Towards Federated Learning on Fresh Datasets

Chen Wu*, Mingjun Xiao*†‡, Jie Wu§, Yin Xu†‡, Jinrui Zhou†‡, and He Sun†‡
*School of Data Science, University of Science and Technology of China, China
†School of Computer Science and Technology, University of Science and Technology of China, China
‡Suzhou Institute for Advanced Research, University of Science and Technology of China, China
§Department of Computer and Information Sciences, Temple University, USA
Corresponding Email: xiaomj@ustc.edu.cn

***Abstract*—Federated Learning (FL) is an emerging privacy-preserving distributed computing paradigm that enables numerous clients collaboratively to train machine learning models without the need of transmitting the private datasets of clients to the FL server. Unlike most existing researches where the local datasets of clients are considered to be unchanging over time during the whole FL process, we consider such scenarios in this paper that the local datasets of clients need to be updated periodically, and the server can stimulate clients to use as fresh as possible datasets to train their local models. Our objective is to determine a client selection strategy to minimize the loss of global model for the FL with a limited budget. To this end, we leverage the concept of Age of Information (AoI) to quantify the freshness of local datasets and theoretically analyze the convergence bound of our AoI-aware FL system. Based on the convergence bound, we formalize our problem as a restless multi-armed bandit problem. Then, we devise a Whittle's-Index-based Client Selection algorithm, called WICS, to tackle the client selection problem. Extensive simulations show that the proposed algorithm can reduce the training loss and improve the learning accuracy compared to those state-of-the-art algorithms.**

***Index Terms*—Federated Learning, Age of Information, Restless Multi-Armed Bandit, Whittle's Index.**

I. Introduction

Federated Learning (FL) [1] is an emerging and promising distributed machine learning paradigm, which enables a potentially large number of clients to collaboratively train a global model under the coordination of a central server. A standard FL procedure usually consists of a certain number of rounds until a satisfactory global model is obtained. In each round, clients train local models on their local datasets, typically by means of Stochastic Gradient Descent (SGD). Then, the central server aggregates these local models to obtain the global model, which is transferred back to each client for the next round of local training. On one hand, FL can efficiently preserve clients' privacy by allowing the training datasets to remain at local. On the other hand, since only local model parameters rather than local datasets are sent to the server, FL can greatly reduce the communication costs. Due to the above advantages, there have been various industrial applications of FL, e.g., WeBank for data analysis in finance and insurance [2], Owkin for biomedical data analysis [3], MELLODDY for drug discovery [4], etc. Meanwhile, much effort has also been devoted to investigating different FL issues [17], [18], including the convergence rate [5]–[7], accuracy [8]–[10], [13], security [11], [12], resource allocation [14]–[16], and so on.

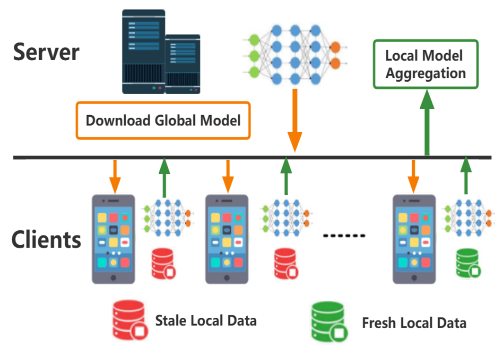


Fig. 1. The architecture of FL training with fresh/stale local data

In most existing works, each client is assumed to hold a dataset in advance and will always use the same dataset to train its local model during the whole FL process. However, in many real-world applications, especially in streaming data scenarios, data are continuously generated along with the time. Clients usually prefer to update their datasets periodically rather than keeping their local datasets unchanging. When participating in FL, clients are encouraged to use as fresh datasets as possible to train local models since fresh data can more accurately characterize the model parameters. For example, a server coordinates some clients to jointly train an object identification model through FL, e.g., recognizing formulas on literatures, identifying traffic signs on photos, etc., where clients can adopt the crowdsourcing technique to periodically recruit mobile users to generate labeled datasets. Intuitively, the fresher the labeled datasets, the more users have participated in the data labelling, and thus the labeled datasets will be more precise. In such kind of FL scenarios, clients will inevitably spend some extra costs in providing fresh datasets, but the total budget from the server is generally limited. Thus, an important problem that needs to be dealt with is how to select clients to use as fresh datasets as possible for the model training in each round of FL under the limited budget, and the server can minimize the loss of global model.

In this paper, we use the well-known "Age-of-Information" (AoI [19]) metric to indicate the freshness of datasets, which is defined as the elapsed time of data from being collected to being trained for updating local models by clients. The smaller the AoI value of a dataset, the fresher the corresponding data, and thus the more precise the trained local models. Accordingly, the above-mentioned problem is actually instantiated as determining a client selection strategy to minimize the loss of the global model under a given budget, while taking the AoI values of the datasets into consideration. Unlike most traditional optimal selection issues with budget constraints, such a problem has two following special challenges. First, although the server can reduce the loss of global model by selecting some clients in each round of FL to update their local datasets and reduce the AoI values, there is no obviously quantitative relationship between the loss of global model and the decrease of the AoI values of clients' datasets. Second, the AoI value of each client's dataset will increase along with the rounds of local training and will return to zero until it is selected to update the dataset. This means that the client selections and the corresponding AoI values are not independent with each other across different rounds of FL. Both of them make the client selection problem much more challenging, especially under the budget constraint.

To address the above challenges, we first derive a convergence upper bound for the novel AoI-aware FL system. The upper bound shows that the loss of global model is positively correlated to the freshness of local datasets, i.e., inversely correlated to the AoI values of local datasets. On this basis, we transform the optimal client selection problem to minimize the loss of global model to the problem of optimal client selection with minimum average AoI value. Furthermore, we model the problem as a Rsetless Multi-Armed Bandit (RMAB) problem, where each client is seen as an arm and the AoI values of clients' local datasets are regarded as the corresponding rewards. By solving the RMAB problem, we propsed a Whittle's-Index-based Client Selection algorithm, called WICS, in which we calculate the Whittle's Index for each client in each round of FL and then adopt a greedy strategy based on Whittle's Index to select clients while ensuring the budget no larger than the given threshold. More specifically, the major contributions are summarized as follows:

- We introduce an AoI-aware FL system, where the server can select some clients to provide fresh datasets for local model training so as to minimize the loss of global model under a budget constraint. To the best of our knowledge, this is the first FL work that takes into consideration the freshness of the local datasets for client selection.
- We derive a convergence upper bound for the novel AoI-aware FL system, whereby we analyze the relationship between the training loss of global model and the AoI values of clients' local datasets. Based on the analysis, we model the client selection problem as a restless multi-armed bandit problem to be solved.
- We deduct the RMAB problem into a decoupled model and theoretically derive the corresponding optimal strategy, based on which, we propose the WICS algorithm by applying the Whittle's Index methodology. Moreover, we prove that WICS can achieve nearly optimal client selection performance.
- We conduct extensive simulations to verify the performance of our proposed algorithm using two popular datasets MNIST and FMNIST. The results show that the performance of WICS is better than those of state-of-the-art algorithms.

