2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Online Federated Learning

Aritra Mitra, Hamed Hassani and George J. Pappas

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683589

***Abstract*— Federated learning (FL) has recently emerged as a popular framework for training a model via periodic coordination between a set of clients and a central server. The training task is abstracted as an optimization problem and solved under the premise that clients have access to their data samples offline, and that such samples are generated statistically. Departing from this paradigm, we initiate the study of FL in uncertain environments, where the clients' local loss functions arrive in an online, streaming manner, and are revealed only once the clients make their model predictions. Moreover, unlike the standard FL setting, we make no statistical assumptions on the loss functions, and our performance measure of interest is an appropriately defined collective regret metric. To minimize this regret metric in a communication-efficient manner, we propose `FedOMD` – an online FL algorithm where, akin to the offline setting, clients perform multiple local processing steps before uploading their model predictions to the server. Crucially, `FedOMD` differs from existing FL algorithms in the nature of its processing step. We (i) prove sublinear regret bounds for `FedOMD` that match their centralized counterparts (up to constants) for both convex and strongly convex losses; and (ii) use our regret guarantees to derive high probability excess risk bounds that characterize the generalization ability of `FedOMD`. Our analysis reveals in a precise way the trade-offs between intermittent communication and performance measures such as regret and excess risk.**

I. Introduction

Federated learning (FL) involves training a statistical model via periodic coordination between a set of clients and a central server while keeping the raw training data localized on each client [1], [2]. As envisioned in [1], the goal of this framework is to exploit the wealth of data typically available on client machines (e.g. mobile phones) to learn a high-quality model. However, to reap the benefits of such shared training, one must tackle some of the core challenges that are intrinsic to the FL setting: (i) *client heterogeneity* that arises due to differences in the data sets of the clients; and (ii) *communication constraints* that are motivated by low bandwidth communication networks and limited battery levels of the devices under consideration. Over the last couple of years, several interesting algorithmic ideas [3]–[7] have been developed to mitigate the challenges described above.

A common assumption that underlies existing works on FL is that each client has access to its data samples offline, and that such samples are drawn i.i.d. from some (client-specific) distribution. In practice, however, the data generating distribution at each client may vary over time, and precise a priori modeling of such distribution shifts may be intractable. Moreover, IoT (internet of things) applications may require operating in complex, uncertain, and dynamic environments where data may not be statistically generated.

Given the above premise, we incorporate for the first time, ideas from online learning [8] into the federated setting. Concretely, we depart from the standard FL paradigm in the following ways. (i) In our formulation, the local loss functions of each client arrive in an online, streaming manner over time. The instantaneous local loss functions of the clients define a *new* global loss function at each time-step. (ii) We make no statistical assumptions on the loss functions; moreover, we do not assume that successive loss functions observed by a client are related in any way. (iii) In standard FL, the task of training a statistical model is abstracted as an optimization problem, that of minimizing a static global loss function. In contrast, the clients in our setting aim to minimize a collective regret metric defined via the sequence of global loss functions that get realized. The above task is complicated by the communication constraints of a typical FL architecture: clients can only communicate indirectly via the server, and such communication is sparse. While recent work [5], [7] has shown that for the standard FL setting one can essentially match centralized rates, analogous results for the online FL counterpart we have just described do not exist. We take the first steps towards bridging this gap.

Contributions: Our contribution is threefold. First, we initiate the study of FL in an online setting and propose `FedOMD` - an online FL algorithm where clients perform multiple local mirror descent steps in isolation before communicating with the server to synchronize their model predictions. While `FedOMD` differs in its local processing step from existing FL algorithms, it resembles them in its sparse communication structure. Our second contribution is to prove that `FedOMD` enjoys sublinear regret, with regret bounds that match their centralized counterparts (up to constants). Specifically, we establish $O(\sqrt{\tau T})$ and $O(\tau \log T)$ regret bounds for convex and strongly convex losses, respectively, where T is the time-horizon, and τ is the maximum communication gap between the clients and the server. Interestingly, interpreting τ as the 'delay' in feedback induced by sparse communication, our regret bounds resemble classical results on online learning under delays [9]–[11]. Finally, for the setting when data is generated in a statistical manner, we use our regret guarantees to derive high-probability bounds on the excess risk. Our result in this context complements related work in FL that typically provides guarantees only in expectation [3]–[5], [7]. Notably, our analysis reveals precise trade-offs between intermittent communication and performance measures such as regret and excess risk.

The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania. Email: {amitra20, hassani, pappasg}@seas.upenn.edu. This work was supported by NSF CPS Grant 1837253, NSF CAREER award CIF 1943064, and the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award FA9550-20-1-0111.

II. Setting and Problem Formulation

We consider a setting involving a central server, and a set $\mathcal{S} = \{1, \ldots, m\}$ of m clients. The model predictions of the clients belong to a compact convex set $\mathcal{X} \subset \mathbb{R}^n$ equipped with a norm $\|\cdot\|$.[1] At each time-step $t \in [T]$, every client $i \in \mathcal{S}$ makes a prediction $x_{i,t} \in \mathcal{X}$ based on the information it has acquired up to time-step $t-1$.[2] Once the predictions $\{x_{i,t}\}_{i\in\mathcal{S}}$ have been made, an instantaneous global loss function denoted by $l_t : \mathcal{X} \to \mathbb{R}$ gets realized. The function $l_t(\cdot)$ can be decomposed into instantaneous local loss functions for the clients as follows:

$$l_t(x) = \frac{1}{m}\sum_{i\in\mathcal{S}} l_{i,t}(x), x \in \mathcal{X}. \tag{1}$$

Here, $l_{i,t} : \mathcal{X} \to \mathbb{R}$ is a convex, differentiable function over $\mathcal{X}$, and represents the local loss function of client i at time t. The feedback that client i receives at time t is the gradient of its local loss function evaluated at $x_{i,t}$, namely $\nabla l_{i,t}(x_{i,t})$. Crucially, this feedback at time t is revealed to client i *only after* it has made its prediction $x_{i,t}$. Given the online, sequential nature of our setting, the **goal** of the clients is to minimize the following collective regret R_T:

$$R_T = \frac{1}{m}\sum_{t=1}^{T}\sum_{i=1}^{m} l_t(x_{i,t}) - \min_{u\in\mathcal{X}}\sum_{t=1}^{T} l_t(u). \tag{2}$$

The above notion of regret measures the difference between the total cost accrued by the clients over the time horizon T, and that of the optimal *fixed* decision in hindsight taken by a centralized entity with access to all the functions $\{l_{i,t}(\cdot)\}_{i\in\mathcal{S},t\in[T]}$ ahead of time.

