p T} + \frac{\eta_p}{2}(G_p^2+\sigma_p^2)\right) \\ &\quad + 4\frac{R_{\mathcal{W}}L}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N} p_n^{(t)}\sqrt{\Psi(q_n^{(t)})} \\ &\quad + \frac{4L\eta_w}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2+\sigma_{w,I}^2), \quad (10)\end{aligned}$$

where $\Psi(q_n^{(t)}) = \left(\frac{(1-q_n^{(t)})G_w^2}{q_n^{(t)}} + \frac{\sigma_{w,I}^2}{q_n^{(t)}}\right)$.

Proof. Please see Appendix B. □

Remark 2. *When* $q_n^{(t)} = 1$ *holds for all* $n \in \mathcal{N}$ *and* $t \in \mathcal{T}$*, i.e., full participation of clients, and we let* $\eta_w = \sqrt{\frac{\Phi_{1/2L}(\boldsymbol{w}^{(0)})}{LT(G_w^2+\sigma_{w,I}^2)}}$ *and* $\eta_p = \frac{R_p}{\sqrt{T(G_p^2+\sigma_p^2)}}$ *in* (9)*, the convergence rate is simplified to* $\mathcal{O}(\frac{1}{\sqrt{T}})$.

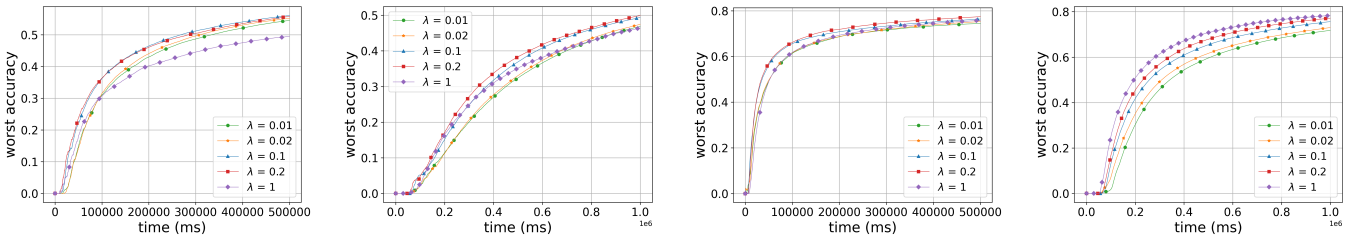
C. Optimization of Client Sampling Probabilities

In this section, we utilize the convergence bounds in Theorem 1 and Theorem 2, and further take into account the communication delay to optimize the client sampling probabilities $\{\boldsymbol{q}^{(t)}\}_{t=0}^{T-1}$. This then fully specifies the CE-MINIMAX algorithm outlined in Section IV-A.

From the bounds in (9) and (10), we can see that only the last two expressions on the right-hand side are related to the sampling probabilities $\{\boldsymbol{q}^{(t)}\}_{t=0}^{T-1}$. However, neither of the terms $\sqrt{\sum_{t=0}^{T-1}\sum_{n=1}^{N} p_n^{(t)}\Psi(q_n^{(t)})}$ and $\sum_{n=1}^{N} p_n^{(t)}\sqrt{\Psi(q_n^{(t)})}$ is a convex function of $\boldsymbol{q}^{(t)}$, making minimization of the upper bounds intractable. Instead, we make a monotonicity-preserving convex relaxation by removing the square root operation. Then, by rearranging terms, we have the following two expressions for convex loss and non-convex loss, respectively: $\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}(\frac{\eta_w}{2}+R_{\mathcal{W}})(\sigma_{w,I}^2+G_w^2) - \frac{\mathcal{W}}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N} p_n^{(t)}G_w^2$ and $\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}(4L(\eta_w + R_{\mathcal{W}})(G_w^2+\sigma_{w,I}^2)) - \frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N} p_n^{(t)}4R_{\mathcal{W}}LG_w^2$. We observe that in both cases only the first term contains the sampling probability $\boldsymbol{q}$, which has the same pattern of $\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}\zeta$, where ζ is some constant independent of $\{\boldsymbol{q}^{(t)}\}_{t=0}^{T-1}$.

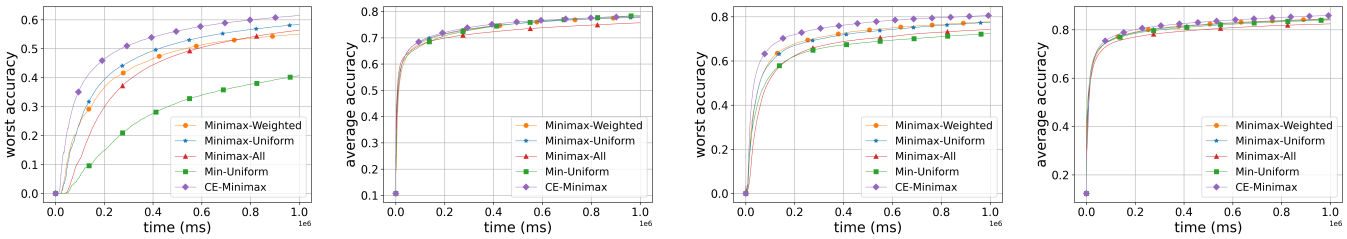
Taking into account the communication time $\Gamma(\cdot)$ as defined in Section III-B, we consider the following minimization objective in each round t:

$$y_0(t) = \sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}} + \lambda\Gamma(\boldsymbol{q}^{(t)}), \quad (11)$$