The remainder of the paper is organized as follows. In Section II, we introduce our model and problem formulation. We carry on the convergence analysis of AoI-aware FL system in III. The detailed design of the WICS algorithm is elaborated in Section IV. Then we present the performance evaluations in Section V. After reviewing the related works in Section VI, we conclude the paper in Section VII. Some complicated proofs of theorems are moved to Appendix.

II. System Overview And Problem Formulation

A. Federated Learning with Data Collection

We consider an AoI-aware FL system, as shown in Fig.1, which is composed of a central server and a set of clients, denoted by $\mathcal{N} = \{1, 2, \cdots, N\}$. In conventional FL systems, the local dataset of each client is generally given in advance and will keep unchanged during the FL process. Unlike these systems, the clients in our system can update their local datasets by spending some costs and use fresh data to train local models. More fresher the local datasets provided by clients, more accurate global model will be obtained by the FL system. Besides, the time is devided into T equivalent-length timeslots, in each of which the server will conduct a round of federated learning under a budget. For simplicity, we assume the server has the same budget in each round, denoted by B, which can be extended to the case with heterogeneous budgets. Specifically, the whole FL system works as follows.

First, the server selects a subset of clients $\mathcal{N}_t$ $(\subseteq \mathcal{N})$ to update their local datasets at the beginning of each timeslot $t \in \mathcal{T} = \{1, \cdots, T\}$. For each client $i \in \mathcal{N}_t$, we denote its local dataset as $\mathcal{D}_t^i$, which can be regarded as the data collected from some fixed Point of Interests (PoIs) or purchased from some preferred data owners by client i. The dataset might be updated on-demand by the client, so that it might vary over different timeslots. For simplicity, we assume that the datasets across different timeslots remain the same size (otherwise, we can randomly sample the same number of data items from different size of datasets), i.e., $|\mathcal{D}_t^i| = |\mathcal{D}_{t'}^i| = n_i$, for any two time slots $t, t' \in \mathcal{T}$, where n_i denotes the size of $\mathcal{D}_t^i$ and $\mathcal{D}_{t'}^i$. Since each client might spend some costs in obtaining its local dataset, the server will pay a reward, denoted by p_i, to client i as the compensation. Meanwhile, the server publicizes global model parameters, denoted by ω_{t-1}, to all clients for their local trainings. Here, ω_{t-1} is the result of the $(t-1)$-th round of federated learning. Specially, ω_0 represents the initial global model parameter.

Second, each client i $(\in \mathcal{N})$ performs local training after receiving the global model parameter ω_{t-1} from the server.

Denote the loss function of local training as

$$F_{t,i}(\omega;\mathcal{D}_t^i)=\frac{1}{|\mathcal{D}_t^i|}\sum_{x\in\mathcal{D}_t^i}f(\omega;x), \tag{1}$$

where ω is the model parameter, $\mathcal{D}_t^i$ is the local training dataset, and $f(\cdot)$ is a server-specified loss function, e.g., mean absolute loss, mean squared loss, or cross entropy loss. Then, based on the received global model parameter ω_{t-1}, client i performs τ steps of mini-batch stochastic gradient descent to compute its local model parameter ω_t^i as follows

$$\omega_t^{i,k+1}=\omega_t^{i,k}-\eta_t\nabla F_{t,i}(\omega_t^{i,k};\xi_t^{i,k}), \tag{2}$$

where $k=0,\cdots,\tau-1$, $\xi_t^{i,k}$ is the mini-batch sampled from $\mathcal{D}_t^i$, η_t is the learning rate in t-th round, $\omega_t^{i,\tau}=\omega_t^i$ and $\omega_t^{i,0}=\omega_{t-1}$. Finally, client i uploads ω_t^i to the server and receives a payment from the server as the compensation of local training, denoted by q_i.

Third, the server aggregates the received local model parameters to obtain the global model parameter ω_t by

$$\omega_t=\sum_{i=1}^{N}\frac{n_i}{n}\omega_t^i, \tag{3}$$

where $n=\sum_{i=1}^{N}n_i$ is the total number of training data in each timeslot. Then, the server sends the updated global model ω_t back to each client for the next round of local training.

Overall, the global loss function is defined as follows

$$F(\omega)\triangleq\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{n_i}{n}F_{t,i}(\omega;\mathcal{D}_t^i). \tag{4}$$

The goal of the whole FL system is to obtain the optimal model parameter vector ω^* so as to minimize $F(\omega)$, i.e.,

$$\omega^*=\arg\min_{\omega}F(\omega). \tag{5}$$

B. Problem Formulation

In this paper, we focus on the data freshness in FL systems. Inspired by sensing systems, we utilize the concept of AoI to indicate the freshness of local dataset, which is defined by the elapsed time since the local data was generated. Specifically, let the current round of federated learning be in the t timeslot and $u_i(t)$ be the latest update time slot of client i's local dataset $\mathcal{D}_t^i$. Then, the AoI value of $\mathcal{D}_t^i$ (hereafter called client i's AoI, for simplicity of presentation) is represented as

$$\Delta_i(t)=t-u_i(t). \tag{6}$$

Especially, $\Delta_i(0)=0$ for all clients. Further, the dynamics of client i's AoI can be described as follows

$$\Delta_i(t)=\begin{cases}\Delta_i(t-1)+1\ , & i\notin\mathcal{N}_t,\\ 0\ , & otherwise.\end{cases} \tag{7}$$

It is worth noting that different client selection strategies will lead to various local data freshness even for the same client. Furthermore, different client selection strategies will also influence the loss of global model since local data freshness affects the quality of local training. Our goal is to minimize the loss of global model after the whole FL process by selecting the optimal $\mathcal{N}_t$ for each time slot $t\in\mathcal{T}$ under the limited budget B. The client selection strategies considered in this paper are non-anticipative, i.e. strategies that do not use future knowledge in selecting clients. Let Π be the class of non-anticipative strategies and $\pi\in\Pi$ be an arbitrary admissible strategy. Specifically, we use $A^\pi(t)=[a_1^\pi(t),\cdots,a_N^\pi(t)]$ ($t\in\mathcal{T}$) to indicate whether each client is selected in the i-th timeslot, i.e., $a_i(t)=1$ means $i\in\mathcal{N}_t$; otherwise, $a_i(t)=0$. Then, we can formalize the problem as follows:

$$\mathbf{P1}: \quad \min_{\pi\in\Pi}\ \mathbb{E}[F(\omega_T)]-F^*, \tag{8}$$

$$\text{s.t.}\quad a_i^\pi(t)\in\{0,1\},\forall i\in\mathcal{N},\forall t\in\mathcal{T}, \tag{8a}$$

$$\Delta_i(t)=\mathbb{1}_{\{a_i^\pi(t)=0\}}\left[\Delta_i(t-1)+1\right], \tag{8b}$$

$$\sum_{i=1}^{N}q_i+a_i^\pi(t)p_i\le B,\ \forall t\in\mathcal{T}. \tag{8c}$$

Here, ω_T in Eq. (8) is the aggregated global model after T rounds and $\mathbb{E}[F(\omega_T)]-F^*$ is the gap between the expected global loss after T rounds and the optimal global loss. Naturally, the closer $\mathbb{E}[F(\omega_T)]-F^*$ is to zero, the better is the performance of ω_T; Eq. (8a) represents that each client can only be selected at most once by the server for updating its local dataset in each timeslot; Eq. (8b) is the reformulation of Eq. (7), i.e., the dynamics of each client's AoI, where $\mathbb{1}_{\{\cdot\}}$ is an indicator function; and Eq. (8c) indicates that the budget constraint in each round of FL. For ease of reference, we list major notations in Table I.

