Optimal Contract Design for Efficient Federated Learning With Multi-Dimensional Private Information

Ningning Ding, Zhixuan Fang, and Jianwei Huang, *Fellow, IEEE*

***Abstract*— As an emerging machine learning technique, federated learning has received significant attention recently due to its promising performance in mitigating privacy risks and costs. While most of the existing work of federated learning focused on designing learning algorithm to improve training performance, the incentive issue for encouraging users' participation is still under-explored. This paper presents an analytical study on the server's optimal incentive mechanism design, in the presence of users' multi-dimensional private information (e.g., training cost and communication delay). Specifically, we consider a multi-dimensional contract-theoretic approach, with a key contribution of summarizing users' multi-dimensional private information into a one-dimensional criterion that allows a complete order of users. We further perform the analysis in three information scenarios to reveal the impact of information asymmetry levels on server's optimal strategy and minimum cost. We show that weakly incomplete information does not increase the server's cost (comparing with the complete information scenario) when training data is IID, but it in general does when data is non-IID. Furthermore, the optimal mechanism design under strongly incomplete information is much more challenging, and it is not always optimal for the server to incentivize the group of users with the lowest training cost and delay to participate.**

***Index Terms*— Federated learning, incentive mechanism, multi-dimensional contract, information asymmetry.**

I. Introduction

A. Background and Motivations

THE unprecedented amount of data generated by users' mobile devices has a great potential in powering intelligent learning models in many aspects of our life. However, the privacy concerns from users often make it risky (or even illegal) to store all the users' data in a centralized location. This motivates the emergence of federated learning, which can enable effective learning while protecting users' privacy.

A typical federated learning application platform (e.g., Google Keyboard, or Gboard in short) usually consists of (i) a population of users who use their local data to collaboratively train a shared learning model and (ii) a central server who coordinates the training. Specifically, in each round of a synchronous training process, each user computes the parameters of the global learning model based on his local data and sends the parameters to the server; the server updates the global model based on the users' inputs and feeds the aggregated global model back to users for their computation in the next round. Users and the server repeat this process until a desired accuracy is achieved, e.g., the training error is smaller than a target threshold [2]. Different from a traditional centralized model training process where the central server acquires and stores users' raw data, federated learning allows users to keep the local data on their own devices and only share the intermediary model parameters, which well protects users' data privacy.

However, with all the promising benefits, federated learning also comes with challenges to tackle. First, most existing studies usually make an optimistic assumption that users are willing to participate in the training process (e.g., [3]). That may not be realistic without proper incentives, as users incur various costs during the training process [4]. Second, the server can selectively incentivize appropriate users' participation to enhance training efficiency as well as effectiveness (e.g., in timely traffic flow prediction [5]). However, this may not be easy to achieve if the server does not know users' multi-dimensional information such as communication delay and training costs. The communication delay depends on each user's device configuration and time availability, which are often unknown to the server, especially when there are a large number of heterogeneous users in federated learning. The training costs are also users' private information and will not be easily accessed by the server due to users' privacy concerns. Nevertheless, the related literature usually only considers one-dimensional private information (e.g., [6] only considers users' training cost.). Therefore, in the presence of users' multi-dimensional private information, it is necessary for the server to design an incentive mechanism to simulate users' participation, encourage honest behaviors, and enhance training efficiency.

Although the server may not know each user's private information, he may have the knowledge about statistics of

Manuscript received July 27, 2020; revised September 27, 2020; accepted October 21, 2020. Date of publication November 9, 2020; date of current version December 16, 2020. This work was supported in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society and in part by the Presidential Fund from the Chinese University of Hong Kong, Shenzhen. This article was presented in part at the International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks *WiOpt'20*. *(Corresponding author: Jianwei Huang.)*

Ningning Ding is with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong.

Zhixuan Fang is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China, and also with the Shanghai Qi Zhi Institute, Shanghai 200232, China.

Jianwei Huang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China, and also with the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), Shenzhen 518172, China (e-mail: jianweihuang@cuhk.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2020.3036944.

Digital Object Identifier 10.1109/JSAC.2020.3036944

such information through market research and past experiences [10]. For example, the server may know the numbers of different types of users (which is denoted as *weakly incomplete information* in this paper) or the user type distribution (which is denoted as *strongly incomplete information* in this paper). Different levels of information asymmetry require the server to design different optimal strategies to achieve the highest possible accuracy with the lowest possible costs.

Moreover, training data is usually non-IID among users in federated learning, which increases the difficulty of the server's incentive mechanism design. The main reason for such non-IID data distribution is that the training data of a given user is typically based on his own usage of the mobile device, which usually is not representative of the entire population distribution. Literature has numerically shown that, non-IID data usually will cause accuracy reduction in federated learning compared with IID data (e.g., [11]). However, the accuracy loss given non-IID data in federated learning cannot be theoretically derived so far. This greatly increases the server's difficulty to evaluate his cost on accuracy.

In this paper, we aim to answer the following key questions:

- *How to incentivize users with multi-dimensional private information to participate and train the federated learning model truthfully and efficiently?*
- *How does the server's knowledge of users' private information influence the server's strategy and cost?*
- *How should the server incentivize users who have non-IID data, to minimize the server's cost on accuracy and payment?*

B. Contributions

We summarize our key results and contributions as follows:

- *Incentive mechanism design with multi-dimensional private information.* To the best of our knowledge, this is one of the first analytical studies on incentive mechanism design for federated learning given users with multi-dimensional private information, considering different levels of information asymmetry.
- *Multi-dimensional contract and server preference characterization.* We analytically solve the server's optimal contract design problem, which is especially challenging under incomplete information due to the non-convexity and high-complexity of the optimization problem. We are able to summarize users' multi-dimensional private information with a single dimension metric, which is the server's complete ordered preference of different users.
- *Investigation on effect of multiple information asymmetry levels.* We reveal the influence of information asymmetry level on the optimal contract, and show that the complexity of contract design increases with information incompleteness. We demonstrate that when users have IID data, 1) comparing with the complete information benchmark, weakly incomplete information does not increase the server's cost, but strongly incomplete information does; 2) choosing the group of users with lowest training cost and delay is not always optimal for the server when information is strongly incomplete, due to poor overall performance or low existence probability of these users. When users have non-IID data, even under weakly incomplete information, server in general will experience a higher cost compared with that in complete information scenario, as the server may choose more than one type of users to avoid highly non-IID training data.
- *Evaluation on accuracy loss given non-IID data.* We characterize the non-IID degree of users' data through the widely adopted method, the earth mover's distance (EMD). Through numerical analysis on CIFAR-10 dataset under various experiment setups, we show that for any given non-IID degree, the accuracy loss given non-IID data has a relatively stable loss coefficient of that given randomly generated IID data. Based on this observation, we extend the model in the IID case to the non-IID case by introducing a loss coefficient. The experiments demonstrate that our optimal incentive mechanism based on this proposed model has a much better performance, compared with the state of art in literature designed for non-IID data.