Communication constraints: To minimize the regret R_T in (2), it is apparent that each client i requires some feedback regarding the observations made by the other clients. However, the clients cannot directly interact among themselves; instead, all information exchanges take place via the server. The server also plays the role of aggregating information received from the clients. Since reducing communication is a key concern in a federated setting, clients interact with the server *infrequently*. In particular, we restrict all server-client interactions to a set $\mathcal{I}$ of synchronization time-steps, where $\mathcal{I} = \{t_1, \ldots, t_r\}$, with $t_1 = 1$ and $t_r = T$. Thus, over the time horizon T, there are precisely r communication rounds. Between two successive rounds, each client performs certain local processing steps *in isolation*, just as in the standard FL setting. By a federated algorithm, we will imply an algorithm that adheres to the above communication structure.

Connections to related settings: Apart from the communication structure, another key feature of the standard FL setting we retain in our formulation is that of heterogeneity in the loss functions across clients. Indeed, we not only allow for different loss functions across clients, but also temporal variations in the functions seen by a given client. The main departures from standard FL setting are as follows. (i) Neither do we make any distributional assumptions on the loss functions $l_{i,t}(\cdot)$, nor do we assume any relationship between successive loss functions observed by a given client. (ii) Our problem is inherently *sequential* in nature, and hence, the goal is to minimize regret, not solve a static optimization problem. Finally, our setting also differs from the fully distributed peer-to-peer framework [12]–[14] where agents can communicate with their neighbors at each time-step. In contrast, our formulation precludes any communication whatsoever between successive communication rounds.

Problem: The problem of interest to us in this paper is to design a federated algorithm that guarantees *sublinear regret while respecting the imposed communication constraints*. In other words, we want R_T to grow at most sublinearly with T subject to the condition that all server-client interactions are restricted to time-steps in $\mathcal{I}$. The main technical challenge in achieving this objective is that each client i has to make a prediction at time t without any knowledge of its own local loss function at time t, while relying on the stale feedback of other clients from the most recent communication round.

Our goal is to also precisely characterize how the effect of stale, delayed information feedback shows up in the regret bounds. To provide a concrete answer to this question, let us define $\tau = \max_{l\in[r-1]} t_{l+1} - t_l$ as the maximum time gap between two successive communication rounds. Note that the quantity τ can be interpreted as the maximum 'delay' in information feedback due to infrequent communication. As such, we aim to formally answer through our analysis: *How does regret R_T depend on the maximum delay τ?*

III. Federated Online Mirror Descent

In this section, we will develop our proposed federated online learning algorithm. Central to our development will be the notion of Bregman Divergence. To introduce this concept, consider a 1-strongly convex function $\psi : \mathcal{X} \to \mathbb{R}$ over $\mathcal{X}$ with respect to the norm $\|\cdot\|$. That is, for all $x, y \in \mathcal{X}$,

$$\psi(y) \geq \psi(x) + \langle y - x, \nabla\psi(x)\rangle + \frac{1}{2}\|y - x\|^2.$$

We can now define the Bregman Divergence w.r.t. $\psi(\cdot)$ as

$$B_\psi(y, x) = \psi(y) - \psi(x) - \langle y - x, \nabla\psi(x)\rangle,$$

where $x, y \in \mathcal{X}$. In words, $B_\psi(y, x)$ is the tail of $\psi(y)$ beyond the first-order Taylor series approximation at x. Since $\psi(\cdot)$ is strongly-convex, it is easy to see that $B_\psi(y, x)$ is always non-negative, and equal to zero if and only if $y = x$. Thus, $B_\psi(y, x)$ can be thought of as a similarity measure between y and x. We are now equipped with all the ingredients needed to describe our algorithm.

The steps of our method are outlined in Algorithm 1. At each time-step t, every client $i \in \mathcal{S}$ makes a prediction $x_{i,t}$, and then receives as feedback $g_{i,t} \triangleq \nabla l_{i,t}(x_{i,t})$. Next, client i performs a local mirror descent step as per (3) to compute an auxiliary local variable $y_{i,t+1}$. In this step, the function

$$\phi_{i,t}(x) \triangleq \langle g_{i,t}, x\rangle + \frac{1}{\eta_t}B_\psi(x, x_{i,t}) \tag{6}$$

that client i minimizes over $\mathcal{X}$ is composed of two parts. The first part $\langle g_{i,t}, x\rangle$ accounts for the most recent piece of

[1] Whenever we talk of a norm in the paper, it will be precisely this norm.
[2] We use $[T]$ as a shorthand for the set $\{1, \ldots, T\}$.

Algorithm 1 FedOMD

Pick any $z \in \mathcal{X}$ and initialize $x_{i,1} = z, \forall i \in \mathcal{S}$.
for $t = 1, \ldots, T-1$ **do**
  for $i = 1, \ldots, m$ **do**
    Predict $x_{i,t}$
    Receive gradient $g_{i,t} \triangleq \nabla l_{i,t}(x_{i,t})$
    Compute auxiliary local variable $y_{i,t+1}$ as

$$y_{i,t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left(\langle g_{i,t}, x \rangle + \frac{1}{\eta_t} B_\psi(x, x_{i,t}) \right) \tag{3}$$

    if $t+1 \in \mathcal{I}$ **then** upload $y_{i,t+1}$ to server, receive $1/m \sum_{i \in \mathcal{S}} y_{i,t+1}$ from server, and update model as

$$x_{i,t+1} = \bar{y}_{t+1} \triangleq \frac{1}{m} \sum_{i=1}^{m} y_{i,t+1} \tag{4}$$

    else

$$x_{i,t+1} = y_{i,t+1} \tag{5}$$

    end if
  end for
end for

information received by client i, namely $g_{i,t}$. The Bregman Divergence term $B_\psi(x, x_{i,t})$, on the other hand, acts as a regularizer that prevents $y_{i,t+1}$ from drifting too far apart from the previous model prediction $x_{i,t}$, thereby lending stability to the updates. The trade-off between the two terms in (6) is controlled via the learning rate η_t. We will see later that the choice of η_t depends on properties of the local loss functions, as well as the maximum communication gap τ.