TABLE I
DESCRIPTION OF MAJOR NOTATIONS

Variable	Description
i, t	the index of client and time slot, respectively.
$\mathcal{N}, \mathcal{T}$	the set of clients and time slots, respectively.
$\mathcal{N}_t$	the set of selected clients to in time slot t.
$\mathcal{D}_t^i, \|\mathcal{D}_t^i\|$	the dataset of client i in time slot t and its size.
$F(\omega)$	the global loss function.
$F_{t,i}(\omega)$	the loss function of client i in time slot t.
ω^*	the optimal model parameter that minimizes $F(\omega)$.
ω_t	the global model parameter in time slot t.
ω_t^i	the local model parameter of client i in time slot t.
τ	the number of local iterations.
η_t	the learning rate in time slot t.
$\bar{\eta}, \widetilde{\eta}$	the min and max of learning rate, respectively.
p_i	the payment of client i for local training.
q_i	the payment of client i for obtaining fresh data.
B	the budget of server per time slot.
$\Delta_i(t)$	the AoI value of client i in time slot t.
Δ	the average AoI value of federated learning system.

III. Convergence Analysis

To identity how each client's AoI affects the global model, we conduct a rigorous convergence analysis of our AoI-aware FL system. We start with several assumptions on the local loss function $F_{t,i}(\omega)$.

Assumption 1 : For all $t \in \{1, 2, \cdots, T\}$, $i \in \{1, 2, \cdots, N\}$, $F_{t,i}$ is β-smooth, that is, for $\forall$ ω_1, ω_2, $F_{t,i}(\omega_2) - F_{t,i}(\omega_1) \leq \langle \nabla F_{t,i}(\omega_1), \omega_2 - \omega_1 \rangle + \frac{\beta}{2} \|\omega_2 - \omega_1\|^2$.

Assumption 2 : For all $t \in \{1, 2, \cdots, T\}$, $i \in \{1, 2, \cdots, N\}$, $F_{t,i}$ is μ-strongly convex, i.e., for $\forall$ ω_1, ω_2, $F_{t,i}(\omega_2) - F_{t,i}(\omega_1) \geq \langle \nabla F_{t,i}(\omega_1), \omega_2 - \omega_1 \rangle + \frac{\mu}{2} \|\omega_2 - \omega_1\|^2$.

Assumption 3 : For all $t \in \{1, 2, \cdots, T\}$, $i \in \{1, 2, \cdots, N\}$, the stochastic gradients of loss function is unbiased, i.e., $\mathbb{E}_\xi[\nabla F_{t,i}(\omega; \xi)] = \nabla F_{t,i}(\omega)$.

Assumption 4 : For all $t \in \{1, 2, \cdots, T\}$, $i \in \{1, 2, \cdots, N\}$, the expected squared norm of stochastic gradients is bounded, i.e. $\mathbb{E}_\xi \|\nabla F_{t,i}(\omega; \xi)\|^2 \leq G_i^2 + \Delta_i(t)\sigma_i^2$.

Assumptions 1-3 are widely-used assumed in many existing convex FL works [7], [20], [21], which ensure that the gradient of $F_{t,i}(\omega)$ does not change arbitrarily quickly or slowly with respect to ω and the stochastic gradients sampled from local datasets are unbiased. It is noteworthy that models with convex loss functions, such as Logistic Regression (LR [22]) and Support Vector Machines (SVM [23]), satisfy Assumption 2. The evaluation results in Section VI show that our algorithm can also work well for the models (e.g., CNN [24]) whose loss functions are non-convex.

Assumption 4, however, is a novel assumption we made for our AoI-aware FL systems. Different from the assumptions made in other FL systems, where those works have assumed that $\mathbb{E}_\xi \|\nabla F_{t,i}(\omega; \xi)\|^2$ is bounded by G_i^2, we take into account the impact of clients's AoI on model training. Specifically, we assume the upper bound of $\mathbb{E}_\xi \|\nabla F_{t,i}(\omega; \xi)\|^2$ is positively correlated with $\Delta_i(t)$, and the coefficient σ_i^2 represents the sensitivity of client i's local dataset to freshness. The potential insight is that a smaller AoI value means a fresher local dataset and better models can be trained, which is consistent with that a smaller gradient norm indicates a better model performance when the loss function is convex. In particular, if client i is selected by the server to update its local dataset in round t, i.e., $\Delta_i(t) = 0$, then Assumption 4 will degrade to $\mathbb{E}_\xi \|\nabla F_{t,i}(\omega; \xi)\|^2 \leq G_i^2$, which is the same as the assumptions in [7], [20], [21].

Theorem 1 (Convergence Upper Bound). *For ease of expression, we define* $\bar{\eta} = \min_t\{\eta_t\}$ *and* $\widetilde{\eta} = \max_t\{\eta_t\}$. *Suppose Assumptions 1 to 4 hold, and the step size* $\bar{\eta} < \frac{2}{\mu}$. *Then, the FL training loss after the initial global model* ω_0 *is updated by Eq.* (3) *for* T *rounds satisfies:*

$$\mathbb{E}[F(\omega_T)] - F^* \leq \frac{\beta}{2}(1 - \frac{\mu\bar{\eta}}{2})^T \|\omega_0 - \omega^*\|^2 + \frac{\beta}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} \alpha_i \left[G_i^2 + \Delta_i(t)\sigma_i^2\right]. \quad (9)$$

where $\alpha_i = \frac{\widetilde{\eta} n_i}{\mu n} + N\widetilde{\eta}\left(\tau^2 \widetilde{\eta} + \frac{2(\tau-1)^2}{\mu} \frac{n_i^2}{n^2}\right)$.

Theorem 1 clearly presents the relationship between various factors and global loss in our AoI-aware FL system.

IV. Problem Deduction And Algorithm Design

In this section, we propose the client selection algorithm WICS. First, we use the convergence upper bound to convert the optimization objective of Problem **P1** to minimize the average AoI value, followed by modeling the AoI minimization problem as a RMAB problem. Next, we relax the RMAB problem and obtain its optimal strategy by using Whittle's approach. Finally, we derive the closed-form expresson for Whittle's Index and present the detailed algorithm.