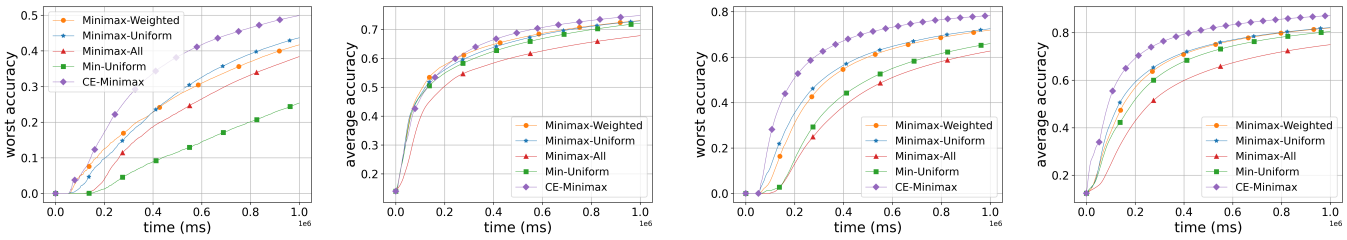
(a) LR for Fashion (b) MLP for Fashion (c) LR for EMNIST (d) MLP for EMNIST

Fig. 1: Worst-case test accuracy over time for various λ values.



(a) Worst accuracy, Fashion (b) Average accuracy, Fashion (c) Worst accuracy, EMNIST (d) Average accuracy, EMNIST

Fig. 2: Worst-case and average test accuracies of LR.



(a) Worst accuracy, Fashion (b) Average accuracy, Fashion (c) Worst accuracy, EMNIST (d) Average accuracy, EMNIST

Fig. 3: Worst-case and average test accuracies of MLP.

where $\lambda \in \mathbb{R}_+$ is a hyper-parameter that represents the relative importance of the communication time. Thus, we obtain the following per-round optimization problem:

$$\textbf{P1:} \underset{\boldsymbol{q}^{(t)}}{\text{minimize}} \quad y_0(t) \tag{12}$$

$$\text{subject to} \quad \sum_{n=1}^{N} q_n^{(t)} = m, \tag{13}$$

$$0 < q_n^{(t)} \leq 1, \quad \forall n \in [N]. \tag{14}$$

Since $\Gamma(\cdot)$ is linear in $\boldsymbol{q}^{(t)}$, the optimization problem **P1** is convex and can be efficiently solved via standard convex solvers [44]. Our proposed algorithm CE-MINIMAX uses the optimal solutions of these per-round problems **P1** as shown in line 3 of Algorithm 1.

V. NUMERICAL EVALUATION

We consider the following benchmarks for comparison with the proposed CE-MINIMAX:

- MINIMAX-UNIFORM. Calculating the stochastic gradient with respect to $\boldsymbol{w}$ in each round t by choosing m client uniformly at random. This is a multi-client extension of the PERDOMAIN scheme in [8].
- MINIMAX-WEIGHTED. Calculating the stochastic gradient with respect to $\boldsymbol{w}$ in each round t with client sampling probabilities proportional to the current weight vector $\boldsymbol{p}^{(t)}$ and m expected clients. This is a multi-client extension of the WEIGHTED scheme [8].
- MINIMAX-ALL. Always selecting all clients, i.e., $q_n^{(t)} = 1$, $\forall n \in [N], \forall t \in \mathcal{T}$.
- MIN-UNIFORM. Solving the minimization problem (1) using federated SGD from [4] with m clients chosen uniformly at random in each round.

Note that we do not numerically compare CE-MINIMAX with the other distributed minimax solutions in [9]–[17]. None of these solutions consider flexible client sampling, and they improve on [8] with extensions or additional techniques (e.g., gradient tracking [9]) that are orthogonal to the client sampling of CE-MINIMAX.

We conduct numerical experiments, using the PyTorch version 2.3.1 [45], on the Fashion-MNIST dataset [46] and the EMNIST/Digits dataset [47], both of which contain images of 10 classes. We consider $N = 10$ clients, where each client contains all data points of exactly one image class. We use cross-entropy loss in all experiments. For learning models,

we use multinomial logistic regression (LR) to represent the scenario of convex loss functions, while we use a fully-connected neural network, i.e., multi-layer perceptron (MLP), with 2-hidden layers of neurons 300 and 100 each and ReLU activation functions to represent the scenario of non-convex loss functions. For heterogeneous communication time, we assume 10 ms for clients 1-5 and 1 ms for clients 6-10. We set $m = 5$ where applicable. To enhance algorithmic stability, we add a chi-squared divergence regularization of strength 0.00001 between the weight vector $\boldsymbol{p}$ and a uniform vector for all methods solving the minimax optimization.