C. Related Work

Studies on federated learning started in 2016, and most literature has focused on improving training efficiency and effectiveness (e.g., [12]), enhancing security (e.g., [13]), and preserving privacy (e.g., [14]). Most of the results are derived under an optimistic assumption that users are willing to participate in federated learning, which may not be realistic without proper incentives.

A carefully designed incentive mechanism can elicit users' honest behaviors and enhance training efficiency in federated learning [15]. Although federated learning has been increasingly widely implemented in practice, there are only a few important earlier work on the incentive mechanism design, with a few limitations. First, these existing work usually modeled the server's cost/profit along one dimension, e.g., time consumption (e.g., [3]) or training data size (e.g., [6]). Second, these literature did not consider various possible information scenarios. For example, Kang *et al.* [3] only considered the weakly incomplete information where the server knows the number of different types of users, while Sarikaya and Ercetin [7] assumed a complete information scenario where the server knows the private information of users (e.g., costs). Third, few work considered the non-IID data scenario (e.g., [9]). Fourth, these studies usually assumed that the server choose users based on one-dimension private information, and the corresponding solutions are not easily generalizable when users have multi-dimensional private information. Moreover, most of them only presented algorithms for deriving the optimal incentive mechanisms without closed-form solutions (e.g., [3], [8], [9]). Building upon these earlier work, we consider a more general and practical model of multi-dimensional private information, provide a more comprehensive mechanism design with closed-form solutions, and investigate the effect of information completeness. Table I summarizes the key differences between our work and the related literature.

II. System Model

We consider a typical federated learning platform (e.g., timely traffic flow prediction [5] or the popular Gboard

TABLE I
COMPARISON OF MODELS IN LITERATURE

Ref.	Users' Private Information		Server's Cost/Profit		Non-IID Data	Multiple Information Scenarios
	Delay	Cost	Time	Data size		
[3], [7]	√	×	√	×	×	×
[6], [8]	×	√	×	√	×	×
[9]	×	√	×	√	√	×
This paper	√	√	√	√	√	√

system [16]) where the model training is distributed over N users and coordinated by a central server. To simulate users' participation under incomplete information, we will propose a contract-based incentive mechanism where the server provides a set of contract items for each user to choose. In the following, we first introduce the federated learning process, then formulate the contract, and finally specify the users' payoff and the server's cost, respectively.

A. Federated Learning Process

As an illustrative example, federated learning are being applied to timely traffic flow prediction [5]. Since the timely traffic data from each user (e.g., traffic station or traffic company) is limited, federated learning relies data from a lot of users to achieve an effective prediction. It asks users to use their local data to cooperatively train a global learning model. Each user only needs to share model parameters with the server without uploading its raw data, which keeps traffic data (including private license plate number, location, route, and so on) inside the walls of each user.

Specifically, consider an example of data (x_i, y_i), where x_i is the input (e.g., traffic flow at the last moment) and y_i is the label (e.g., traffic flow at this moment). The objective of learning is to find the proper model parameter w that can predict the label y_i based on the input x_i. Denote the prediction value as $\tilde{y}(x_i; w)$. The gap between the prediction $\tilde{y}(x_i; w)$ and the ground truth label y_i is characterized by the prediction loss function $f_i(w)$. If user k uses a set $\mathcal{S}_k$ of data with data size s_k to train the model, the loss function of user k is the average prediction loss on all data $i \in \mathcal{S}_k$, i.e.,

$$F_k(w) = \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} f_i(w).$$

The optimal model parameter w^* minimizes global loss function, which is a weighted average of all users' loss functions:

$$w^* = \arg\min_w f(w) = \arg\min_w \sum_{k=1}^{N} \frac{s_k}{s} F_k(w), \tag{1}$$

where s is the total data size of all users [2].

We consider the widely adopted synchronous update scheme that proceeds in rounds of communication, i.e., all users enter a new global training round simultaneously; the server sends the global parameter to all users at the same time and waits for all users' updates. The key advantage of the synchronous algorithms is that they have provable convergence (e.g., [15], [17]). A typical synchronous federated learning algorithm with one-step local update works as Algorithm 1 [2].

Algorithm 1 Synchronous Federated Learning

Input : Number of iterations D, learning rate η, number of users N, and users' data.
Output: Model parameter w_D

initialize w_0
for *round* $d = 0$; $d < D$; $d++$ **do**
  Server executes:
  select a set $\mathcal{K}$ of users
  send current global parameter w_d to users
  Each user $k \in \mathcal{K}$ **executes:**
  compute local parameter: $w_{d+1}^k \leftarrow w_d - \eta \nabla F_k(w_d)$
  return w_{d+1}^k to server
  Server executes:
  aggregate all users' updates: $w_{d+1} \leftarrow \sum_{k \in \mathcal{K}} \frac{s_k}{s} w_{d+1}^k$

Users have to perform both communication and computation in the federated learning. Communication usually takes time. McMahan *et al.* [2] shows that mobile users usually have a limited upload bandwidth and may even wait for some time before uploading. Meanwhile, computation time becomes shorter and shorter, as each user's on-device dataset is small compared to the total dataset size and modern mobile phones have relatively fast processors. In this paper, we focus on the synchronous federated learning that each user only conducts one step of gradient update in each round, in which case we can reasonably assume that communication time dominates in each round of training. We will discuss the more general case of multiple-step local updates in Appendix XI of the technical report [18]. Moreover, we assume that the powerful server has a large enough bandwidth, so that the communications from multiple users to the server do not interfere with each other.

Because users will suffer time/energy costs due to the model training, the server needs to properly incentivize users' participation by providing rewards. An effective incentive mechanism usually offers heterogeneous rewards for different types of users.

B. User's Types

We consider a population of N users on the federated learning platform. Users are distinguished by two-dimensional private information: the marginal data-usage cost θ and the communication time t. For the convenience of presentation, we refer to a user with $\pi_i \triangleq (\theta_i, t_i)$ as a type-i user. We consider all users belonging to a set $\mathcal{I} = \{1, \ldots, I\}$ of I types. Each type $i \in \mathcal{I}$ has N_i users, with $\sum_{i \in \mathcal{I}} N_i = N$. Though each user could have data different from others,

we first assume that the data is i.i.d. among all users and each user's type does not change in the entire training process. We will further study the case where users have non-IID data in Section IV.[1]

In the presence of users' private information, it is difficult for the server to predict the users' behaviors without complete information about each user's type. To this end, we propose to design a contract mechanism to elicit the private information.