A. Using the Convergence Bound to Convert Problem

According to Theorem 1, we obtain the convergence bound of the global model after T rounds. Then, we convert the objective of Problem **P1** using this convergence bound as follows

$$\mathbb{E}[F(\omega_T)] - F^* \leq \frac{\beta}{2}(1 - \frac{\mu\bar{\eta}}{2})^T \|\omega_0 - \omega^*\|^2 + \frac{\beta}{2} \sum_{t=1}^{T} \sum_{i=1}^{N} \alpha_i G_i^2 + \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \phi_i \Delta_i(t) \quad (10)$$

where $\phi_i = \frac{\alpha_i \sigma_i^2 \beta N T}{2}$. Neglecting the constant term, the objective of Problem **P1** can be converted to

$$\min_{\pi \in \Pi} \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \phi_i \Delta_i(t), \quad (11)$$

In addition, the total reward paid by the server to the clients for their local training in each timeslot, i.e., $\sum_{i=1}^{N} q_i$, is fixed. Therefore, by defining $B' = B - \sum_{i=1}^{N} q_i$, we can simplify Eq. (8c) as follows

$$\sum_{i=1}^{N} a_i(t) p_i \leq B'. \quad (12)$$

As a result, we can convert the original Problem **P1** to the following AoI minimization problem:

$$\mathbf{P2}: \quad \min_{\pi \in \Pi} \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \phi_i \Delta_i(t), \quad (13)$$

$$\text{s.t.} \quad a_i^\pi(t) \in \{0, 1\}, \forall i \in \mathcal{N}, \forall t \in \mathcal{T}, \quad (13b)$$

$$\Delta_i(t) = \mathbb{1}_{\{a_i^\pi(t)=0\}} \left[\Delta_i(t-1) + 1\right], \quad (13c)$$

$$\sum_{i=1}^{N} a_i^\pi(t) p_i \leq B', \ \forall t \in \mathcal{T}. \quad (13d)$$

B. RMAB Modeling and Solution

To solve Problem **P2**, we cast it as a Restless Multi-Armed Bandit (RMAB) problem [25] by means of the stochastic control theory. Different from classic MAB problems, where the unused arms neither yield rewards nor change states and the states of all arms are known at any time, the arms in RMAB might continue to change states according to different transition rules even if they are not being pulled. In this paper, we regard each client as a restless bandit and the AoI value as its state since the AoI value changes in every timeslot even if the cilent is not selected. However, the RMAB problem is usually PSPACE-hard [25]. To this end, we adopt the Whittle's methodology to solve this problem [26].

First, we relax Problem $\mathbf{P2}$ by replacing the constraint Eq. (14d) with a relaxed version as follows

$$\mathbf{P3}: \quad \min_{\pi \in \Pi} \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \phi_i \Delta_i(t), \tag{14}$$

$$\text{s.t.} \quad a_i^{\pi}(t) \in \{0,1\}, \forall i \in \mathcal{N}, \forall t \in \mathcal{T}, \tag{14b}$$

$$\Delta_i(t) = \mathbb{1}_{\{a_i^{\pi}(t)=0\}} \left[\Delta_i(t-1)+1\right], \tag{14c}$$

$$\frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} a_i^{\pi}(t) p_i \leq \frac{B'}{N}, \ \forall t \in \mathcal{T}. \tag{14d}$$

Second, after the relaxation, we apply the Lagrangian approach to decouple the RMAB problem (P3) into multiple sub-problems. The Lagrange dual function is given by

$$\mathcal{L}(\pi, \lambda) = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \left[\phi_i \Delta_i(t) + \lambda a_i^{\pi}(t) p_i\right] - \frac{\lambda B'}{N}. \tag{15}$$

Then, our goal turns to find a client selection strategy that minimizes $\mathcal{L}(\pi, \lambda)$. Notice that $\mathcal{L}(\pi, \lambda)$ is separable and thus can be solved for each individual client. The problem associated with each client is called the decoupled model, whose goal is to determine whether or not the client should be selected for updating its local dataset in each round. Specifically, we formalize the decoupled model over an infinite-horizon as the following problem

$$\mathbf{P4}: \quad \min_{\pi \in \Pi} \left\{ \lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} \left[\frac{\phi_i}{p_i} \Delta_i(t) + \lambda a_i^{\pi}(t) \right] \right\} \tag{16}$$

$$\text{s.t.} \quad a_i^{\pi}(t) \in \{0,1\}, \forall i \in \mathcal{N}, \forall t \in \mathcal{T}, \tag{16b}$$

$$\Delta_i(t) = \mathbb{1}_{\{a_i^{\pi}(t)=0\}} \left[\Delta_i(t-1)+1\right], \tag{16c}$$

$$\lambda \geq 0. \tag{16d}$$

Here, Eq. (16) is actually the decoupled version of Eq. (14).

Third, to address Problem $\mathbf{P4}$, we formulate the Decoupled Model as a Markov Decision Process (MDP), which consists of the AoI state $\Delta_i(t)$, the control variable $a_i^{\pi}(t)$, the state transition functions $\mathbb{P}(\cdot)$, and the cost function $\mathbb{C}_i(\cdot)$. Specifically, the state transition from time slot t to time slot $t+1$ in MDP is deterministic as follows

$$\begin{aligned} &\mathbb{P}(\Delta_i(t+1) = \Delta_i(t)+1 | a_i^{\pi}(t)=0) = 1; \\ &\mathbb{P}(\Delta_i(t+1) = 0 | a_i^{\pi}(t)=0) = 0; \\ &\mathbb{P}(\Delta_i(t+1) = \Delta_i(t)+1 | a_i^{\pi}(t)=1) = 0; \\ &\mathbb{P}(\Delta_i(t+1) = 0 | a_i^{\pi}(t)=1) = 1; \end{aligned} \tag{17}$$

Moreover, we can see the objective of Problem $\mathbf{P4}$ as the cost function of MDP. The cost function on the transition from timeslot t to timeslot $t+1$ is defined as

$$C_i(\Delta_i(t), a_i^{\pi}(t)) \triangleq \frac{\phi_i}{p_i} \Delta_i(t) + \lambda a_i^{\pi}(t), \tag{18}$$

where the first part is associated with the resulting AoI value in timeslot t and λ can be regarded as the service charge, which is generated only when $a_i^{\pi}(t) = 1$.

Finally, we derive the optimal strategy of this MDP and prove that it is a special type of deterministic strategy.

Algorithm 1: Whittle's Index based Client Selection

Input: AoI value of each client $\{\Delta_1(t), \cdots, \Delta_N(t)\}$, weight of each client $\{\phi_1, \cdots, \phi_N\}$, payment of each client $\{p_1, \cdots, p_N\}$, budget B'

Output: The index set of selected clients $\mathcal{N}_{t+1}$

for each client i in $\mathcal{N}$ **do**
  Calculates its WI value $WI_{i,t}$ according to Eq.(20) and sends it to the server
end for
The server sorts the clients into $(i_1, i_2, \cdots, i_N)$ such that $WI_{i_1,t} \geq WI_{i_2,t} \geq \cdots \geq WI_{i_N,t}$, and initializes an empty set $\mathcal{N}_{t+1}$, an initial index $k=1$
while $\sum_{i \in \mathcal{N}_{t+1}} p_i + p_{i_k} < B'$ **do**
  $\mathcal{N}_{t+1} \leftarrow \mathcal{N}_{t+1} \cup \{i_k\}$, $k = k+1$
end while
return the index set $\mathcal{N}_{t+1}$ of the selected clients

Theorem 2 (Optimal Strategy for Problem $\mathbf{P4}$)**.** *Consider the decoupled model over an infinite time-horizon. The optimal strategy π^* is selecting client i in each timeslot t to update its local dataset only when $\Delta_i(t) > H-1$, where*

$$H = \left\lfloor -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{2\lambda p_i}{\phi_i}} \right\rfloor. \tag{19}$$

Note that the threshold H is a function of λ. Intuitively, we expect that the server selects client i when $\Delta_i(t)$ is high to reduce the AoI value and does not select client i when $\Delta_i(t)$ is low, so as to avoid the service charge λ.