A. Effect of λ on CE-Minimax Performance

In our per-round optimization **P1**, λ is a tunable hyper-parameter. In Fig. 1, we show the worst-case test accuracy among the N clients/distributions under different λ values for both LR and MLP on both the Fashion-MNIST and EMNIST datasets. We observe that a λ value between 0.1 and 1 gives the best performance in most scenarios. In the results below for CE-Minimax, we choose the best λ value for each scenario.

B. Convex Loss

We study the results of LR on the Fashion-MNIST dataset and the EMNIST dataset. We set $\lambda = 0.1$ for Fashion-MNIST and $\lambda = 0.2$ for EMNIST. In Fig. 2, we show both the worst-case test accuracy among the N clients/distributions, and the average test accuracy. We observe that CE-Minimax achieves significantly better tradeoff between worst-case test accuracy and communication time than the other algorithms. For example, to reach 55% worst-case test accuracy for Fashion-MNIST, CE-MINIMAX requires 443.102 seconds, compared with 666.402 seconds for MINIMAX-UNIFORM, 995.895 seconds for MINIMAX-WEIGHTED, 872.795 seconds for MINIMAX-ALL while MIN-UNIFORM does not reach the required worst-case test accuracy within 1000 seconds. To reach 70% worst-case test accuracy for EMNIST, CE-MINIMAX requires 149.238 seconds, compared with 308.615 seconds for MINIMAX-UNIFORM, 275.691 seconds for MINIMAX-WEIGHTED, 492.030 seconds for MINIMAX-ALL and 677.229 seconds for MIN-UNIFORM. Furthermore, CE-MINIMAX also achieves average test accuracy competitive with the best among the other algorithms, so its distribution robustness does not need to come from any loss in the average performance.

C. Non-convex Loss

We study the results of MLP on the Fashion-MNIST dataset and the EMNIST dataset. We set $\lambda = 0.2$ for Fashion-MNIST and $\lambda = 1$ for EMNIST. We observe from Fig. 3 that CE-MINIMAX again substantially outperforms the other algorithms in terms of the tradeoff between the worst-case test accuracy and communication time. Furthermore, unlike the case of convex loss, we observe that CE-Minimax in the case of non-convex loss also achieves the best average test accuracy. This may be due to the intricate generalization abilities of neural networks.

VI. CONCLUSION

In this work, we propose the CE-MINIMAX algorithm to solve the distributed minimax optimization problem in a communication-efficient manner. CE-MINIMAX formulates and solves per-round optimization problems that take into account both the client data contribution and its communication delay to perform random client sampling. We derive convergence bounds for both convex loss and non-convex loss functions, which enables optimization of the client sampling probabilities. Our experiments on classification tasks demonstrate that our proposed CE-MINIMAX can substantially outperform other state-of-the-arts sampling benchmarks for the distributed minimax problems, achieving higher worst-case test accuracy for less communication time.

APPENDIX A
PROOF OF THEOREM 1

Proof. We first bound the terms related to updates of $\boldsymbol{w}$. Recall that the stochastic gradient with respect to $\boldsymbol{w}$ in round t is $\tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) = \sum_{n\in\mathcal{N}} \frac{p_n^{(t)} a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}} \ell_{n, J_n^{(t)}}(\boldsymbol{w}^{(t)})$. We have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}\|^2] \\
&= \mathbb{E}[\|\Pi_{\mathcal{W}}(\boldsymbol{w}^{(t)} - \eta_w \tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})) - \boldsymbol{w}\|^2] \\
&\overset{(a)}{\leq} \mathbb{E}[\|\boldsymbol{w}^{(t)} - \eta_w \tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) - \boldsymbol{w}\|^2] \\
&\overset{(b)}{=} \mathbb{E}[\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|^2] + \eta_w^2 A_1^{(t)} - 2\eta_w A_2^{(t)},
\end{aligned} \tag{15}$$

where (a) is by projection onto a convex set and (b) is by introducing $A_1^{(t)} = \mathbb{E}[\|\sum_{n\in\mathcal{N}} \frac{p_n^{(t)} a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}} \ell_{n, J_n^{(t)}}(\boldsymbol{w})\|^2]$ and $A_2^{(t)} = \mathbb{E}[(\boldsymbol{w}^{(t)} - \boldsymbol{w})^T (\sum_{n\in\mathcal{N}} \frac{p_n^{(t)} a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}} \ell_{n, J_n^{(t)}}(\boldsymbol{w}))]$.