C. Contract Formulation

Contract theory is a promising and widely adopted theoretic tool for dealing with problems with private information. Therefore, we propose a contract theoretic framework to tackle the incentive mechanism design problem.

1) Server's Contract: The server will propose a contract that specifies the relationship among users' communication time, training data size, and reward for the entire training process. Specifically, the contract $\mathcal{C} = (t_{\max}, \boldsymbol{\phi})$ contains a maximum communication time $t_{\max}$ (for all user types) and I contract items $\boldsymbol{\phi} = \{\phi_i\}_{i\in\mathcal{I}}$ (one for each type). The term $t_{\max}$ is the maximum communication time in each global round set by the server, i.e., users with $t_i \leq t_{\max}$ are able to finish the transmission of the parameters in time. Each contract item $\phi_i \triangleq (s_i, r_i)$ specifies the relationship between each type-i user's data size and reward. The term s_i is the required training data size for each type-i user in each global round. The term r_i is the reward (e.g., money) for each type-i user in each global round, if the user completes the training task with required time and data size.[2] The server offers a zero contract item for any user type i with $t_i > t_{\max}$.

2) Users' Choices: At the beginning of the training process, each user decides whether to participate in the training and (if yes) which contract item to choose. If a user chooses the contract item ϕ_i, he needs to use s_i data examples to train the model and sends local updates to the server in time $t_{\max}$. In return, he will get r_i reward in this global round. Users will not participate if their payoff (defined in Section II-D) is negative.

Under such a contract, we specify the users' payoff and the server's cost in the following.

D. Users' Payoff

Each user's payoff in each global round is the difference between the reward offered by the server and the cost of data usage in model training.

We assume that a user's training cost (e.g., time and energy costs) is proportional to the used data size, i.e., $\theta_i s_i$ [4]. Hence, if a type-i user chooses the contract item ϕ_i, his payoff is[3]:

$$U(\theta_i, t_i, \phi_i) = \begin{cases} r_i - \theta_i s_i, & \text{if } t_i \leq t_{\max}, \\ -\theta_i s_i, & \text{if } t_i > t_{\max}. \end{cases} \tag{2}$$

We assume that in each global iteration, every user locally performs one step of mini-batch stochastic gradient decent (SGD) to compute model parameters. Thus from the global perspective, it is equivalent to a centralized mini-batch SGD with batch size $B = \sum_{i\in\mathcal{I}} \mathbb{1}_{t_i\leq t_{\max}} N_i s_i$, where

$$\mathbb{1}_{t_i\leq t_{\max}} = \begin{cases} 1, & \text{if } t_i \leq t_{\max}, \\ 0, & \text{if } t_i > t_{\max}, \end{cases}$$

means that only users with $t_i \leq t_{\max}$ are eligible to train the model. Note that we are able to analyze the general case where users perform multiple steps of local updates and the results turn out to be similar to the one-step case. Due to space limit, both analysis and results of the general case are given in Appendix XI of the technical report [18].[4]

E. Server's Cost

With a fixed training time, the server's cost is determined by the accuracy loss of global model and the total payment to users.

First, we characterize the expected accuracy loss of the global model. We use T to denote the total training time. Thus, the number of global iterations is denoted by $D = T/t_{\max}$. The accuracy loss after D rounds is measured by the difference between the prediction loss with parameter w_D and that with the optimal parameter w^*, i.e., $f(w_D) - f(w^*)$ (defined in Section II-A). The expected difference is bounded by $O(1/\sqrt{BD} + 1/D)$ when users use mini-batch SGD and have IID data [23], [24], where B is the batch size. Thus, server's expected loss in accuracy decreases as the number of iterations D and batch size B increase.

Next, we consider the server's total payment to all users in the entire training process. If all users choose the respective contract items, the total payment is the product of the number of global iterations and the payment to all users in each iteration, i.e., $D \cdot \sum_{i\in\mathcal{I}} \mathbb{1}_{t_i\leq t_{\max}} N_i r_i$.

To summarize, the server's cost is:

$$W(t_{\max}, \boldsymbol{\phi}) = \gamma_1 \min\{\left(\frac{1}{\sqrt{\frac{T}{t_{\max}}\sum_{i\in\mathcal{I}} \mathbb{1}_{t_i\leq t_{\max}} N_i s_i}} + \frac{t_{\max}}{T}\right), C\} + \gamma_2 \frac{T}{t_{\max}} \sum_{i\in\mathcal{I}} \mathbb{1}_{t_i\leq t_{\max}} N_i r_i. \tag{3}$$

[1]In fact non-IID is the general case, and IID is the special case. However, the study of IID case is valuable. First, the IID assumption allows us to derive the close-form solution and provide some useful insights that might go beyond the IID case. That is the reason for quite many federated learning literature (e.g., [19], [20]) analyzing IID data. Second, non-IID data adds additional challenges in terms of modeling and analysis. We will solve the problem by extending our IID model to the non-IID model by introducing an accuracy loss coefficient. Our experimental results demonstrate the effectiveness of our general modeling of the non-IID case.

[2]The server is able to know each user's training data size which is the weight in parameter aggregation step (i.e., (1)) of federated learning [21].

[3]Since we consider synchronous federated learning, our model can be applied to the case where users' payoffs include an additional homogeneous time cost term $\alpha t_{\max}$. Such a time cost only makes the optimal rewards increase by a constant. Thus, we normalize the time cost to zero.

[4]Considering one-step update is for the convenience of modeling the global training accuracy in Section II-E, so that we can derive explicit solutions as well as comprehensible insights. That is because there is no theoretical results of the global accuracy of federated learning in the presence of users' multiple updates in each global iteration. However, this assumption is not restrictive. First, if users perform multiple steps of local updates, we are able to derive similar results by using a general accuracy function $f(S, E, K)$, where S is the total training data size, E is the number of data passes, K is the number of global iterations, and $f(S, E, K)$ is a convex decreasing function [22], [23]. Second, we will show in simulation (Fig. 7) that even if users perform multiple steps of updates in each global iteration, our proposed mechanism still has a good performance.

The first term on the right hand side of (3) characterizes the server's expected loss in accuracy, where γ_1 indicates the server's valuation on accuracy loss. The second term on the right hand side of (3) represents the server's payment to users, where γ_2 indicates the server's valuation on payment. We use $C \in (\frac{1}{\sqrt{\frac{T}{t_{\max}}}} + \frac{t_{\max}}{T}, \infty)$ to characterize the server's finite (and possibly large) accuracy loss when there is no data for training (i.e., $\sum_{i\in\mathcal{I}} \mathbb{1}_{t_i \le t_{\max}} N_i s_i = 0$).