Notice that in the local processing step (3), client i only uses its own privately observed gradient $g_{i,t}$. However, to minimize R_T as defined in (2), we require the model predictions of client i to also incorporate information about the observations made by the other clients. To achieve this, at each synchronization step $t+1 \in \mathcal{I}$, each client $i \in \mathcal{S}$ uploads its local auxiliary variable $y_{i,t+1}$ to the server. The server then averages these local variables and broadcasts the "blended model prediction" $\bar{y}_{t+1} = 1/m \sum_{i \in \mathcal{S}} y_{i,t+1}$ to each client. This mixing operation at the server constitutes the only indirect feedback that each client obtains about the observations of the other clients. Between successive communication rounds, the prediction of each client is its own local auxiliary variable, as is evident from (5). We call the algorithm described above Federated Online Mirror Descent, or simply FedOMD.

At this stage, it is instructive to point out that the key difference between FedOMD, and existing FL algorithms such as FedAvg [1], FedProx [3], SCAFFOLD [5], FedSplit [6], and FedLin [7], is in the nature of the local processing step.[3] Indeed, while the latter algorithms are designed with the aim of solving an optimization problem, the local step (3) of our proposed algorithm FedOMD has a completely different objective, that of minimizing regret. The overall communication structure of FedOMD, however, resembles that of the existing FL algorithms mentioned above.

Having formally introduced FedOMD, we now proceed to show in the subsequent sections that it provides strong theoretical guarantees.

[3]For instance, in the popular FedAvg algorithm, a local processing step simply involves a client performing one iteration of SGD using an unbiased estimate of the gradient of its local loss function. The local iterates are then periodically synchronized via the server, similar to FedOMD.

IV. Regret Guarantees of FedOMD

To state our main results concerning the regret bounds of FedOMD, we will first need to make a few assumptions.

Assumption 1. *For every $i \in \mathcal{S}$, $l_{i,t}(\cdot)$ is L-Lipschitz w.r.t. $\|\cdot\|$, $\forall t \in [T]$, i.e., $\forall x, y \in \mathcal{X}$,*

$$|l_{i,t}(x) - l_{i,t}(y)| \leq L\|x - y\|.$$

Assumption 2. *Consider any $u \in \mathcal{X}$, and any m points $z_1, \ldots, z_m \in \mathcal{X}$. Let $w_1, \ldots, w_m$ be convex weights, i.e., $w_j \in [0,1], \forall j \in \{1, \ldots, m\}$, and $\sum_{j=1}^m w_j = 1$. Then, the Bregman Divergence $B_\psi(\cdot, \cdot)$ satisfies*

$$B_\psi(u, \sum_{j=1}^{m} w_j z_j) \leq \sum_{j=1}^{m} w_j B_\psi(u, z_j).$$

While Assumption 1 is standard in the analysis of online learning, Assumption 2 holds for common Bregman divergences such as the Euclidean distance and the KL-divergence. Recall that $\tau = \max_{l \in [r-1]} t_{l+1} - t_l$, and define $D^2 \triangleq \sup_{x,y \in \mathcal{X}} B_\psi(x, y)$. With the above assumptions and notation in place, we can now state our first main result.

Theorem 1. *Suppose the local loss functions $l_{i,t}(\cdot)$ are convex over $\mathcal{X}$, $\forall i \in \mathcal{S}$, and $\forall t \in [T]$. Let Assumptions 1 and 2 hold. Then, with $\eta_t = \eta, \forall t \in [T]$, FedOMD guarantees*

$$R_T \leq \frac{D^2}{\eta} + 5\eta L^2 \tau T. \tag{7}$$

In particular, with $\eta = \frac{D}{\sqrt{5\tau T L}}$, we have

$$R_T \leq 5LD\sqrt{\tau T}. \tag{8}$$

In a variety of applications, the local loss functions may exhibit strong convexity (for instance, in online least squares problems). For centralized online learning, it is well known that one can achieve $O(\log T)$ regret for the strongly convex setting [15], improving upon the $O(\sqrt{T})$ regret guarantee for the convex case [8]. It is natural to thus ask whether one can achieve logarithmic regret in the federated setting of interest to us by exploiting the curvature of the local loss functions. In what follows, we will show that this is indeed the case. To this end, we will make use of the following notion of strong convexity w.r.t. the function $\psi(\cdot)$.

Definition 1. *A function $f : \mathcal{X} \to \mathbb{R}$ is σ-strongly convex w.r.t. $\psi(\cdot)$ over $\mathcal{X}$ if $\forall x, y \in \mathcal{X}$,*

$$f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\sigma}{2} B_\psi(y, x). \tag{9}$$

As an example, when $\psi(x) = \|x\|_2^2$, (i.e., when the norm under consideration is the standard Euclidean norm) it is easy to verify that $B_\psi(y, x) = \|y - x\|_2^2$. In this case, Definition 1 boils down to the standard definition of strong convexity.

Theorem 2. *Suppose the local loss functions $l_{i,t}(\cdot)$ are σ-strongly convex over $\mathcal{X}$ w.r.t. $\psi(\cdot)$, $\forall i \in \mathcal{S}$, and $\forall t \in [T]$. Let Assumptions 1 and 2 hold. Then, with $\eta_t = \frac{2}{\sigma t}, \forall t \in [T]$, FedOMD guarantees*

$$R_T \leq \frac{17L^2\tau}{\sigma}(1 + \log T). \tag{10}$$

The proofs of Theorems 1 and 2 are deferred to Section VI. Next, we discuss some key implications of our results.

Discussion: Assuming τ is a constant, our results in Theorems 1 and 2 indicate that FedOMD enjoys sublinear regret, where the dependence of the regret bound on L, D, and T exactly matches the respective centralized counterparts [8], [15]. The most interesting connection of our results is to the literature on online learning under delayed feedback [9]–[11]. These works typically consider a single-agent that receives the loss associated with its current action at a future round, i.e., with some delay. When this delay is at most τ, one obtains regret bounds of $O(\sqrt{\tau T})$ and $O(\tau \log T)$ for the convex and strongly convex settings, respectively [9] - the *exact* same bounds that we obtain in our multi-client federated setting (see equations (8) and (10)). It is important to recognize that unlike [9]–[11] where communication does not play a role, delay in our model is *induced* precisely due to sparse communication, τ being the maximum such delay. Moreover, unlike [9]–[11], a client only gets to observe the average prediction as delayed feedback, but not the losses that its predictions incur on the functions of the other clients.