C. The WICS Algorithm

Based on the optimal strategy given in Theorem 2 for Problem $\mathbf{P4}$, we design the WICS algorithm for Problem $\mathbf{P2}$ by applying the Whittle's Index methodology as follows.

First, we show that Problem $\mathbf{P2}$ is indexable and thus applicable to define Whittle's Index for it. Let $\mathcal{P}(\lambda) = \{\Delta_i \in \mathbb{N} | \Delta_i < H\}$ be the set of AoI states in which client i will not be selected by the optimal strategy given in Theorem 2. Then, the indexability analysis of Problem $\mathbf{P2}$ is as follows.

Definition 1 (Indexability [27])**.** *The decoupled model is indexable if the set $\mathcal{P}(\lambda)$ increases monotonically from $\emptyset$ to $\mathbb{N}$ when λ increases from 0 to ∞. Moreover, Problem $\mathbf{P2}$ is indexable if the decoupled model is indexable for each client.*

Theorem 3 (Indexability of Problem $\mathbf{P2}$)**.** *Problem $\mathbf{P2}$ is indexable.*

Proof. Obviously, the threshold H in Eq. (19) is monotonically increasing with λ. Specifically, substituting $\lambda = 0$ yields $H = 0$, which implies $\mathcal{P}(\lambda) = \emptyset$, and $\lambda \to \infty$ gives $H \to \infty$, which suggests $\mathcal{P}(\lambda) = \mathbb{N}$. Therefore, the decoupled model for each client is indexable, which also implies that Problem $\mathbf{P2}$ is indexable according to Definition 1. □

Next, we define Whittle's Index (WI) for Problem $\mathbf{P2}$ as the infimum service charge λ that makes both of the decisions on selecting client i or not equally desirable. Specifically, we use $\lambda_i(\Delta_i(t))$ to denote the Whittle's Index of client i in state $\Delta_i(t)$, where λ is extended as a function on $\Delta_i(t)$. According to Theorem 2, to make both selection decisions equally desirable in state $\Delta_i(t)$, the threshold should satisfy $H = \Delta_i(t) + 1$. Substituting Eq. (19), we can solve this equation to derive a closed-form expression for $\lambda_i(\Delta_i(t))$, i.e., the Whittle's Index of client i, as follows:

$$WI_{i,t} \triangleq \lambda_i(\Delta_i(t)) = \frac{(\Delta_i(t)+1)(\Delta_i(t)+2)\phi_i}{2p_i}. \quad (20)$$

Now, based on the Whittle's Index, we can design the WICS algorithm for Problem $\mathbf{P2}$. The basic idea is to select the clients with higher WI values in each timeslot, while ensuring that the budget is not exceeded. As shown in Algorithm 1, we first calculate the WI value for each client according to Eq.(20) and then sort all clients in $\mathcal{N}$ into the set $(i_1, i_2, \cdots, i_N)$ such that $WI_{i_1,t} \geq WI_{i_2,t} \geq \cdots \geq WI_{i_N,t}$ (Steps 1-3). Next, we greedily select the clients into a winning set $\mathcal{N}_t$ and give the corresponding payments for the winning clients until the remaining budget cannot afford the next client (Steps 4-6).

Finally, we analyze the performance of the WICS algorithm. Obviously, the computational complexity of WICS is dominated by the sorting operation on clients' WI values, i.e., $O(N \log N)$. In addition, we define the ratio $\rho^\pi \triangleq \frac{U_B^\pi}{L_B}$ to indicate the performance of strategy π, where L_B is a lower bound to the optimal performance of Problem $\mathbf{P2}$ and U_B^π is an upper bound to the performance of Problem $\mathbf{P2}$ under strategy π, and say that strategy π is ρ^π-optimal. Then, the WICS algorithm satisfies the following theorem.

Theorem 4 (Approximate Optimality)**.** *The solution produced by the WICS algorithm to the Problem* $\mathbf{P2}$ *over an infinite time-horizon is* ρ^{WI}*-optimal, where*

$$\rho^{WI} = \frac{2\left[\frac{1}{N}\sum_{i=1}^{N}(\sqrt{2\phi_i}+\sqrt{\frac{\phi_i}{2}})\right]^2 - \frac{1}{N}\sum_{i=1}^{N}\phi_i}{\frac{1}{2NM}\left[\sum_{i=1}^{N}\sqrt{\phi_i}\right]^2 - \frac{1}{2N}\sum_{i=1}^{N}\phi_i}, \quad (21)$$

$M = \left\lfloor \frac{B'}{p_{min}} \right\rfloor$, *and* $p_{min} = \min_i\{p_i\}$.

Note that the objective of Problem $\mathbf{P2}$ (i.e., Eq.(13)) is derived from the objective of Problem $\mathbf{P1}$ (i.e., Eq.(8)) according to the convergence bound analysis. Thus, the WICS algorithm is at least ρ^{WI}-optimal for Problem $\mathbf{P1}$.

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of with extensive simulations on real world datasets.

A. Evaluation Methodology

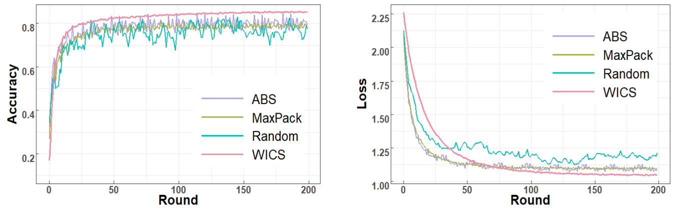
1) Simulation Setup: We conduct extensive simulations on two widely used real datasets MNIST [28] and Fashion-MNIST (FMNIST [29]). The MNIST dataset contains 60,000 handwritten digits for training and 10,000 for the test, while the Fashion-MNIST dataset contains 60,000 fashion clothes for training and 10,000 for the test. We adopt both the convex model (i.e., LR) and the non-convex model (i.e., CNN). The CNN consists of two 5×5 convolution layers (32, 64 channels), each of which is followed by 2×2 max pooling, two fully-connected layers with 3136 and 512 units, and a ReLU layer with 10 units. We first let the number of clients $N = 10$ and the number of time slots $T = 200$. Next, we generate the simplified budget in each time slot (i.e., B') from $\{25, 40, 55, 70\}$. Then, we determine the cost p_i. We assume that the cost p_i is proportional to the number of local data and let the cost for each client not exceed $[5, 15]$. For all experiments, we initialize our model with $\omega_0 = \mathbf{0}$ and use an SGD batch size of $b = 16$. Without loss of generality, we set the learning rate of LRas $\eta_t = 0.005$ and the learning rate of CNN as $\eta_t = 0.01$ for all time slots and each client performing $\tau = 10$ local iterations. After that, we can appropriately set the weight $\phi_i \in (0, 1)$ according to Eq. (10), which is similar to the method in [41]. In order to reflect the impact of AoI on the local data, we mislabel some local data of each client according to its AoI value in each time slot. Specifically, we will mislabel more data if the client has a larger AoI value.