To bound $A_1^{(t)}$, we have

$$\begin{aligned}
A_1^{(t)} &\overset{(a)}{\leq} \sum_{n=1}^N p_n^{(t)} \mathbb{E}[\|\frac{a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}} \ell_{n, J_n^{(t)}}(\boldsymbol{w})\|^2] \\
&= \sum_{n=1}^N p_n^{(t)} \mathbb{E}[|\frac{a_n^{(t)}}{q_n^{(t)}}|^2 \|\nabla_{\boldsymbol{w}} \ell_{n, J_n^{(t)}}(\boldsymbol{w})\|^2] \\
&\overset{(b)}{=} \sum_{n=1}^N p_n^{(t)} \mathbb{E}[|\frac{a_n^{(t)}}{q_n^{(t)}}|^2] \mathbb{E}[\|\nabla_{\boldsymbol{w}} \ell_{n, J_n^{(t)}}(\boldsymbol{w})\|^2] \\
&\overset{(c)}{\leq} \sum_{n=1}^N \frac{p_n^{(t)}}{q_n^{(t)}} (G_w^2 + \sigma_{w,I}^2),
\end{aligned} \tag{16}$$

where (a) is by Jensen's inequality, (b) is by the independence of client sampling and mini-batch data sampling, and (c) follows Assumptions 2 and 3. Substituting the bound of $A_1^{(t)}$ in (16) into (15) and rearranging terms, we have a bound for $A_2^{(t)}$:

$$\begin{aligned}
A_2^{(t)} \leq \frac{1}{2\eta_w} &(\mathbb{E}[\|\boldsymbol{w}^{(t)} - \boldsymbol{w}\|^2] - \mathbb{E}[\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}\|^2]) \\
&+ \frac{\eta_w}{2} \sum_{n=1}^N \frac{p_n^{(t)}}{q_n^{(t)}} (G_w^2 + \sigma_{w,I}^2).
\end{aligned} \tag{17}$$

Summing (17) over t from 0 to $T-1$, dividing both sides by T and taking total expectation, we have

$$\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1} A_2^{(t)}] \leq \frac{R_{\mathcal{W}}^2}{2\eta_w T} + \frac{\eta_w}{2T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2+\sigma_{w,I}^2). \quad (18)$$

Applying similar techniques to the updates of $\boldsymbol{p}$, we obtain

$$\mathbb{E}[\sum_{t=0}^{T-1}(\boldsymbol{p}-\boldsymbol{p}^{(t)})^T(\tilde{\nabla}_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})] \leq \frac{R_{\mathcal{P}}^2}{2\eta_p} + \frac{\eta_p T}{2}(G_p^2+\sigma_p^2). \quad (19)$$

We now bound the expected duality gap based on the bounds of updates on $\boldsymbol{w}$ and $\boldsymbol{p}$.

$$\begin{aligned}
&\mathbb{E}[\max_{\boldsymbol{p}\in\mathcal{P}} F(\hat{\boldsymbol{w}},\boldsymbol{p}) - \min_{\boldsymbol{w}\in\mathcal{W}} F(\boldsymbol{w},\hat{\boldsymbol{p}})] \\
&= \mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{p}\in\mathcal{P}}(F(\hat{\boldsymbol{w}},\boldsymbol{p}) - F(\boldsymbol{w},\hat{\boldsymbol{p}}))] \\
&\overset{(a)}{\leq} \mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{p}\in\mathcal{P}}\frac{1}{T}\sum_{t=0}^{T-1}(F(\boldsymbol{w}^{(t)},\boldsymbol{p}) - F(\boldsymbol{w},\boldsymbol{p}^{(t)}))] \\
&= \mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{p}\in\mathcal{P}}\frac{1}{T}\sum_{t=0}^{T-1}(F(\boldsymbol{w}^{(t)},\boldsymbol{p}) - F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) \\
&\quad + F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - F(\boldsymbol{w},\boldsymbol{p}^{(t)}))] \\
&\overset{(b)}{\leq} \mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{p}\in\mathcal{P}}(\frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}-\boldsymbol{p}^{(t)})^T\nabla_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) \\
&\quad + \frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{w}^{(t)}-\boldsymbol{w})^T\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})] \\
&\overset{(c)}{\leq} \mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W},\boldsymbol{p}\in\mathcal{P}} \mathrm{C1}+\mathrm{C2}+\mathrm{C3}+\mathrm{C4}+\mathrm{C5}+\mathrm{C6}], \quad (20)
\end{aligned}$$

where (a) and (b) are from the fact that $F(\cdot,\cdot)$ is convex-concave, and (c) is by introducing $\mathrm{C1} = \frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}-\boldsymbol{p}^{(t)})^T\tilde{\nabla}_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})$, $\mathrm{C2} = \frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{w}^{(t)}-\boldsymbol{w})^T\tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})$, $\mathrm{C3} = \frac{1}{T}\sum_{t=0}^{T-1}((-\boldsymbol{w})^T(\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})))$, $\mathrm{C4} = \frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}^T(\nabla_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})))$, $\mathrm{C5} = \frac{1}{T}\sum_{t=0}^{T-1}((-\boldsymbol{p}^{(t)})^T(\nabla_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})))$, $\mathrm{C6} = \frac{1}{T}\sum_{t=0}^{T-1}((\boldsymbol{w}^{(t)})^T(\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})))$.