V. Statistical Guarantees of FedOMD

In Section IV, we showed that FedOMD guarantees sublinear regret without making any distributional assumptions on the clients' loss functions. In this section, we consider a more favorable scenario where the loss functions of the clients are statistically generated, and use the regret guarantees from Section IV to derive high-probability performance bounds for FedOMD. To formalize the setting, for each client i, let ξ_i be a random variable drawn from some space $\mathcal{D}_i$ according to the distribution $\mathcal{P}_i$. At each client i, we consider a bounded loss function $l_i : \mathcal{X} \times \mathcal{D}_i \to [0, B]$ that is convex in the first argument. Let us define the joint sample space $\mathcal{D} \triangleq \Pi_{i \in \mathcal{S}} \mathcal{D}_i$ and the product measure $\mathcal{P} \triangleq \Pi_{i \in \mathcal{S}} \mathcal{P}_i$. Let

$$L_i(x) = \mathbb{E}[l_i(x; \xi_i)]; \quad L(x) = \frac{1}{m}\sum_{i=1}^{m} L_i(x), \tag{11}$$

where the expectation is taken w.r.t. $\mathcal{P}$. We will refer to $L(x)$ as the *risk*. With the above notations in place, we can now describe the problem of interest to us in this section. Suppose at each time-step t, an m-tuple $(\xi_{1,t}, \ldots, \xi_{m,t})$ is drawn i.i.d. from $\mathcal{P}$. These samples define the instantaneous loss function at each client i, namely $l_{i,t}(x) \triangleq l_i(x; \xi_{i,t})$. Now suppose FedOMD is run on the above loss functions, and generates the sequence of predictions $\{x_{i,t}\}_{t \in [T]}$ at each client i. Based on these predictions, our goal is to obtain a *single* predictor $\hat{x}_T$ such that the *excess risk*

$$L(\hat{x}_T) - \min_{x \in \mathcal{X}} L(x) \tag{12}$$

is small with high probability. In other words, we seek to find a predictor $\hat{x}_T$ that generalizes well w.r.t. the joint distribution $\mathcal{P}$. Our next result identifies such a predictor; for the proof of this result, see Section VI.

Theorem 3. *For each $i \in \mathcal{S}$, let $l_i : \mathcal{X} \times \mathcal{D}_i \to [0, B]$ be a bounded function that is convex in its first argument. Suppose T m-tuples $(\xi_{1,t}, \ldots, \xi_{m,t}), t \in [T]$ are drawn i.i.d. from $\mathcal{P}$, and set $l_{i,t}(x) \triangleq l_i(x; \xi_{i,t}), \forall i \in \mathcal{S}, \forall t \in [T]$. With $\eta_t = \eta = \frac{D}{\sqrt{5\tau T L}}$, let $\{x_{i,t}\}_{t \in [T]}$ be the sequence of predictions generated by FedOMD at client i, when run on the above loss functions. Then, if Assumptions 1 and 2 hold, given any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,*

$$L(\hat{x}_T) - \min_{x \in \mathcal{X}} L(x) \leq 5LD\sqrt{\frac{\tau}{T}} + 2B\sqrt{\frac{2}{T}\log\frac{2m}{\delta}}, \tag{13}$$

where

$$\hat{x}_T = \frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m} x_{i,t}. \tag{14}$$

Discussion: There are a few salient features of the above result. First, it can be seen as an extension of the classical work [16] that studies the generalization capabilities of online learning algorithms, to the federated setting. Second, as a byproduct of the mirror-descent style of analysis we conduct for FedOMD, we are able to derive high probability excess risk bounds that do not require smoothness of the local loss functions, and work for any arbitrary norm $\|\cdot\|$. This stands in contrast to existing works on FL [3]–[5], [7] that typically only provide guarantees in expectation, while assuming smoothness and relying crucially on properties of the Euclidean norm for their analysis. Third, note that Eq. (13) reveals in a transparent way the dependence of our high probability guarantees on the maximum communication-induced delay τ. In this way, we precisely characterize the trade-offs between communication-efficiency (sparse communication) and performance measures such as regret (Theorems 1 and 2) and excess risk (Theorem 3). Fourth, using the regret guarantee from Theorem 2, one can derive a risk bound for strongly convex losses exactly in the same way as for convex losses in Theorem 3. Finally, it is easy to see that the predictor $\hat{x}_T$ in (14) can be easily computed at the server: it suffices for each client to maintain a running average of its predictions that it uploads to the server at time-step T.

VI. Analysis

In this section, we provide the proofs of all our technical results, starting with that of Theorem 1. We begin with a simple regret decomposition lemma.

Lemma 1. *Fix any $i \in \mathcal{S}$, $t \in [T]$, and $u \in \mathcal{X}$. Suppose the conditions in Theorem 1 hold. Then, we have*

$$l_t(x_{i,t}) - l_t(u) \leq \frac{1}{m}\sum_{j=1}^{m}\left(l_{j,t}(x_{j,t}) - l_{j,t}(u)\right) + L\|x_{i,t} - \bar{x}_t\| + \frac{L}{m}\sum_{j=1}^{m}\|x_{j,t} - \bar{x}_t\|, \tag{15}$$

where $\bar{x}_t = \frac{1}{m}\sum_{j=1}^{m} x_{j,t}$.

Proof. We start by noting that

$$\begin{aligned} l_t(x_{i,t}) - l_t(u) &= \frac{1}{m}\sum_{j=1}^{m}\left(l_{j,t}(x_{i,t}) - l_{j,t}(u)\right) \\ &= \frac{1}{m}\sum_{j=1}^{m}\left(l_{j,t}(x_{j,t}) - l_{j,t}(u)\right) \\ &\quad + \frac{1}{m}\sum_{j=1}^{m}\left(l_{j,t}(x_{i,t}) - l_{j,t}(x_{j,t})\right). \end{aligned} \tag{16}$$

To bound the second summation, we use the L-Lipschitz property of the local losses in Assumption 1 to obtain

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^{m}\left(l_{j,t}(x_{i,t}) - l_{j,t}(x_{j,t})\right) &\leq \frac{L}{m}\sum_{j=1}^{m}\|x_{i,t} - x_{j,t}\| \\ &\leq \frac{L}{m}\sum_{j=1}^{m}\left(\|x_{i,t} - \bar{x}_t\| + \|x_{j,t} - \bar{x}_t\|\right) \\ &\leq L\|x_{i,t} - \bar{x}_t\| + \frac{L}{m}\sum_{j=1}^{m}\|x_{j,t} - \bar{x}_t\|. \end{aligned} \tag{17}$$

Plugging in the above bound in (16) completes the proof. □

Observe that the last two terms in (15) depend on how much the local model predictions of the clients "drift" from their average value. We now proceed to bound this drift in the following lemma.