2) Algorithms for Comparison: Since WICS consider the freshness of local datasets in FL, there are no existing algorithms that can be directly applied to our problem. To the best of our knowledge, the closest algorithm that can be adapted to our setting is the ABS algorithm proposed by [30], which is also an index based strategy. However, since the ABS algorithm considers the age-of-update (AoU) rather than AoI, we need to modify it to deal with the AoI in our model. More specifically, the modified ABS index of client i in timeslot t is given by $\frac{\Delta_i(t)\phi_i}{p_i}$. Similar to our WICS strategy, the ABS algorithm greedily selects clients with larger modified ABS index values while ensuring that the budget is not exceeded in each timeslot. Moreover, we implement the MaxPack algorithm [31] and the Random algorithm for better comparison. The MaxPack algorithm is the comparison algorithm of ABS algorithm in [30].

B. Evaluation Results

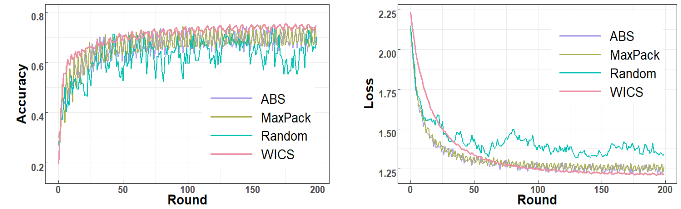
In this section, we train different models (i.e., LR, CNN) on both MNIST and FMNIST to compare the performance of different algorithms. Notably, we conduct experiments with variant budget B', which show a similar performance. Due to the limited space, we only illustrate the result of $B' = 40$ in the paper.

First, we exhibit the performance of different algorithms for LR on MNIST and FMNIST in terms of both accuracy and loss, as shown in Fig.2 and Fig.3, respectively. In Fig.2, we can observe that the achieved accuracy of all four algorithms rises along with the increase of rounds, while the achieved loss of all four algorithms descends with the increase of rounds. Moreover, the performance of WICS in terms of both accuracy and loss is better than the three compared algorithms. In Fig.3, we conduct the same experiments of LR on FMNIST and obtain the similar results. We see that WICS can also achieve the best results in all algorithms. This means that WICS is effective for models with convex loss function, which



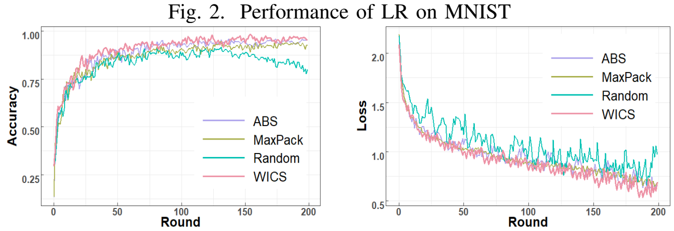
(a) Accuracy of LR on MNIST (b) Loss of LR on MNIST

Fig. 2. Performance of LR on MNIST



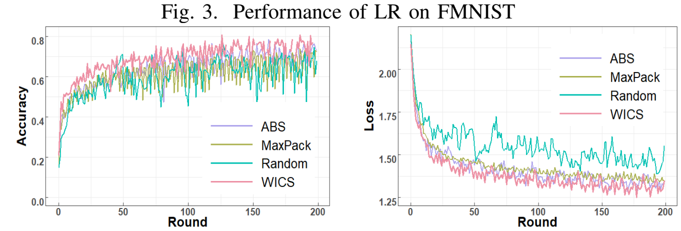
(a) Accuracy of LR on FMNIST (b) Loss of LR on FMNIST

Fig. 3. Performance of LR on FMNIST



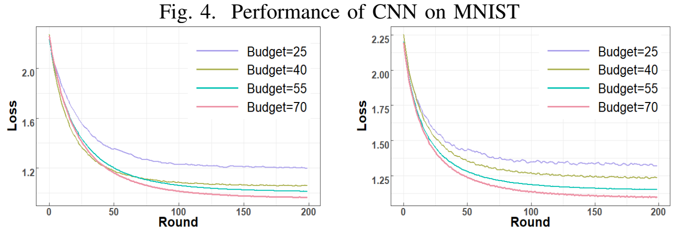
(a) Accuracy of CNN on MNIST (b) Loss of CNN on MNIST

Fig. 4. Performance of CNN on MNIST



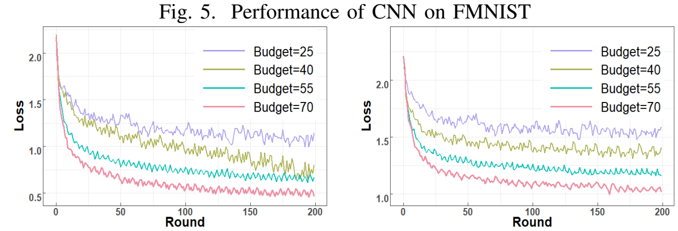
(a) Accuracy of CNN on FMNIST (b) Loss of CNN on FMNIST

Fig. 5. Performance of CNN on FMNIST



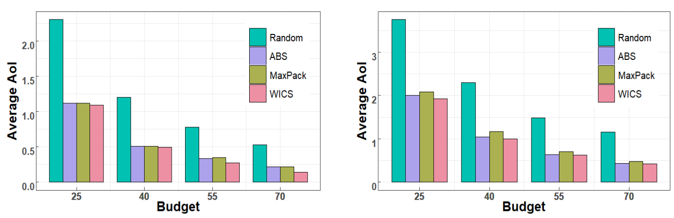
(a) N=10 (b) N=20

Fig. 6. Loss vs. the number of clients N



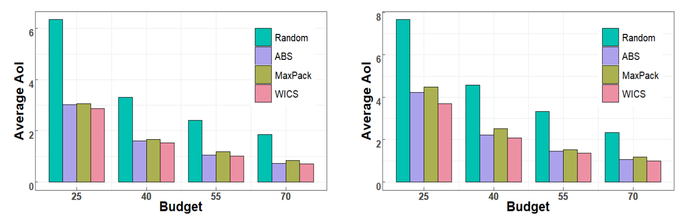
(a) N=30 (b) N=40

Fig. 7. Loss vs. the number of clients N



(a) N=10 (b) N=20

Fig. 8. Average AoI vs. the number of clients N



(a) N=30 (b) N=40

Fig. 9. Average AoI vs. the number of clients N

matches with the theoretical convergence bound. To verify the effectiveness of WICS when the loss function is non-convex, we further train CNN on MNIST and FMNIST. Figs.4-5 show that the performances of WICS are still better than other algorithms when the loss function does not satisfy the convex assumption.

Furthermore, we analyze the influence of different budget B' on all four algorithms in terms of loss, where we set the number of clients $N = 10$. We take WICS as an example, and display the results in Figs.6-7. The figures indicate that whether the loss function of the model is convex or not, the larger B' is, the smaller loss of model can be achieved. This can be explained by the reason that a larger B' allows more clients to update their local datasets in each time slot, so as to make the local datasets more fresh and get better learning performance, which also matches with the convergence upper bound analysis.