III. Multi-Dimensional Contract Design for Users With IID Data

In this section, we analyze the server's optimal incentive mechanism for users with IID data. To understand the impact of incomplete information, we consider three information scenarios:

1) *Complete information scenario (benchmark):* The server knows each user's type. This provides a lower bound of the server's minimum cost for all information scenarios.
2) *Weakly incomplete information scenario:* The server knows the total number of users and the specific number of each user type, but he does not know which user belongs to which type.
3) *Strongly incomplete information scenario:* The server knows the total number of users and the distribution of user types, but he does not know the specific number of each user type.

In each scenario, we first derive the condition for a feasible contract, and then characterize the optimal contract. Feasibility and optimality of the contract are defined as follows:

Definition 1 (Contract Feasibility): A contract is feasible *if each user achieves the maximum payoff under the contract item designed for his type.*

Definition 2 (Contract Optimality): A contract is optimal *if it minimizes the server's cost among all feasible contracts.*

A. Complete Information Scenario

In this subsection, we study the server's optimal contract in the scenario where the server knows the type of each user. This makes it possible for the server to monitor and make sure that each type of users accepts will not accept any contract item not designed for that type. Even in this case, the server still needs to ensure that each user achieves a non-negative payoff, so that the user will accept the corresponding contract item. In other words, a contract is feasible if and only if it satisfies Individual Rationality (IR) constraints:

Definition 3 (Individual Rationality): A contract is individually rational if each type-i user receives a non-negative payoff by accepting the contract item ϕ_i intended for his type, i.e.,

$$U(\theta_i, t_i, \phi_i) \ge 0, \quad \forall i \in \mathcal{I}. \tag{4}$$

Thus, in the complete information scenario, the optimal contract $\mathcal{C}^{opt}_{complete} = (t^*_{\max}, \boldsymbol{\phi}^*)$ is the solution to the following optimization problem:

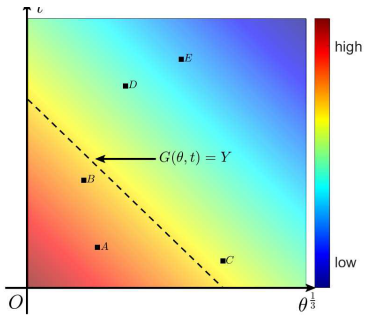


Fig. 1. Server's preference order.

Problem 1 (Contract Design for IID Users Under Complete Information):

$$\begin{aligned} &\min_{t_{\max}, \boldsymbol{\phi}} W(t_{\max}, \boldsymbol{\phi}) \\ &\text{s.t. IR} \quad \text{Constaints in (4).} \end{aligned} \tag{5}$$

We will solve Problem 1 in two steps. First, for any given data size s_i, we derive the server's optimal reward $r^*_i(s_i)$ (Lemma 1). Second, we substitute the optimal reward $r^*_i(s_i)$ into the server's objective function and derive the optimal data size s^*_i as well as the optimal maximum communication time $t^*_{\max}$ (Theorem 1).

Lemma 1: For any given data size s_i (even if it is not optimal), it is optimal for the server to choose the reward as $r^*_i(s_i) = \theta_i s_i, \ \forall i \in \mathcal{I}$.

Proof of Lemma 1 is given in Appendix I of the technical report [18]. Lemma 1 shows that the server will design the contract such that all users get a zero payoff under complete information.

Based on Lemma 1, we can derive the optimal data size for each type that minimizes the server's cost. To illustrate the impact of choosing each user type on the server's cost, we have the following lemma:

Lemma 2: The server's cost of only choosing type i is

$$G(\theta_i, t_i) \triangleq \frac{\gamma_1 t_i}{T} + \left(2^{\frac{1}{3}} + 2^{-\frac{2}{3}}\right) \gamma_2^{\frac{1}{3}} \gamma_1^{\frac{2}{3}} \theta_i^{\frac{1}{3}}, \tag{6}$$

Proof of Lemma 2 is given in Appendix II of the technical report [18]. Lemma 2 characterizes the server's trade-off between users' different dimensions of private information (i.e., θ and t). Thus, we can transform users' two-dimensional private information into a one-dimensional criterion, which indicates the server's preference on different user types:

Definition 4 (Preference): The server has a higher preference on type j than type i (denoted by $j \succ i$) if and only if $G(\theta_j, t_j) < G(\theta_i, t_i)$.

Fig. 1 illustrates how the server's preference on user types changes over the parameter space of (θ, t). More specifically, the server's preference on users' types decreases from the red area (low cost and delay) to the blue area (high cost and delay). The server has the same preference on users whose $(\theta_i^{\frac{1}{3}}, t_i)$ on the same line $G(\theta, t) = Y$, where Y is an arbitrary constant. We denote the set of user types which have the same highest

server preference as follows:

$$\mathcal{J}^{prefer} \triangleq \arg\min_{j\in\mathcal{I}} G(\theta_j, t_j).$$

For example, suppose that there are five user types A, B, C, D, and E with (θ, t) shown in Fig 1. The server's preference order is $A \succ B \succ C \succ D \succ E$ and $\mathcal{J}^{prefer} = \{A\}$.

Theorem 1 characterizes the optimal contract for the server in the complete information scenario under different cases of the set $\mathcal{J}^{prefer}$:

Theorem 1: In the complete information scenario, when users have IID data,

1) *if $\mathcal{J}^{prefer} = \{j\}$, the server's optimal contract is $t^*_{\max} = t_j$, $\phi^*_j = (\frac{1}{N_j \frac{T}{t_j}[\frac{2\gamma_2\theta_j}{\gamma_1}]^{\frac{2}{3}}}, \frac{\theta_j}{N_j \frac{T}{t_j}[\frac{2\gamma_2\theta_j}{\gamma_1}]^{\frac{2}{3}}})$, and $\phi^*_i = \mathbf{0}, \forall i \neq j$.*
2) *if $|\mathcal{J}^{prefer}| > 1$, the server's optimal contract is to select any one type $j \in \mathcal{J}^{prefer}$ with $t^*_{\max} = t_j$, $\phi^*_j = (\frac{1}{N_j \frac{T}{t_j}[\frac{2\gamma_2\theta_j}{\gamma_1}]^{\frac{2}{3}}}, \frac{\theta_j}{N_j \frac{T}{t_j}[\frac{2\gamma_2\theta_j}{\gamma_1}]^{\frac{2}{3}}})$, and $\phi^*_i = \mathbf{0}, \forall i \neq j$.*
3) *if $|\mathcal{J}^{prefer}| > 1$, offering the only positive contract to one type $j \in \mathcal{J}^{prefer}$ leads to the same minimum server cost.*

Proof of Theorem 1 is given in Appendix III of the technical report [18]. Theorem 1 shows that the server only provides a positive contract item for the single most preferred user type and offers the same zero contract item for all other user types.[5] Moreover, the optimal contact always exists but may not be unique, as the most preferred type may not be unique. On the other hand, it is never optimal to select (provide a positive contract item) to multiple user types, even if they all belong to the set $\mathcal{J}^{prefer}$, as this would increase the cost of the server. Also, we can verify that our incentive mechanism under complete information maximizes the social welfare.