Next, we further derive bounds for the expectations of C1 to C6. From (19), we have

$$\mathbb{E}[\max_{\boldsymbol{p}\in\mathcal{P}}\mathrm{C1}] \leq \frac{R_{\mathcal{P}}^2}{2\eta_p T} + \frac{\eta_p}{2}(G_p^2+\sigma_p^2). \quad (21)$$

From (18), we have

$$\mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W}}\mathrm{C2}] \leq \frac{R_{\mathcal{W}}^2}{2\eta_w T} + \frac{\eta_w}{2T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2+\sigma_{w,I}^2). \quad (22)$$

We now bound C3.

$$\begin{aligned}
&\mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W}}\mathrm{C3}] \\
&\leq \mathbb{E}[-\frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{w}^T(\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})))] \\
&\leq \mathbb{E}[\frac{1}{T}\|\boldsymbol{w}\|\|\sum_{t=0}^{T-1}(\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}))\|] \\
&\leq \mathbb{E}[\frac{1}{T}\|\boldsymbol{w}\|\sum_{t=0}^{T-1}\|\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})\|] \\
&\leq \frac{R_{\mathcal{W}}}{T}\sqrt{\mathbb{E}[(\sum_{t=0}^{T-1}\|\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})\|)^2]} \\
&\leq \frac{R_{\mathcal{W}}}{T}\sqrt{\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})\|^2]}.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)}) - \tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})\|^2] \\
&= \mathbb{E}[\|\sum_{n\in\mathcal{N}}\frac{p_n^{(t)}a_n^{(t)}}{q_n^{(t)}}\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)}) - \nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})\|^2] \\
&\overset{(a)}{\leq} \sum_{n\in\mathcal{N}} p_n^{(t)}\mathbb{E}[\|\frac{a_n^{(t)}}{q_n^{(t)}}\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)}) - \nabla_{\boldsymbol{w}}f_n(\boldsymbol{w}^{(t)})\|^2] \\
&= \sum_{n\in\mathcal{N}} p_n^{(t)}\big(\mathbb{E}[\|\frac{a_n^{(t)}}{q_n^{(t)}}\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})\|^2] + \mathbb{E}[\|\nabla_{\boldsymbol{w}}f_n(\boldsymbol{w}^{(t)})\|^2] \\
&\quad - 2\mathbb{E}[(\frac{a_n^{(t)}}{q_n^{(t)}}\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)}))^T\nabla_{\boldsymbol{w}}f_n(\boldsymbol{w}^{(t)})]\big) \\
&\overset{(b)}{=} \sum_{n\in\mathcal{N}} p_n^{(t)}(\mathbb{E}[\|\nabla_{\boldsymbol{w}}\ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})\|^2]/q_n^{(t)} - \|\nabla_{\boldsymbol{w}}f_n(\boldsymbol{w}^{(t)})\|^2) \\
&\overset{(c)}{\leq} \sum_{n\in\mathcal{N}} p_n^{(t)}\Psi(q_n^{(t)}), \quad (23)
\end{aligned}$$

where (a) is by Jensen's inequality, (b) is by the independence of client sampling and mini-batch data sampling, and (c) follows Assumptions 2 and 3 and the definition $\Psi(q_n^{(t)}) = (\frac{(1-q_n^{(t)})G_w^2}{q_n^{(t)}} + \frac{\sigma_{w,I}^2}{q_n^{(t)}})$.

Therefore, we have the following bound for C3

$$\mathbb{E}[\max_{\boldsymbol{w}\in\mathcal{W}}\mathrm{C3}] \leq \frac{R_{\mathcal{W}}}{T}\sqrt{\sum_{t=0}^{T-1}\sum_{n=1}^{N} p_n^{(t)}\Psi(q_n^{(t)})}. \quad (24)$$

Following similar proof techniques, we obtain the following bound for C4:

$$\mathbb{E}[\max_{\boldsymbol{p}\in\mathcal{P}}\mathrm{C4}] \leq \frac{R_{\mathcal{P}}\sigma_p}{\sqrt{T}}. \quad (25)$$

Finally, we note that C5 and C6 are not related to $\boldsymbol{w}$ and $\boldsymbol{p}$. Taking expectation on C5 and C6, we have $\mathbb{E}[\mathrm{C5}] = \mathbb{E}[\mathrm{C6}] = 0$ since $\tilde{\nabla}_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})$ and $\tilde{\nabla}_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})$ are unbiased estimators of $\nabla_{\boldsymbol{w}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})$ and $\nabla_{\boldsymbol{p}}F(\boldsymbol{w}^{(t)},\boldsymbol{p}^{(t)})$, respectively. Combining the bounds for C1-C6, we complete the proof. □