Finally, we evaluate the performance of all four algorithms in terms of average AoI, which is computed by $\Delta = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{n_i}{n}\Delta_i(t)$. The evaluation results are depicted in Figs.8-9, where we scale B' from 25 to 70 with an increment of 15 and evaluate the effects of N. The figures show that WICS can achieve the lowest weighted average AoI in all four algorithms. More specifically, ABS, MaxPack and WICS are far better than Random algorithm and the performance of ABS is the closest algorithm compared to WICS. In addition, the weighted average AoI exhibits an uptrend with the increasing of N. This is because when we keep the budget fixed, the number of clients who are not selected by the server in each timeslot will increase with N, i.e., the increment of AoI values in each time slot will become larger. Hence, the weighted average AoI is also increasing with the increment of N.

VI. Related Work

We review the related work from the following two aspects:

Client Selection for FL: Client selection has been widely investigated in the literatures of FL [1], [32]–[34], considering various facets of the system, such as statistical heterogeneity and system heterogeneity. Different optimization objectives like importance sampling and resource-aware optimization-based approaches have also been considered. For example, existing works use clients local gradient [32], [33] or local loss [34] information to measure the importance of clients local data, and then select the clients according to the data importance. In addition, resource-aware optimization-based approaches, such as CPU frequency allocation [35], communication bandwidth allocation [36] and straggler-aware client scheduling [37], select the clients to optimize the different aspects of the federated learning system.

Age of Information: AoI is a novel application-layer metric for measuring freshness that was initially conceived by [19]. Since its inception, there has been active research on AoI (see an online bibliography in [38]), which includes a wide range of problems. An important class of problems that has attracted much attention is how to design schedulers to minimize AoI [39], [40]. For instance, [39] studies how to minimize the average AoI of the deployed sensor nodes in data collection by mobile crowdsensing. [40] studies the problem of minimizing AoI in general single-hop and multihop wireless networks.

VII. CONCLUSION

In this paper, we introduce a novel AoI-aware FL systems, where clients might use fresh datasets to perform local model training and the server try to select some clients to provide fresh datasets in each timeslot but constrained by a limited budget. We use AoI to indicate the freshness of datasets and theoretically analyze the convergence upper bound of the AoI-aware FL system. On this basis, we model the corresponding client selection issue as a restless multi-armed bandit problem, and propose a Whittle's-Index-based client selection algorithm, i.e., WICS, to solve this problem. Moreover, we prove that the WICS can achieve nearly optimal performance on client selection. Finally, we also conduct extensive simulations on two real datasets and the simulation results demonstrate the effectiveness of our algorithm.

VIII. ACKNOWLEDGMENTS

This paper was supported in part by the National Natural Science Foundation of China (NSFC) (Grant No. 61572457, 61872330, 62172386), the Natural Science Foundation of Jiangsu Province in China (Grant No. BK20191194, BK20131174), and NSF Grants CNS 2128378, CNS 2107014, CNS 1824440, CNS 1828363, CNS 1757533, CNS 1629746, and CNS 1651947.

IX. APPENDIX

A. Proof of Theorem 1

Proof. First, we analyze how the difference between $\mathbb{E}[F(\omega_t)]$ and F^* (i.e., $F(\omega^*)$) changes in each round. Due to β-smooth and by using the fact that $\nabla F(\omega^*) = 0$, we have

$$\Omega_t = \Omega_{t-1} + \underbrace{\mathbb{E}\,\| \sum_{i=1}^{N} \frac{n_i}{n}(\omega_t^i - \omega_{t-1})\|^2}_{A_1} + \underbrace{2\mathbb{E}\left\langle \omega_{t-1} - \omega^*, \sum_{i=1}^{N} \frac{n_i}{n}(\omega_t^i - \omega_{t-1})\right\rangle}_{A_2}. \tag{22}$$

For A_1, we can bound it by using the AM-GM inequality and the Cauchy-Schwarz inequality:

$$\begin{aligned} A_1 &= \| - \sum_{i=1}^{N} \frac{n_i \eta_t}{n} \sum_{k=0}^{\tau-1} \nabla F_{t,i}(\omega_t^{i,k}; \xi_t^i)\|^2 \\ &\le N\tau \sum_{i=1}^{N} \frac{n_i^2 \eta_t^2}{n^2} \sum_{k=0}^{\tau-1} \|\nabla F_{t,i}(\omega_t^{i,k}; \xi_t^i)\|^2. \end{aligned} \tag{23}$$

Further, according to Assumption 4, it follows:

$$\mathbb{E}[A_1] \le N\tau^2\eta_t^2 \sum_{i=1}^{N} \frac{n_i^2}{n^2} \left[G_i^2 + \Delta_i(t)\sigma_i^2\right]. \tag{24}$$

For A_2, we have:

$$A_2 = \underbrace{\left\langle \omega_{t-1} - \omega^*, -\sum_{i=1}^{N} \frac{n_i \eta_t}{n} \nabla F_{t,i}(\omega_{t-1}; \xi_t^i)\right\rangle}_{B_1} + \underbrace{\left\langle \omega_{t-1} - \omega^*, -\sum_{i=1}^{N} \frac{n_i \eta_t}{n} \sum_{k=1}^{\tau-1} \nabla F_{t,i}(\omega_t^{i,k}; \xi_t^i)\right\rangle}_{B_2}. \tag{25}$$

Next, we bound $\mathbb{E}[B_1]$ and $\mathbb{E}[B_2]$, respectively. Using the μ-strongly convex of $F_{t,i}(\cdot)$ and the fact that $F_{t,i}^* \le F_{t,i}(\omega_{t-1})$, we can bound $\mathbb{E}[B_1]$ as follows:

$$\begin{aligned} &\mathbb{E}[B_1] = \sum_{i=1}^{N} \frac{n_i \eta_t}{n} - \langle \omega_{t-1} - \omega^*, \nabla F_{t,i}(\omega_{t-1})\rangle \\ &\le \sum_{i=1}^{N} \frac{n_i \eta_t}{n} \left(F_{t,i}(\omega^*) - F_{t,i}(\omega_{t-1}) - \frac{\mu}{2}\mathbb{E}\|\omega_{t-1} - \omega^*\|^2\right) \\ &\le \sum_{i=1}^{N} \frac{n_i \eta_t}{n} \left[F_{t,i}(\omega^*) - F_{t,i}^*\right] - \frac{\mu\eta_t}{2}\Omega_{t-1} \\ &\le \sum_{i=1}^{N} \frac{n_i \eta_t}{2n\mu} \|\nabla F_{t,i}(\omega^*)\|^2 - \frac{\mu\eta_t}{2}\Omega_{t-1} \\ &\le \sum_{i=1}^{N} \frac{n_i \eta_t}{2n\mu} \left[G_i^2 + \Delta_i(t)\sigma_i^2\right] - \frac{\mu\eta_t}{2}\Omega_{t-1} \end{aligned} \tag{26}$$