Intuitively, having less information would lead to a different behavior of the server. However, we will show in next section that server's optimal contract under weakly incomplete information is the same as that in complete information scenario.

B. Weakly Incomplete Information Scenario

In this subsection, we study the server's optimal contract in the weakly incomplete information scenario. The server does not know which user belongs to which type, but knows the specific number of each user type (i.e., $N_i, \forall i \in \mathcal{I}$).

Since the server cannot force a user to accept certain contract item in this case, he needs to design the contract to further ensure the Incentive Compatibility (IC) constraints:

Definition 5 (Incentive Compatibility): A contract is incentive compatible if each type-i user maximizes his own payoff by choosing the contract item ϕ_i intended for his type, i.e.,

$$U(\theta_i, t_i, \phi_i) \geq U(\theta_i, t_i, \phi_j), \forall i, j \in \mathcal{I}. \tag{7}$$

[5]We assume that each type has enough number of users. One may concern that in reality, there may not be enough users of each type. As a solution, the server can divide users into several groups and each group has an appropriate number of users. Then, the server approximates each group as one type to design the contract. As validated by the experiments in Section V-C, the approximation will not significantly increase the server's payment.

The optimal contract $\mathcal{C}^{opt}_{W-incomplete} = (t^*_{\max}, \boldsymbol{\phi}^*)$ under weakly incomplete information is the solution to Problem 2:

Problem 2 (Contract Design for IID Users Under Weakly Incomplete Information):

$$\min_{t_{\max}, \boldsymbol{\phi}} W(t_{\max}, \boldsymbol{\phi})$$
$$\text{s.t. IR Constaints in (4), IC Constaints in (7).} \tag{8}$$

As the total number of IR and IC constraints is I^2, it is quite complex to solve Problem 2 directly. In the following, we first transform IR and IC constraints into a smaller number of equivalent constraints (Lemma 3). Then, for any given data size s_i, we derive the server's optimal reward $r^*_i(s_i)$ (Lemma 4). Finally, we derive the optimal data size s^*_i and the optimal maximum communication time $t^*_{\max}$ (Theorem 2).

We use $\mathcal{I}'$ to denote the set of user types with communication time no larger than $t_{\max}$, i.e., $\mathcal{I}' = \{i | t_i \leq t_{\max}\}$. We denote $I' = |\mathcal{I}'|$ and reindex the user types in $\mathcal{I}'$ by $\{i_{\mathcal{I}'}\}_{i\in\{1,\ldots,I'\}}$ in the ascending order of marginal cost θ, because as long as the communication time t is no larger than $t_{\max}$, it does not matter anymore. Lemma 3 characterizes contract feasibility:

Lemma 3: Under weakly incomplete information, a contract $\mathcal{C} = (t_{\max}, \boldsymbol{\phi})$ is feasible if and only if the followings are true:

a) for user types in $\mathcal{I}'$, the contract items satisfy the following three conditions:
- *a.1) $r_{I'_{\mathcal{I}'}} - \theta_{I'_{\mathcal{I}'}} s_{I'_{\mathcal{I}'}} \geq 0$;*
- *a.2) $r_{1_{\mathcal{I}'}} \geq \ldots \geq r_{I'_{\mathcal{I}'}} \geq 0$ and $s_{1_{\mathcal{I}'}} \geq \ldots \geq s_{I'_{\mathcal{I}'}} \geq 0$;*
- *a.3) $r_{i+1_{\mathcal{I}'}} + \theta_{i_{\mathcal{I}'}}(s_{i_{\mathcal{I}'}} - s_{i+1_{\mathcal{I}'}}) \leq r_{i_{\mathcal{I}'}} \leq r_{i+1_{\mathcal{I}'}} + \theta_{i+1_{\mathcal{I}'}}(s_{i_{\mathcal{I}'}} - s_{i+1_{\mathcal{I}'}})$, $i \in \{1, \ldots, I'\}$.*

b) for any user type $i \notin \mathcal{I}'$, $s_i = r_i = 0$.

Proof of Lemma 3 is given in Appendix IV of the technical report [18].

Constraint $(a.1)$ ensures that each type of users can get a non-negative payoff by accepting the contract item of type-$I'_{\mathcal{I}'}$ users (with maximum marginal cost $\theta_{I'_{\mathcal{I}'}}$ in $\mathcal{I}'$) as $r_{I'_{\mathcal{I}'}} - \theta_{i_{\mathcal{I}'}} s_{I'_{\mathcal{I}'}} \geq r_{I'_{\mathcal{I}'}} - \theta_{I'_{\mathcal{I}'}} s_{I'_{\mathcal{I}'}} \geq 0, \forall i \in \{1, \ldots, I'\}$. This corresponds to the IR constraints. Both constraints $(a.2)$ and $(a.3)$ are related to IC constraints. Constraint $(a.2)$ shows that the server should request more data from a user type with a lower marginal cost and provide more reward in return. Constraint $(a.3)$ characterizes the relationship between any two neighbor contract items. The results in (b) mean that users with $t_i > t_{\max}$ cannot finish the communication in time, so the required data size and reward in contract are zero.

Based on Lemma 3, Lemma 4 characterizes the optimal rewards for any feasible data size:

Lemma 4: For any given data size $\boldsymbol{s} = \{s_i\}_{i\in\mathcal{I}}$ (even if it is not optimal), it is optimal to choose reward satisfy:

- *for any user type $i \in \mathcal{I}'$,*

$$r^*_i(\boldsymbol{s}) = \begin{cases} \theta_i s_i, & \text{if } i = I'_{\mathcal{I}'}; \\ \theta_i s_i + \sum_{j=i+1_{\mathcal{I}'}}^{I'_{\mathcal{I}'}} (\theta_j - \theta_{j-1}) s_j, & \\ \qquad \text{if } i = 1_{\mathcal{I}'}, \ldots, (I'-1)_{\mathcal{I}'}. & \end{cases}$$

- *for any user type $i \notin \mathcal{I}'$, $r^*_i(\boldsymbol{s}) = 0$.*

Proof of Lemma 4 is given in Appendix V of the technical report [18]. Based on Lemma 3 and Lemma 4, we can significantly simplify Problem 2. The following theorem characterizes the server's optimal contract in weakly incomplete information scenario:

Theorem 2: Given IID data, the server's optimal contract under weakly incomplete information $\mathcal{C}^{opt}_{W-incomplete}$ is the same as that in complete information scenario in Theorem 1, i.e., $\mathcal{C}^{opt}_{complete}$.