Lemma 2. *Fix any $i \in \mathcal{S}$, $t \in [T]$, and $u \in \mathcal{X}$. Suppose the conditions in Theorem 1 hold. Then, we have*

$$\|x_{i,t} - \bar{x}_t\| \leq 2\eta L\tau. \tag{18}$$

Proof. The proof of this lemma relies on two simple observations: (i) at the beginning of each round, all clients start from a synchronized iterate, and (ii) the Bregman divergence term in (3) ensures that the local predictions of each client do not change too much in each step. To formalize the argument, let us define

$$h(t) = \max\{t_p \in \mathcal{I} : t_p \leq t\} \tag{19}$$

as the most recent synchronization time-step prior to time-step t, inclusive of t. Now suppose $t \notin \mathcal{I}$, for if $t \in \mathcal{I}$, then $x_{i,t} = \bar{x}_t$ based on the mixing step in line 7 of FedOMD, and the bound in (18) holds trivially. For $t \notin \mathcal{I}$, observe that

$$x_{i,t} = \bar{x}_{h(t)} + \sum_{l=h(t)}^{t-1}\left(x_{i,l+1} - x_{i,l}\right), \tag{20}$$

where we used the fact that $x_{i,h(t)} = \bar{x}_{h(t)}, \forall i \in \mathcal{S}$. Averaging the above equation across clients yields

$$\bar{x}_t = \bar{x}_{h(t)} + \frac{1}{m}\sum_{j=1}^{m}\sum_{l=h(t)}^{t-1}\left(x_{j,l+1} - x_{j,l}\right). \tag{21}$$

Subtracting (21) from (20) and taking norms on both sides, we obtain

$$\begin{aligned} \|x_{i,t} - \bar{x}_t\| &\leq \left(1 - \frac{1}{m}\right)\sum_{l=h(t)}^{t-1}\|x_{i,l+1} - x_{i,l}\| \\ &\quad + \frac{1}{m}\sum_{j\in\mathcal{S}\setminus\{i\}}\sum_{l=h(t)}^{t-1}\|x_{j,l+1} - x_{j,l}\|. \end{aligned} \tag{22}$$

To bound $\|x_{i,l+1} - x_{i,l}\|$, note that $l+1 \notin \mathcal{I}$. This follows from the definition of $h(t)$, the fact that $t \notin \mathcal{I}$, and that $h(t) < l+1 \leq t$. Since $l+1 \notin \mathcal{I}$, we have $x_{i,l+1} = y_{i,l+1}$ based on line 8 of FedOMD. Based on (3), and the first order necessary condition for optimality, we must then have

$$\langle \nabla\phi_{i,l}(x_{i,l+1}), z - x_{i,l+1}\rangle \geq 0, \forall z \in \mathcal{X}, \tag{23}$$

where $\phi_{i,l}(\cdot)$ is as defined in (6). With $z = x_{i,l}$, the above inequality implies

$$\langle g_{i,l} + \frac{1}{\eta}\left(\nabla\psi(x_{i,l+1}) - \nabla\psi(x_{i,l})\right), x_{i,l} - x_{i,l+1}\rangle \geq 0. \tag{24}$$

Rearranging the above inequality, we obtain

$$\begin{aligned} &\langle \nabla\psi(x_{i,l+1}) - \nabla\psi(x_{i,l}),\, x_{i,l+1} - x_{i,l}\rangle \\ &\qquad \leq \eta\langle g_{i,l}, x_{i,l} - x_{i,l+1}\rangle \\ &\qquad \leq \eta\|x_{i,l+1} - x_{i,l}\|\|g_{i,l}\|_* \\ &\qquad \leq \eta L\|x_{i,l+1} - x_{i,l}\|. \end{aligned} \tag{25}$$

For the second step, we used Hölder's inequality; here, $\|\cdot\|_*$ is the dual norm of the norm $\|\cdot\|$. For the final step, we used the fact that the local loss function of client i is L-Lipschitz, which, in turn, implies a uniform bound of L on the magnitude of the gradient $g_{i,l}$ measured w.r.t. the dual norm $\|\cdot\|_*$. From the strong convexity of $\psi(\cdot)$, we also have

$$\langle \nabla\psi(x_{i,l+1}) - \nabla\psi(x_{i,l}),\, x_{i,l+1} - x_{i,l}\rangle \geq \|x_{i,l+1} - x_{i,l}\|^2.$$

Combining the above inequality with (25) yields

$$\|x_{i,l+1} - x_{i,l}\| \leq \eta L. \tag{26}$$

The above analysis applies identically to each client, and hence, plugging the bound from (26) in (22), and simplifying, we obtain

$$\|x_{i,t} - \bar{x}_t\| \leq 2\eta L(t - h(t)) \leq 2\eta L\tau, \tag{27}$$

where we used the fact that the maximum gap between two consecutive synchronization steps is τ. □

In the following lemma, we obtain a bound on the first term in (15).

Lemma 3. *Fix any $u \in \mathcal{X}$ and $t \in [T]$, and suppose the conditions in Theorem 1 hold. We then have*

$$\frac{1}{m}\sum_{j=1}^{m}(l_{j,t}(x_{j,t}) - l_{j,t}(u)) \leq \frac{1}{\eta m}\sum_{j=1}^{m}(B_\psi(u, x_{j,t}) - B_\psi(u, x_{j,t+1})) + \frac{1}{2}\eta L^2. \tag{28}$$

Proof. Fix any $j \in \mathcal{S}$. Based on the convexity of $l_{j,t}(\cdot)$,

$$\begin{aligned} l_{j,t}(x_{j,t}) - l_{j,t}(u) &\leq \langle x_{j,t} - u, g_{j,t}\rangle \\ &= \langle x_{j,t} - y_{j,t+1}, g_{j,t}\rangle + \langle y_{j,t+1} - u, g_{j,t}\rangle. \end{aligned} \tag{29}$$