For $\mathbb{E}[B_2]$, we have:

$$\mathbb{E}[B_2] \leq \frac{\mu\eta_t}{4}\mathbb{E}\|\omega_{t-1} - \omega^*\|^2 + \frac{1}{\mu\eta_t}\mathbb{E}\|\sum_{i=1}^{N}\frac{n_i\eta_t}{n}\sum_{k=1}^{\tau-1}\nabla F_{t,i}(\omega_t^{i,k};\xi_t^i)\|^2$$
$$\leq \frac{\mu\eta_t}{4}\Omega_{t-1} + \frac{N\eta_t(\tau-1)^2}{\mu}\sum_{i=1}^{N}\frac{n_i^2}{n^2}\left[G_i^2 + \Delta_i(t)\sigma_i^2\right]. \quad (27)$$

Combining (22)-(26), we can obtain that:

$$\Omega_t \leq (1-\frac{\mu\eta_t}{2})\Omega_{t-1} + \frac{\eta_t}{\mu}\sum_{i=1}^{N}\frac{n_i}{n}\left[G_i^2+\Delta_i(t)\sigma_i^2\right] + N\eta_t\left(\tau^2\eta_t + \frac{2(\tau-1)^2}{\mu}\right)\sum_{i=1}^{N}\frac{n_i^2}{n^2}\left[G_i^2+\Delta_i(t)\sigma_i^2\right]$$
$$\leq (1-\frac{\mu\bar{\eta}}{2})\Omega_{t-1} + \sum_{i=1}^{N}\alpha_i\left[G_i^2+\Delta_i(t)\sigma_i^2\right]. \quad (28)$$

Then, by induction, we can prove:

$$\Omega_T \leq (1-\frac{\mu\bar{\eta}}{2})^T\Omega_0 + \sum_{t=1}^{T}\sum_{i=1}^{N}\alpha_i\left[G_i^2+\Delta_i(t)\sigma_i^2\right] \quad (29)$$

Therefore, the theorem holds. □

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 *AISTATS*, 2017.

[2] Life Insurance, "Swiss Re and Chinese firm WeBank partner to explore use of AI in reinsurance," https://www.lifeinsuranceinternational.com/news/swiss-re-webank/, accessed Jan. 11, 2021.

[3] Owkin, "Owkin solutions," https://owkin.com/solutions/#data-connect, accessed Jan. 11, 2021.

[4] MELLODDY, "Machine learning ledger orchestration for drug discovery," https://www.melloddy.eu, accessed Jan. 11, 2021.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *ICML*, 2020.

[6] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *MLSys*, 2020.

[7] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in *ICLR*, 2020.

[8] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," *arXiv preprint arXiv:1806.00582*, 2018.

[9] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *IEEE INFOCOM*, 2020.

[10] C. Chen et al., "Communication-Efficient Federated Learning with Adaptive Parameter Freezing," in *ICDCS*, 2021.

[11] X. Zhang, F. Li, Z. Zhang, Q. Li, C. Wang, and J. Wu, "Enabling execution assurance of federated learning at untrusted participants," in *IEEE INFOCOM*, 2020.

[12] X. Yuan, X. Ma, L. Zhang, Y. Fang and D. Wu, "Beyond Class-Level Privacy Leakage: Breaking Record-Level Privacy in Federated Learning," in *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2555-2565, 2022.

[13] Tuli, Shreshth et al. "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data." *VLDB Endow*. 2022.

[14] N. H. Tran, W. Bao, A. Zomaya, N. M. NH, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *IEEE INFOCOM*, 2019.

[15] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546-3557, 2020.

[16] Y. Zhan, P. Li, Z. Qu, D. Zeng and S. Guo, "A Learning-Based Incentive Mechanism for Federated Learning," in *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6360-6368, 2020.

[17] R. Zeng, S. Zhang, J. Wang and X. Chu, "FMore: An Incentive Scheme of Multi-dimensional Auction for Federated Learning in MEC," in *ICDCS*, 2020.

[18] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys & Tuts.*, vol. 22, no. 3, pp. 2031-2063, 2020.

[19] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *IEEE SECON*, 2011.

[20] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD for non-convex optimization with faster convergence and less communication," in *AAAI*, 2019.

[21] Z. Qu, K. Lin, J. Kalagnanam, Z. Li, J. Zhou, and Z. Zhou, "Federated learning's blessing: Fedavg has linear speedup," *arXiv preprint arXiv:2007.05690*, 2020.

[22] D. W. Hosmer, Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, vol. 398, 2013.

[23] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, 1998.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[25] J. Gittins, K. Glazebrook, and R. Weber, *Multi-Armed Bandit Allocation Indices*. John Wiley & Sons, 2011.

[26] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *Journal of applied probability*, pp. 287-298, 1988.

[27] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh and E. Modiano, "Scheduling Policies for Minimizing Age of Information in Broadcast Wireless Networks," in IEEE/ACM Transactions on Networking, vol. 26, no. 6, pp. 2637-2650, 2018.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[29] Xiao, Han, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747* (2017).

[30] H. H. Yang, A. Arafa, T. Q. S. Quek and H. Vincent Poor, "Age-Based Scheduling Policy for Federated Learning in Mobile Edge Networks," in *ICASSP*, 2020.

[31] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, 2019.

[32] Cho, Yae Jee et al. "Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies." *ArXiv* abs/2010.01243 2020.

[33] Ribero, Mónica and Haris Vikalo. "Communication-Efficient Federated Learning via Optimal Client Sampling." *ArXiv* abs/2007.15197 2020.

[34] Lai, Fan et al. "Oort: Efficient Federated Learning via Guided Participant Selection." in *USENIX OSDI*, 2020.

[35] Y. Xu, M. Xiao, J. Wu, S. Zhang and G. Gao, "Incentive Mechanism for Spatial Crowdsourcing With Unknown Social-Aware Workers: A Three-Stage Stackelberg Game Approach," in *IEEE Transactions on Mobile Computing*, vol. 22, no. 8, pp. 4698-4713, 1 Aug. 2023, doi: 10.1109/TMC.2022.3157687.

[36] Y. Xu, M. Xiao, H. Tan, A. Liu, G. Gao and Z. Yan, "Incentive Mechanism for Differentially Private Federated Learning in Industrial Internet of Things," in *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 6927-6939, Oct. 2022, doi: 10.1109/TII.2021.3134257.

[37] Y. Xu, M. Xiao, J. Wu, H. Tan and G. Gao, "A Personalized Privacy Preserving Mechanism for Crowdsourced Federated Learning," in *IEEE Transactions on Mobile Computing*, doi: 10.1109/TMC.2023.3237636.

[38] Y. Sun, "A Collection of Recent Papers on the Age of Information," available at http://www.auburn.edu/%7eyzs0078

[39] Z. Dai, H. Wang, C. H. Liu, R. Han, J. Tang and G. Wang, "Mobile Crowdsensing for Data Freshness: A Deep Reinforcement Learning Approach," *IEEE INFOCOM*, 2021.

[40] V. Tripathi and E. Modiano, "Age Debt: A General Framework for Minimizing Age of Information," in *IEEE INFOCOM WKSHPS*, 2021.

[41] S. Wang et al., "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *IEEE INFOCOM)*, 2018.