APPENDIX B
PROOF OF THEOREM 2

We start by bounding the terms related to the updates of $\boldsymbol{p}$. Following the update rule of $\boldsymbol{p}$, the fact that $F(\boldsymbol{w},\cdot)$ is

concave, and the Assumptions 2 and 3, we obtain, for all $\boldsymbol{p} \in \mathcal{P}$,

$$\mathbb{E}[\|\boldsymbol{p}^{(t+1)} - \boldsymbol{p}\|^2] \leq \|\boldsymbol{p}^{(t)} - \boldsymbol{p}\|^2 + \eta_p^2(G_p^2 + \sigma_p^2) + 2\eta_p(F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p})). \quad (26)$$

Rearranging the terms of (26), we have, for all $\boldsymbol{p} \in \mathcal{P}$,

$$-F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) \leq \frac{1}{2\eta_p}(\|\boldsymbol{p}^{(t)} - \boldsymbol{p}\|^2 - \mathbb{E}[\|\boldsymbol{p}^{(t+1)} - \boldsymbol{p}\|^2]) + \frac{\eta_p}{2}(G_p^2 + \sigma_p^2) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p}). \quad (27)$$

Let $\boldsymbol{p}^*(\boldsymbol{w}^{(t)}) = \arg\max_{\boldsymbol{p}\in\mathcal{P}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p})$. Setting $\boldsymbol{p}$ as $\boldsymbol{p}^*(\boldsymbol{w}^{(t)})$, adding $F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^*(\boldsymbol{w}^{(t)}))$ on both sides of (27), and using the fact $\Phi(\boldsymbol{w}^{(t)}) = \max_{\boldsymbol{p}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p})$, we obtain

$$\Phi(\boldsymbol{w}^{(t)}) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) \leq \|\boldsymbol{p}^{(t)} - \boldsymbol{p}^*(\boldsymbol{w}^{(t)})\|^2/(2\eta_p) - \mathbb{E}[\|\boldsymbol{p}^{(t+1)} - \boldsymbol{p}^*(\boldsymbol{w}^{(t)})\|^2]/(2\eta_p) + \frac{\eta_p}{2}(G_p^2 + \sigma_p^2). \quad (28)$$

Summing (28) over t from 0 to $T-1$, dividing both sides by T, and taking total expectation, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\Phi(\boldsymbol{w}^{(t)}) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\right] \leq \frac{R_{\mathcal{P}}^2}{2\eta_p T} + \frac{\eta_p}{2}(G_p^2 + \sigma_p^2). \quad (29)$$

We then bound the term with respect to the updates of $\boldsymbol{w}$. Let $\boldsymbol{w}_*^{(t)} = \arg\min_{\boldsymbol{w}\in\mathcal{W}} \Phi(\boldsymbol{w}) + L\|\boldsymbol{w} - \boldsymbol{w}^{(t)}\|^2$. We have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t+1)}\|^2] \\
&\overset{(a)}{\leq} \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)} + \eta_w \tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\|^2] \\
&= \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)} + \eta_w \tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) \\
&\quad - \eta_w \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) + \eta_w \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\|^2] \\
&\leq \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] + \eta_w^2 \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2) \\
&\quad + 2\eta_w \mathbb{E}[(\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)})^T (\tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}))] \\
&\overset{(b)}{=} \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] + \eta_w^2 \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2) + 2\eta_w A_3^{(t)} \\
&\quad + 2\eta_w \mathbb{E}[(\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)})^T (\tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) - \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}))] \\
&\overset{(c)}{\leq} \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] + \eta_w^2 \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2) + 2\eta_w A_3^{(t)} \\
&\quad + \eta_w \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\| \|\tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) - \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\|] \\
&\overset{(d)}{\leq} \mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] + \eta_w^2 \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2) + 2\eta_w A_3^{(t)} \\
&\quad + \eta_w R_{\mathcal{W}} A_4^{(t)}, \quad (30)
\end{aligned}$$

where (a) is by the projection onto a convex set $\mathcal{W}$, (b) is by defining $A_3^{(t)} = \mathbb{E}[(\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)})^T \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})]$, (c) is by the Cauchy-Schwartz inequality, and (d) is by denoting $A_4^{(t)} = \mathbb{E}[\|\tilde{\nabla}_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) - \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\|]$.

We next bound $A_3^{(t)}$.