Proof of Theorem 2 is given in Appendix VI of the technical report [18]. Here we discuss some intuitions about Theorem 2. Recall that under complete information, the server's optimal contract is to only choose the most preferred user type. Under weakly incomplete information, the server knows the exact number of each user type. Thus, the server can focus on designing a contract to only attract the most preferred type, so that it achieves the same minimum cost as complete information scenario. Next, we will show that when the server does not know the number of each type, he needs to design a more complex contract to deal with all possible situations.

C. Strongly Incomplete Information Scenario

In this section, we consider a scenario where the information about users' types is strongly incomplete. The server does not know the specific number of each user type, but only knows the total number of users N and the distribution of users' types, i.e., the probability of a user being type i (θ_i, t_i) as p_i.

Due to the uncertainty, the server needs to minimize his expected cost in this scenario. Consider the case where $N_i = n_i$ for each user type i with $\sum_{i\in\mathcal{I}} n_i = N$. The probability for this case is

$$P(n_1,\ldots,n_I) = \frac{N! p_1^{n_1}\ldots p_{I-1}^{n_{I-1}} p_I^{N-\sum_{i=1}^{I-1} n_i}}{n_1!\ldots n_{I-1}!(N-\sum_{i=1}^{I-1} n_i)!},$$

and the server's corresponding cost (if all users choose the respective contract items) is

$$W(t_{\max},\boldsymbol{\phi};n_1,\ldots,n_I) = \gamma_2 \frac{T}{t_{\max}} \sum_{i\in\mathcal{I}} \mathbb{1}_{t_i\le t_{\max}} n_i r_i + \gamma_1 \min\{(\frac{1}{\sqrt{\frac{T}{t_{\max}}\sum_{i\in\mathcal{I}} \mathbb{1}_{t_i\le t_{\max}} n_i s_i}} + \frac{t_{\max}}{T}), C\}.$$

Then, the optimal contract $\mathcal{C}^{opt}_{S-incomplete} = (t^*_{\max}, \boldsymbol{\phi}^*)$ is the solution to the following problem:

Problem 3 (Contract Design for IID Users Under Strongly Incomplete Information):

$$\min_{t_{\max},\boldsymbol{\phi}} \mathbb{E}[W(t_{\max},\boldsymbol{\phi})] = \sum_{(n_1,\ldots,n_I)} P(n_1,\ldots,n_I) W(t_{\max},\boldsymbol{\phi};n_1,\ldots,n_I)$$
$$\text{s.t. IR Constaints in (4), IC Constaints in (7).}$$

It is very challenging to directly solve Problem 9 analytically. First, we can show that even after simplifying Problem 3 based on Lemma 3 and Lemma 4, the new optimization problem is not necessarily convex. Second, even the problem is convex in some special cases, there is no closed-form optimal solution due to the high order polynomial equations in KKT conditions. This motivates us to consider a more tractable approach to compute a suboptimal contract.

If the server adopts the previously derived optimal contracts (under complete and weakly incomplete information) in strongly incomplete information scenario, he will have no data for training with a probability of $(1-p_j)^N$ when the user type j with positive contract item turns out to have $n_j = 0$. The server would be very likely to get the no data training cost when p_j and N are not large enough. Inspired by the structure of the previously derived optimal contracts that only choose the most preferred user type, we consider a simplified contract where the server only offers two kinds of contract items, one is positive for a group $\chi \subseteq \mathcal{I}$ of user types and the other is zero for the rest user types in $\mathcal{I}\backslash\chi$. We name such a contract structure as Two-Part Uniform (TPU) contract.

The optimal TPU contract $\mathcal{C}^{TPU,opt}_{S-incomplete}$ is the solution to the following problem:

Problem 4 (TPU Contract Design for IID Users Under Strongly Incomplete Information):

$$\min_{t_{\max},\boldsymbol{\phi},\chi} \mathbb{E}[W(t_{\max},\boldsymbol{\phi})] = \sum_{(n_1,\ldots,n_I)} P(n_1,\ldots,n_I) W(t_{\max},\boldsymbol{\phi};n_1,\ldots,n_I)$$
$$\text{s.t. IR Constaints in (4), IC Constaints in (7),}$$
$$\phi_i = \phi_j > \mathbf{0}, \quad \forall i,j\in\chi; \phi_k = \mathbf{0}, \ \forall k\in\mathcal{I}\backslash\chi.$$

The performance of the optimal TPU contract turns out to be close to the optimal contract $\mathcal{C}^{opt}_{S-incomplete}$ (i.e., the optimal solution of Problem 3), which can be shown through both analytical performance bounds and simulation results.

We denote by χ_m an arbitrary subset of user types in $\mathcal{I}$, and we denote χ^* as the type set that leads to the minimum server cost under the optimal TPU contract. In the following, we will first show the optimal TPU contract given an arbitrary type set χ_m, i.e., $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$ in Lemma 5, then evaluate the performance of $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$ in Theorem 3, and finally provide the guideline for finding the optimal type set χ^*.

First, we characterize the optimal TPU contract under type set χ_m, i.e., $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$. The probability of having n_{χ_m} users belonging to the types in χ_m is:

$$P(n_{\chi_m}) = \binom{N}{n_{\chi_m}} P_{\chi_m}^{n_{\chi_m}} (1-P_{\chi_m})^{N-n_{\chi_m}}, \tag{9}$$

where $P_{\chi_m} = \sum_{i\in\chi_m} p_i$ is the probability that a user belongs to a type in χ_m. We denote by $T_{\chi_m} = \max\{t_i\}_{i\in\chi_m}$ the maximum communication time of user types in χ_m, and we denote by $\Theta_{\chi_m} = \max\{\theta_i\}_{i\in\chi_m}$ the maximum marginal cost of user types in χ_m. Lemma 5 presents the $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$:

Lemma 5: The optimal TPU contract given an arbitrary type set χ_m under strongly incomplete information (i.e., $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$) is: $t^*_{\max} = T_{\chi_m}$,

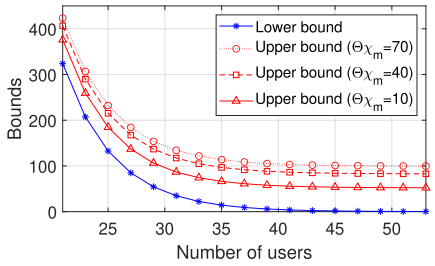


Fig. 2. Server's bounds of cost difference $\Delta W(\chi_m)$.