To bound the first term in the resulting inequality above, we use the Fenchel-Young inequality to obtain:

$$\begin{aligned} \langle x_{j,t} - y_{j,t+1}, g_{j,t}\rangle &\leq \frac{1}{2\eta}\|x_{j,t} - y_{j,t+1}\|^2 + \frac{1}{2}\eta\|g_{j,t}\|_*^2, \\ &\leq \frac{1}{2\eta}\|x_{j,t} - y_{j,t+1}\|^2 + \frac{1}{2}\eta L^2, \end{aligned} \tag{30}$$

where in the second step we used Assumption 1. As for the second term in (29), we use (3), and appeal to the first order condition of optimality[4] to obtain

$$\begin{aligned} \langle y_{j,t+1} - u, g_{j,t}\rangle &\leq \frac{1}{\eta}\langle \nabla\psi(y_{j,t+1}) - \nabla\psi(x_{j,t}), u - y_{j,t+1}\rangle \\ &\overset{(a)}{=} \frac{1}{\eta}\left(B_\psi(u, x_{j,t}) - B_\psi(u, y_{j,t+1})\right) \\ &\quad - \frac{1}{\eta}B_\psi(y_{j,t+1}, x_{j,t}) \\ &\overset{(b)}{\leq} \frac{1}{\eta}\left(B_\psi(u, x_{j,t}) - B_\psi(u, y_{j,t+1})\right) \\ &\quad - \frac{1}{2\eta}\|x_{j,t} - y_{j,t+1}\|^2. \end{aligned} \tag{31}$$

For (a), we used the "three-point equality" for Bregman divergences which states that for any $x, y, z \in \mathcal{X}$,

$$\langle \nabla\psi(x) - \nabla\psi(y), x - z\rangle = B_\psi(x, y) + B_\psi(z, x) - B_\psi(z, y).$$

For (b), we used the strong convexity of $\psi(\cdot)$. Combining the bounds in (29), (30), and (31), performing simple algebraic manipulations, and averaging across clients, we obtain

$$\begin{aligned} \frac{1}{m}\sum_{j=1}^{m}(l_{j,t}(x_{j,t}) - l_{j,t}(u)) &\leq \frac{1}{\eta m}\sum_{j=1}^{m}(B_\psi(u, x_{j,t}) - B_\psi(u, x_{j,t+1})) \\ &\quad + \underbrace{\frac{1}{\eta m}\sum_{j=1}^{m}(B_\psi(u, x_{j,t+1}) - B_\psi(u, y_{j,t+1}))}_{P} \\ &\quad + \frac{1}{2}\eta L^2. \end{aligned} \tag{32}$$

To prove the lemma, it remains to show that $P \leq 0$ in the above inequality. To see that this is indeed true, first consider the case when $t+1 \notin \mathcal{I}$. From (5), we have $x_{j,t+1} = y_{j,t+1}, \forall j \in \mathcal{S}$, and hence, $P = 0$. When $t+1 \in \mathcal{I}$, we have from (4) that $x_{j,t+1} = \bar{y}_{t+1} = 1/m\sum_{i=1}^{m} y_{i,t+1}, \forall j \in \mathcal{S}$. In this case, we have

$$P = \frac{1}{\eta}\left(B_\psi(u, \bar{y}_{t+1}) - \frac{1}{m}\sum_{j=1}^{m}B_\psi(u, y_{j,t+1})\right).$$

The fact that $P \leq 0$ then follows immediately by appealing to Assumption 2. This completes the proof. □

We now have all the pieces required to prove Theorem 1.

Proof. (**Theorem 1**) Substituting the bounds in (18) and (28) from Lemmas 2 and 3, respectively, into (15), we obtain

$$\begin{aligned} l_t(x_{i,t}) - l_t(u) &\leq \frac{1}{\eta m}\sum_{j=1}^{m}(B_\psi(u, x_{j,t}) - B_\psi(u, x_{j,t+1})) \\ &\quad + (4\tau + 1/2)\eta L^2. \end{aligned} \tag{33}$$

Summing the above inequality from $t = 1$ to T leads to a telescoping sum on the right, and we are left with

$$\begin{aligned} \sum_{t=1}^{T}(l_t(x_{i,t}) - l_t(u)) &\leq \frac{1}{\eta m}\sum_{j=1}^{m}(B_\psi(u, x_{j,1}) - B_\psi(u, x_{j,T+1})) \\ &\quad + 5\eta L^2\tau T \\ &\leq \frac{D^2}{\eta} + 5\eta L^2\tau T, \end{aligned} \tag{34}$$

where we used $\tau \geq 1$ and the definition of D. Averaging the final inequality over clients, we obtain

$$\frac{1}{m}\sum_{t=1}^{T}\sum_{i=1}^{m}l_t(x_{i,t}) - \sum_{t=1}^{T}l_t(u) \leq \frac{D^2}{\eta} + 5\eta L^2\tau T. \tag{35}$$

Noting that the above analysis holds for any $u \in \mathcal{X}$, we obtain the desired result. □

We now turn to the proof of Theorem 2.

Proof. (**Theorem 2**) The key difference in the analysis w.r.t. that of Theorem 1 is that now the learning rate is time-varying. Nonetheless, the regret decomposition result in Lemma 1 remains unchanged. Following the drift analysis in Lemma 2, we now have $\|x_{i,l+1} - x_{i,l}\| \leq \eta_l L$ in place of (26), where $h(t) \leq l \leq t-1$. Substituting this bound in (22) and simplifying, we obtain $\forall i \in \mathcal{S}$, and $\forall t \in [T]$,

$$\|x_{i,t} - \bar{x}_t\| \leq 2L\sum_{l=h(t)}^{t-1}\eta_l. \tag{36}$$

Next, using the fact that $l_{j,t}(\cdot)$ is strongly convex w.r.t. $\psi(\cdot), \forall i \in \mathcal{S}, \forall t \in [T]$, and repeating the analysis in Lemma

[4]Specifically, we use (23) with $y_{i,t+1}$ in place of $x_{i,l+1}$, and with $z = u$.