$$\begin{aligned}
&A_3^{(t)} \\
&\overset{(a)}{\leq} \mathbb{E}[F(\boldsymbol{w}_*^{(t)}, \boldsymbol{p}^{(t)})] - \mathbb{E}[F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})] + \frac{L}{2}\mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] \\
&\overset{(b)}{\leq} \mathbb{E}[\Phi(\boldsymbol{w}_*^{(t)})] + L\mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] - \mathbb{E}[F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})] \\
&\quad - \frac{L}{2}\mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] \\
&\overset{(c)}{\leq} \mathbb{E}[\Phi(\boldsymbol{w}^{(t)})] + L\mathbb{E}[\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t)}\|^2] - \mathbb{E}[F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})] \\
&\quad - \frac{L}{2}\mathbb{E}[\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] \\
&\overset{(d)}{\leq} \mathbb{E}[\Phi(\boldsymbol{w}^{(t)}) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}) - \frac{1}{8L}\|\nabla\Phi_{1/2L}(\boldsymbol{w}^{(t)})\|^2], \quad (31)
\end{aligned}$$

where (a) follows from the L-smoothness of $F(\cdot,\cdot)$, (b) is from the definition of $\Phi(\cdot)$, (c) is from the definition of $\boldsymbol{w}_*^{(t)}$, and (d) is because $\|\nabla\Phi_{1/2L}(\boldsymbol{w}^{(t)})\| = \|2L(\boldsymbol{w}^{(t)} - \boldsymbol{w}_*^{(t)})\|$.

We then bound $A_4^{(t)}$.

$$\begin{aligned}
A_4^{(t)} &= \mathbb{E}[\|\sum_{n\in\mathcal{N}} \frac{p_n^{(t)} a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}} \ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)})] - \nabla_{\boldsymbol{w}} F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\|] \\
&\leq \sum_{n\in\mathcal{N}} p_n^{(t)} \mathbb{E}[\|\frac{a_n^{(t)}}{q_n^{(t)}} \nabla_{\boldsymbol{w}} \ell_{n,J_n^{(t)}}(\boldsymbol{w}^{(t)}) - \nabla_{\boldsymbol{w}} f_n(\boldsymbol{w}^{(t)})\|] \\
&\leq \sum_{n\in\mathcal{N}} p_n^{(t)} \sqrt{\frac{(1-q_n^{(t)})G_w^2}{q_n^{(t)}} + \frac{\sigma_{w,I}^2}{q_n^{(t)}}}. \quad (32)
\end{aligned}$$

Now, we have

$$\begin{aligned}
&\mathbb{E}[\Phi_{1/2L}(\boldsymbol{w}^{(t+1)})] \\
&= \mathbb{E}[\min_{\boldsymbol{w}\in\mathcal{W}} \Phi(\boldsymbol{w}) + L\|\boldsymbol{w} - \boldsymbol{w}^{(t+1)}\|^2] \\
&\leq \mathbb{E}[\Phi(\boldsymbol{w}_*^{(t)}) + L\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t+1)}\|^2] \\
&\overset{(a)}{\leq} \mathbb{E}[\Phi(\boldsymbol{w}_*^{(t)}) + L\|\boldsymbol{w}_*^{(t)} - \boldsymbol{w}^{(t)}\|^2] + \mathbb{E}[2\eta_w L A_3^{(t)}] \\
&\quad + \eta_w^2 \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2)L + \eta_w R_{\mathcal{W}} L A_4^{(t)} \\
&\overset{(b)}{\leq} \mathbb{E}[\Phi_{1/2L}(\boldsymbol{w}^{(t)})] + \mathbb{E}[2\eta_w L(\Phi(\boldsymbol{w}^{(t)}) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)}))] \\
&\quad + \eta_w^2 \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2)L + \eta_w R_{\mathcal{W}} A_4^{(t)} L \\
&\quad - \frac{\eta_w}{4}\mathbb{E}[\|\nabla\Phi_{1/2L}(\boldsymbol{w}^{(t)})\|^2], \quad (33)
\end{aligned}$$

where (a) is by the bound in (30), and (b) is by the definition of $\nabla\Phi_{1/2L}(\cdot)$ and the bound of $A_3^{(t)}$ in (31). Rearranging the terms of (33), we have

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\Phi_{1/2L}(\boldsymbol{w}^{(t)})\|^2\right] &\leq 4\eta_w \sum_{n=1}^{N} \frac{p_n^{(t)}}{q_n^{(t)}}(G_w^2 + \sigma_{w,I}^2)L \\
&\quad + 4R_{\mathcal{W}} L A_4^{(t)} + \frac{4}{\eta_w}(\mathbb{E}[\Phi_{1/2L}(\boldsymbol{w}^{(t)})] - \mathbb{E}[\Phi_{1/2L}(\boldsymbol{w}^{(t+1)})]) \\
&\quad + 8L(\mathbb{E}[\Phi(\boldsymbol{w}^{(t)})] - \mathbb{E}[F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})]). \quad (34)
\end{aligned}$$

Summing (34) over t from 0 to $T-1$, dividing both sides by T, taking total expectation, and then utilizing the bound of $A_4^{(t)}$ in (32) and the bound of $\frac{1}{T}\sum_{t=0}^{T-1}\left[\Phi(\boldsymbol{w}^{(t)}) - F(\boldsymbol{w}^{(t)}, \boldsymbol{p}^{(t)})\right]$ in (29), we complete the proof.

REFERENCES

[1] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, 2019.

[2] T. Li, A. K. Sahu, A. S. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, pp. 50–60, 2020.

[3] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet *et al.*, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1-2, pp. 1–210, 2021.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017.