- *for all user types in* χ_m*:*

$$\phi^* = \left(\frac{1}{\frac{T}{T_{\chi_m}}\left[\frac{2\gamma_2\Theta_{\chi_m}}{\gamma_1}\right]^{\frac{2}{3}}}\left(\frac{\sum_{n_{\chi_m}=1}^{N} P(n_{\chi_m})\frac{1}{\sqrt{n_{\chi_m}}}}{\sum_{n_{\chi_m}=1}^{N} P(n_{\chi_m}) n_{\chi_m}}\right)^{\frac{2}{3}}, \right.$$
$$\left. \frac{\Theta_{\chi_m}}{\frac{T}{T_{\chi_m}}\left[\frac{2\gamma_2\Theta_{\chi_m}}{\gamma_1}\right]^{\frac{2}{3}}}\left(\frac{\sum_{n_{\chi_m}=1}^{N} P(n_{\chi_m})\frac{1}{\sqrt{n_{\chi_m}}}}{\sum_{n_{\chi_m}=1}^{N} P(n_{\chi_m}) n_{\chi_m}}\right)^{\frac{2}{3}}\right);$$

- *for all user types not in* χ_m*,* $\phi^* = \mathbf{0}$.

We use $\Delta W(\chi_m)$ to denote the cost gap between the one achieved under optimal TPU contract given χ_m $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$ and the minimum cost achieved under the optimal contract with complete information $\mathcal{C}^{opt}_{complete}(\chi_m)$ (as in Theorem 1). Such a gap is due to two reasons: (i) the strongly incomplete information, and (ii) the simplification of TPU contract. Theorem 3 shows the bounds of $\Delta W(\chi_m)$:

Theorem 3: In the strongly incomplete information scenario, the cost difference $\Delta W(\chi_m)$ *has the following bounds:*

- *lower bound*:

$$LB \triangleq (1 - P_{\chi_m})^N \gamma_1 \left(C - \frac{T_{\chi_m}}{T}\right)$$

- *upper bound*:

$$UB \triangleq \left(2^{\frac{1}{3}} + 2^{\frac{2}{3}}\right)\gamma_1^{\frac{2}{3}}\gamma_2^{\frac{1}{3}}\Theta_{\chi_m}^{\frac{1}{3}}\left[\left(1.05 + \frac{\sqrt{NP_{\chi_m}}}{e^{0.02P_{\chi_m}^{-2}N}}\right)^{\frac{2}{3}} - 1\right] + LB$$

Proof of Theorem 3 is given in Appendix VII of the technical report [18].

Note that the gap between the minimum cost achieved under $\mathcal{C}^{TPU,opt}_{S-incomplete}(\chi_m)$ and the one under $\mathcal{C}^{opt}_{S-incomplete}$ is no larger than $\Delta W(\chi_m)$. Theorem 3 shows that if the server adopts the optimal TPU contract, he will have a bounded cost difference compared with the complete information scenario. First, both the lower and upper bounds decrease in the number of users N. When N becomes very large (i.e., goes to infinity), the lower bound approaches 0 and the upper bound approaches the constant $(2^{\frac{1}{3}} + 2^{\frac{2}{3}})\gamma_1^{\frac{2}{3}}\gamma_2^{\frac{1}{3}}\Theta_{\chi_m}^{\frac{1}{3}}[(1.05)^{\frac{2}{3}} - 1]$. Moreover, the server can decrease the upper bound by choosing a type set χ_m with a lower marginal cost Θ_{χ_m} as illustrated in Fig. 2. Especially, when N is large, the upper bound approaches to the constant, which is dominated by the Θ_{χ_m} and not related to other parameters of χ_m. This provides the guideline for us to find the optimal type set χ^* under the optimal TPU contract.

Next, we derive the optimal type set χ^*. Since there is a large number of users on federated learning platform like Gboard, we first study the asymptotic behavior under a large user population:

Proposition 1: As the number of users N *approaches infinity, the server will only set a positive contract item for a most preferred type in* $\mathcal{C}^{TPU,opt}_{S-incomplete}$ *(i.e.,* $\chi^* = \{j\}$ *where type* j *can be any type in* $\mathcal{J}^{prefer}$*), which achieves zero cost gap (i.e.,* $\lim_{N\to\infty} \Delta W(\chi^*) = 0$*).*

Proof of Proposition 1 is given in Appendix VIII of the technical report [18]. By the law of large numbers, the empirical value of the number of a type users approaches the expected value computed based on the distribution when N becomes large. Thus, the server will not encounter the situation of having no training data when he only chooses the most preferred type.

Next, we present the insight about which types are in the optimal type set χ^* under any value of N.[6] We may naturally presume that the server would prefer types with higher preference (based on Definition 4). However, the following results show that choosing some user types with lower preference (while excluding other user types with higher preferences) may minimize the server's cost, which is counter-intuitive.

Proposition 2: For the optimal TPU contract under strongly incomplete information $\mathcal{C}^{TPU,opt}_{S-incomplete}$*, it is possible to exist user types* i *and* j *such that* $i \in \chi^*$*,* $j \notin \chi^*$*, and* $G(\theta_i, t_i) > G(\theta_j, t_j)$.

Proof of Proposition 2 is given in Appendix IX of the technical report [18]. The insights behind Proposition 2 are 1) selecting a user type with higher preference may not be optimal when the existence probability of this user type is small; 2) the server's cost is determined by the maximum communication time and maximum marginal cost of user types in the type set. Thus, the combination of several high-preference types may not have a good overall performance.

IV. Multi-Dimensional Contract Design for Users With Non-IID Data

In this section, we analyze the server's incentive mechanism for users with non-IID data. First, we specify the updates on system model in Section IV-A. Then, we derive the server's optimal contracts under different information scenarios in Section IV-B.

A. Model Updates

There are some differences between the model in the non-IID case and that in the IID case. In the following, we first introduce a criterion for evaluating the non-IID degree of users' data, then redefine the user types in the non-IID case, and finally reformulate the server's cost based on the non-IID degree. Other settings of the model remain unchanged from Section II.