3, one can easily verify that

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}\left(l_{j,t}(x_{j,t})-l_{j,t}(u)\right) &\leq \frac{1}{\eta_t m}\sum_{j=1}^{m}\left(B_\psi(u,x_{j,t})-B_\psi(u,x_{j,t+1})\right)\\ &\quad+\frac{1}{2}\eta_t L^2-\frac{\sigma}{2m}\sum_{j=1}^{m}B_\psi(u,x_{j,t})\\ &=\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-\frac{\sigma}{2}\right)B_\psi(u,x_{j,t})\\ &\quad+\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{\eta_{t-1}}B_\psi(u,x_{j,t})-\frac{1}{\eta_t}B_\psi(u,x_{j,t+1})\right)\\ &\quad+\frac{1}{2}\eta_t L^2.\end{aligned} \tag{37}$$

Combining the drift bound with the above inequality yields

$$\begin{aligned}l_t(x_{j,t})-l_t(u) &\leq \left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-\frac{\sigma}{2}\right)D^2\\ &\quad+\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{\eta_{t-1}}B_\psi(u,x_{j,t})-\frac{1}{\eta_t}B_\psi(u,x_{j,t+1})\right)\\ &\quad+\frac{1}{2}\eta_t L^2+4L^2\sum_{l=h(t)}^{t-1}\eta_l,\end{aligned} \tag{38}$$

where we used the definition of D. Summing the above inequality from $t=1$ to T, and using $1/\eta_0 \triangleq 0$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\left(l_t(x_{j,t})-l_t(u)\right) &\leq \left(\frac{1}{\eta_T}-\frac{\sigma}{2}T\right)D^2+\frac{1}{2}L^2\sum_{t=1}^{T}\eta_t\\ &\quad+4L^2\sum_{t=1}^{T}\sum_{l=h(t)}^{t-1}\eta_l.\end{aligned} \tag{39}$$

Setting $\eta_t=\frac{2}{\sigma t}$ in the above inequality causes the first term to vanish. For the second term, we have

$$\frac{1}{2}L^2\sum_{t=1}^{T}\eta_t=\frac{L^2}{\sigma}\sum_{t=1}^{T}\frac{1}{t}\leq\frac{L^2}{\sigma}\left(1+\log T\right).$$

It remains to bound the third term in (39). Observe that

$$\begin{aligned}\sum_{t=1}^{T}\sum_{l=h(t)}^{t-1}\eta_l &= \left(\sum_{t=t_1}^{t_2-1}\sum_{l=h(t)}^{t-1}\eta_l+\cdots+\sum_{t=t_{r-1}}^{T}\sum_{l=h(t)}^{t-1}\eta_l\right)\\ &\leq\frac{2}{\sigma}\left(\sum_{t=t_1}^{t_2-1}\sum_{l=t_1}^{t-1}\frac{1}{l}+\cdots+\sum_{t=t_{r-1}}^{T}\sum_{l=t_{r-1}}^{T}\frac{1}{l}\right)\\ &\leq\frac{2}{\sigma}\Bigg((t_2-t_1)\left(1+\log\left(t_2-1\right)\right)\\ &\quad+\cdots+(T-t_{r-1}+1)\log\frac{T}{t_{r-1}-1}\Bigg)\\ &\leq\frac{2(\tau+1)}{\sigma}\left(1+\log T\right).\end{aligned} \tag{40}$$

For the second inequality above, we used the fact that for t such that $t_p\leq t\leq t_{p+1}-1$, we have $h(t)=t_p$. For the final step, we used the fact that $t_{p+1}-t_p\leq\tau,\forall p\in[r-1]$. Plugging in the bound (40) in equation (39), averaging the resulting inequality across clients, and simplifying, we are led to the assertion of Theorem 2. □

The proof of Theorem 3, which we provide next, is an adaptation of the arguments in [16] to our setting.

Proof. (**Theorem 3**) The key idea behind the proof is to define an appropriate martingale sequence and then appeal to the Azuma-Hoeffding inequality for martingales with bounded increments [17]. To get started, define $V_0=0$, and

$$V_t=\frac{1}{m}\sum_{r=1}^{t}\sum_{i=1}^{m}\left(L(x_{i,r})-l_r(x_{i,r})\right),\forall t\in[T].$$

Next, let $\mathcal{F}_0=\{\emptyset,\mathcal{D}\}$ be the trivial σ-algebra, and $\mathcal{F}_t=\sigma(\{\xi_{i,r}\}_{i\in\mathcal{S},r\in[t]})$ be the σ-algebra generated by all the data samples up to time t. We claim that the process $V=(V_t:0\leq t\leq T)$ is a martingale relative to the filtration $\{\mathcal{F}_t:0\leq t\leq T\}$. To see this, observe that

$$\begin{aligned}\mathbb{E}\left[l_{t+1}(x_{i,t+1})|\mathcal{F}_t\right] &= \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[l_j(x_{i,t+1};\xi_{j,t+1})|\mathcal{F}_t\right]\\ &=\frac{1}{m}\sum_{j=1}^{m}L_j(x_{i,t+1})=L(x_{i,t+1}),\end{aligned} \tag{41}$$

where we used the fact that $x_{i,t+1}$ is $\mathcal{F}_t$-measurable. Based on the definition of V_t, it is then easy to see that $\mathbb{E}[V_{t+1}|\mathcal{F}_t]=V_t$, establishing the martingale property. Under the boundedness assumption on the loss functions, we also have $|V_{t+1}-V_t|\leq B,\forall t\in\{0,\ldots,T-1\}$. The Azuma-Hoeffding inequality for martingales then tells us that

$$\mathcal{P}\{V_T-V_0\geq\epsilon\}\leq\exp\left(-\frac{\epsilon^2T}{2B^2}\right). \tag{42}$$

We then have with probability at least $1-\delta/2$ that

$$\begin{aligned}\frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m}L(x_{i,t}) &\overset{(a)}{\leq}\frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m}l_t(x_{i,t})+B\sqrt{\frac{2}{T}\log\frac{2}{\delta}}\\ &\overset{(b)}{\leq}5LD\sqrt{\frac{\tau}{T}}+\frac{1}{T}\sum_{t=1}^{T}l_t(x^*)+B\sqrt{\frac{2}{T}\log\frac{2}{\delta}},\end{aligned} \tag{43}$$

where $x^*=\operatorname{argmin}_{x\in\mathcal{X}}L(x)$. For (a), we used $V_0=0$, and for (b), we used the regret bound from Theorem 1 by noting that our regret analysis applies w.r.t. *any* fixed comparator $u\in\mathcal{X}$, and hence $u=x^*$ in particular (see equation (35)). It remains to argue that $\frac{1}{T}\sum_{t=1}^{T}l_t(x^*)$ concentrates around $L(x^*)$. To this end, start by noting that

$$\frac{1}{T}\sum_{t=1}^{T}l_t(x^*)=\frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m}l_i(x^*;\xi_{i,t}). \tag{44}$$