[5] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[6] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[7] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *Ann. Stat.*, vol. 49, no. 3, pp. 1378–1406, 2021.

[8] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019.

[9] Z. Sun and E. Wei, "A communication-efficient algorithm with linear convergence for federated minimax learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[10] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[11] Y. Deng and M. Mahdavi, "Local stochastic gradient descent ascent: Convergence analysis and communication efficiency," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2021.

[12] Y. Zhang, M. Qiu, and H. Gao, "Communication-efficient stochastic gradient descent ascent with momentum algorithms," in *Proc. Int. Jt. Conf. Artif. Intell. (IJCAI)*, 2023.

[13] P. Sharma, R. Panda, G. Joshi, and P. Varshney, "Federated minimax optimization: Improved convergence analyses and algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.

[14] X. Zhang, G. Mancino-Ball, N. S. Aybat, and Y. Xu, "Jointly improving the sample and communication complexities in decentralized stochastic minimax optimization," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2024.

[15] P. Sharma, R. Panda, and G. Joshi, "Federated minimax optimization with client heterogeneity," *Trans. Mach. Learn. Res.*, 2023.

[16] A. Beznosikov, V. Samokhin, and A. V. Gasnikov, "Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms," *ArXiv preprint 2010.13112*, 2022.

[17] A. Beznosikov, G. Scutari, A. Rogozin, and A. Gasnikov, "Distributed saddle-point problems under data similarity," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021.

[18] J. von Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische Annalen*, vol. 119, no. 1, pp. 295–320, 1928.

[19] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic game theory*. Cambridge University Press, 2007.

[20] A. Nemirovski, "Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM J. Optim.*, vol. 15, no. 1, pp. 229–251, 2004.

[21] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *J. Optim. Theory Appl.*, vol. 142, pp. 205–228, 2009.

[22] Y. Chen, G. Lan, and Y. Ouyang, "Optimal primal-dual methods for a class of saddle point problems," *SIAM J. Optim.*, vol. 24, no. 4, pp. 1779–1814, 2014.

[23] E. Y. Hamedani and N. S. Aybat, "A primal-dual algorithm with line search for general convex-concave saddle point problems," *SIAM J. Optim.*, vol. 31, no. 2, pp. 1299–1329, 2021.

[24] V. Dem'yanov and A. Pevnyi, "Numerical methods for finding saddle points," *USSR Comput. Math. Math. Phys.*, vol. 12, no. 5, pp. 11–52, 1972.

[25] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

[26] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, "Efficient algorithms for smooth minimax optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[27] T. Lin, C. Jin, and M. I. Jordan, "Near-optimal algorithms for minimax optimization," in *Proc. Conf. Learn. Theory (COLT)*, 2020.

[28] G. M. Korpelevich, "The extragradient method for finding saddle points and other problems," *Matecon*, vol. 12, pp. 747–756, 1976.

[29] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo *et al.*, "Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[30] A. Mokhtari, A. Ozdaglar, and S. Pattathil, "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2020.

[31] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, "Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems," *SIAM J. Optim.*, vol. 30, no. 4, pp. 3230–3251, 2020.

[32] C. Daskalakis and I. Panageas, "The limit points of (optimistic) gradient descent in min-max optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018.

[33] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019.

[34] Z. Chai, A. Ali, S. Zawad, S. Truex *et al.*, "Tifl: A tier-based federated learning system," in *Proc. Int. Symp. High-Perform. Parallel Distrib. Comput.*, 2020.

[35] M. Chen, Z. Yang, W. Saad, C. Yin *et al.*, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2021.

[36] C. T. Dinh, N. H. Tran, M. N. H. Nguyen, C. S. Hong *et al.*, "Federated learning over wireless networks: Convergence analysis and resource allocation," *IEEE/ACM Trans. Netw.*, vol. 29, no. 1, pp. 398–409, 2021.

[37] W. Chen, S. Horváth, and P. Richtárik, "Optimal client sampling for federated learning," *Trans. Mach. Learn. Res.*, 2022.

[38] J. Ren, Y. He, D. Wen, G. Yu *et al.*, "Scheduling for cellular federated edge learning with importance and channel awareness," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.

[39] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, 2021.

[40] B. Luo, W. Xiao, S. Wang, J. Huang *et al.*, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2022.

[41] M. Zhang, G. Zhu, S. Wang, J. Jiang *et al.*, "Communication-efficient federated edge learning via optimal probabilistic device scheduling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8536–8551, 2022.

[42] W. Xu, B. Liang, G. Boudreau, and H. Sokun, "Probabilistic client sampling and power allocation for wireless federated learning," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2023.

[43] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM J. Optim.*, vol. 29, no. 1, pp. 207–239, 2019.

[44] S. Diamond and S. Boyd, "Cvxpy: a python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2909–2913, 2016.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.

[46] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[47] G. Cohen, S. Afshar, J. Tapson, and A. V. Schaik, "Emnist: Extending mnist to handwritten letters," in *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, 2017.