Fix a client i. Since $\mathbb{E}[l_i(x^*;\xi_{i,t})]=L_i(x^*),\forall t\in[T]$, a simple application of Hoeffding's inequality yields

$$\mathcal{P}\Big\{\frac{1}{T}\sum_{t=1}^{T}l_i(x^*;\xi_{i,t})-L_i(x^*)\geq\epsilon\Big\}\leq\exp\left(-\frac{\epsilon^2T}{2B^2}\right). \tag{45}$$

Thus, with probability at least $1-\frac{\delta}{2m}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}l_i(x^*;\xi_{i,t})\leq L_i(x^*)+B\sqrt{\frac{2}{T}\log\frac{2m}{\delta}}.$$

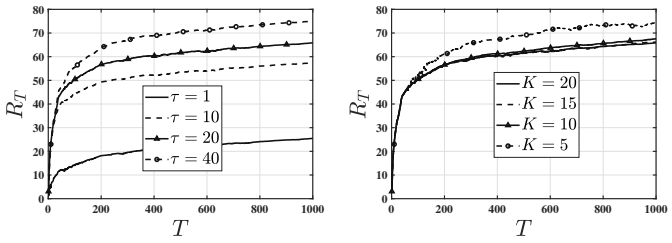

Fig. 1. Plots of regret R_T versus time horizon T for an instance with $m = 20$ clients, and loss functions as defined in (46). Communication is periodic with period τ. **Left:** Variation in regret with communication period τ. **Right:** Variation in regret with number of participating clients K.

Taking an union bound over the m clients, and using (44), observe that with probability at least $1 - \delta/2$, we have

$$\frac{1}{T}\sum_{t=1}^{T} l_t(x^*) \leq L(x^*) + B\sqrt{\frac{2}{T}\log\frac{2m}{\delta}}.$$

Combining the above inequality with (43), and applying the union bound, we have with probability at least $1 - \delta$,

$$\frac{1}{mT}\sum_{t=1}^{T}\sum_{i=1}^{m} L(x_{i,t}) - L(x^*) \leq 5LD\sqrt{\frac{\tau}{T}} + 2B\sqrt{\frac{2}{T}\log\frac{2m}{\delta}}.$$

The claim in (13) then follows by noting that $L(\cdot)$ is convex, and using Jensen's inequality. □

VII. Simulations

We consider a setting with $m = 20$ clients that communicate periodically with the server (as is typically done in FL) with period equal to τ, i.e., $t_{l+1} - t_l = \tau, \forall l \in [r-1]$. The set $\mathcal{X}$ is the interval $[-3, 3]$. At each time-step t, for each client $i \in \mathcal{S}$, we first generate $a_{i,t} \sim \mathcal{N}(2, 5)$. The local loss function $l_{i,t}(x)$ at client i is then given by

$$l_{i,t}(x) = \begin{cases} \frac{1}{2}(x - a_{i,t})^2, & \text{if } t \text{ is even} \\ \frac{1}{2}(x + a_{i,t})^2, & \text{if } t \text{ is odd.} \end{cases} \tag{46}$$

Thus, at each time-step t, the loss functions across clients vary since $a_{i,t}$ is chosen randomly. Moreover, for a given client i, its loss function switches every time-step. In this way, the sequence of loss functions we consider capture heterogeneity both across clients, and over time. We first study the effect of increasing the communication period τ. From Fig. 1, we note that larger τ leads to larger regret, as one would naturally expect. Next, with τ kept fixed at 20, we explore the robustness of `FedOMD` to *partial client participation* - a common feature in FL. Specifically, at each $t \in \mathcal{I}$, suppose $1 \leq K \leq m$ clients are chosen by the server uniformly at random from the set $\mathcal{S}$. Only these clients upload their predictions to the server, the server averages these predictions, and broadcasts the average to each client. From Fig. 1, we observe that smaller K leads to larger regret, aligning with intuition. For all our simulations, we observe from Fig. 1 that R_T scales logarithmically with T, conforming with Theorem 2.

VIII. Conclusion

We studied, for the first time, federated learning in an online setting, i.e., without making any statistical assumptions on the clients' data. We proposed an online FL algorithm `FedOMD` that retains the sparse communication structure of its offline counterparts, while differing in its local processing step. We proved that `FedOMD` guarantees sublinear regret that matches centralized regret bounds for both convex and strongly convex losses, and characterized how sparse communication inflates regret. Finally, we used our regret guarantees to obtain high-probability excess risk bounds for `FedOMD`. Our work establishes a simple template for synthesizing and analyzing online FL algorithms, paving the way for studying more interesting settings as future work: adaptive online algorithms and bandit feedback.

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] 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.

[3] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "On the convergence of federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, vol. 3, 2018.

[4] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2021–2031.

[5] 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.

[6] R. Pathak and M. J. Wainwright, "Fedsplit: An algorithmic framework for fast federated optimization," *arXiv:2005.05238*, 2020.

[7] A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani, "Achieving linear convergence in federated learning under objective and systems heterogeneity," *arXiv preprint arXiv:2102.07053*, 2021.

[8] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the 20th international conference on machine learning (icml-03)*, 2003, pp. 928–936.

[9] M. Zinkevich, A. J. Smola, and J. Langford, "Slow learners are fast," in *NIPS*, 2009.

[10] P. Joulani, A. Gyorgy, and C. Szepesvári, "Online learning under delayed feedback," in *International Conference on Machine Learning*. PMLR, 2013, pp. 1453–1461.

[11] K. Quanrud and D. Khashabi, "Online learning with adversarial delays." in *NIPS*, 2015, pp. 1270–1278.

[12] D. Mateos-Núnez and J. Cortés, "Distributed online convex optimization over jointly connected digraphs," *IEEE Transactions on Network Science and Engineering*, vol. 1, no. 1, pp. 23–37, 2014.

[13] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 714–725, 2017.

[14] S. Lee, A. Nedić, and M. Raginsky, "Stochastic dual averaging for decentralized online optimization on time-varying communication graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6407–6414, 2017.

[15] E. Hazan, A. Agarwal, and S. Kale, "Logarithmic regret algorithms for online convex optimization," *Machine Learning*, vol. 69, no. 2-3, pp. 169–192, 2007.

[16] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of on-line learning algorithms," *IEEE Transactions on Information Theory*, vol. 50, no. 9, pp. 2050–2057, 2004.

[17] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math Journal, 2nd Series*, vol. 19, no. 3, pp. 357–367, 1967.