1) Non-IID Degree: To study the impact of non-IID data on accuracy, we use the widely adopted average earth mover's distance (EMD) to measure the distribution heterogeneity of data among different user types [11]. Consider a Y class

[6]Due to space limit, we provide detailed analysis regarding how to choose type set χ^* when N is finite in Appendix XII of the technical report [18].

classification problem defined over a compact space $\mathcal{X}$ and a label space $\mathcal{Y} = \{1, \ldots, Y\}$. The data points $\{\boldsymbol{x}, y\}$ are distributed over $\mathcal{X} \times \mathcal{Y}$ following the distribution p. Then, the EMD of user $k \in [K]$ is defined as

$$EMD_k \triangleq \sum_{i=1}^{Y} \left\| p^{(k)}(y=i) - p(y=i) \right\|, \tag{10}$$

where $p^{(k)}(y=i)$ is the proportion of data with label i in user k's data and $p(y=i)$ is the proportion of data with label i in all users' data. Then, the average EMD (denoted by $\overline{EMD}$) is defined as a weighted sum of all users' EMD as follows:

$$\begin{aligned}\overline{EMD} &\triangleq \sum_{k=1}^{K} \frac{s_k}{\sum_{k=1}^{K} s_k} EMD_k \\ &= \sum_{k=1}^{K} \frac{s_k}{\sum_{k=1}^{K} s_k} \sum_{i=1}^{Y} \left\| p^{(k)}(y=i) - p(y=i) \right\|, \end{aligned} \tag{11}$$

where s_k is user k's data size.

Since users with the same marginal data-usage cost θ and communication time t may have different EMD, we need to redefine the user types next.

2) User Types: We still consider a population of N users on the federated learning platform. In the non-IID case, users are distinguished by four-dimensional private information: the marginal data-usage cost θ, the communication time t, the EMD, and the total data size $s^{\max}$. For the convenience of presentation, we denote the parameters of a type-i user as $(\theta_i, t_i, EMD_i, s_i^{\max})$.[7] We consider all users belonging to a set $\mathcal{I} = \{1, \ldots, I\}$ of I types. Each type $i \in \mathcal{I}$ has N_i users, with $\sum_{i \in \mathcal{I}} N_i = N$. We assume that each user's type does not change in the entire training process and each user's data label(s) is (are) independent of the user's (θ, t).[8]

3) Server's Cost: Since the accuracy loss in the non-IID case cannot be analytically derived so far, we will start from the IID model and perform experiments to explore how the non-IID data affects the accuracy.

We use CIFAR-10 dataset to study how the $\overline{EMD}$ affects accuracy loss. The dataset has 50000 training images and 10 output classes (5000 images per class). Fig. 3 gives the experiment results under a given experiment setup.[9] It shows that when the number of training round is not very small (e.g., after 100 rounds), for any given $\overline{EMD}$, the accuracy loss given non-IID data has a relatively stable loss coefficient of that given IID data (i.e., $\overline{EMD} = 0$). The loss coefficient is denoted by $\Pi(\overline{EMD})$. A larger $\overline{EMD}$ increases the relative accuracy loss. Moreover, when we change the experiment setup (e.g., in model structure, number of users, batch size, epoch, fraction of users in each round, balanced/unbalanced data distribution, or label distribution[10]), the relative accuracy loss under each $\overline{EMD}$ only varies in a small range. The average value and fluctuation range of the relative accuracy loss $\Pi(\overline{EMD})$ under different $\overline{EMD}$ are given in Fig. 4.

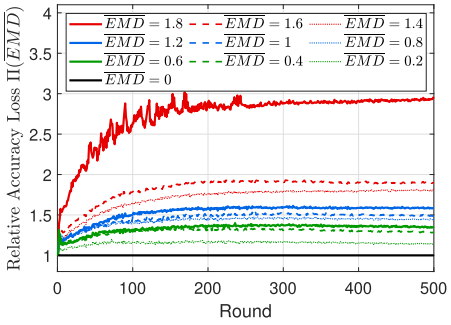


Fig. 3. Relative accuracy loss under different $\overline{EMD}$: an example under a given experiment setup.

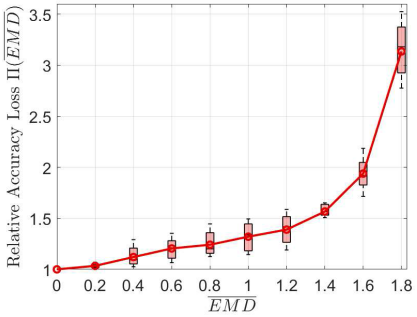


Fig. 4. Relative accuracy loss $\Pi(\overline{EMD})$: average value under various experiment setups.

Based on these observations, we model the server's accuracy loss given non-IID data having a loss coefficient (i.e., $\Pi(\overline{EMD})$) of that given IID data. The specific value of the loss coefficient depends on the $\overline{EMD}$ of the non-IID data. Then, the server's expected cost in non-IID case is

$$\begin{aligned}\tilde{W}(t_{\max}, \boldsymbol{\phi}) &= \gamma_2 \frac{T}{t_{\max}} \sum_{i \in \mathcal{I}} \mathbb{1}_{t_i \le t_{\max}} N_i r_i \\ &+ \gamma_1 \min\{\Pi(\overline{EMD})\left(\frac{1}{\sqrt{\frac{T}{t_{\max}} \sum_{i \in \mathcal{I}} \mathbb{1}_{t_i \le t_{\max}} N_i s_i}} + \frac{t_{\max}}{T}\right), C\}, \end{aligned} \tag{12}$$

The first term on the right hand side of (12) characterizes the server's expected accuracy loss in non-IID case, where γ_1 indicates the server's valuation on accuracy loss. Compared with the IID case, the additional loss coefficient $\Pi(\overline{EMD})$ is the loss coefficient of non-IID data, which increases in $\overline{EMD}$. The second term on the right hand side of (12) represents the server's payment to users, which is the same as the IID case. Term γ_2 indicates the server's valuation on payment.

[7]It is possible to remove $s^{\max}$ from the user type definition in the non-IID case, i.e., assuming that each user has enough data like the IID case. In that case, the key insights in the non-IID case will still hold. The detailed discussion is presented in Appendix XVIII of the technical report [18].

[8]This is reasonable because a user's data should not be related to his marginal cost and communication time.

[9]A convolutional neural network (CNN) model consists of six 3×3 convolution layers (with 64, 64, 128, 128, 256, 256 channels, respectively, and every two followed with 2×2 max pooling), a Drop-out layer (0.5), a fully-connected layer with 10 units and ReLU activation, and a final softmax output layer. Other settings: the number of users $N = 10$, batch size $B = 100$, epoch $E = 1$, fraction of users in each round $C = 1$, learning rate $\eta = 0.01$, decay 0.992, balanced data among users, and same number of labels of each user.

[10]The specific experiment setups are given in Appendix XIII of the technical report [18].

B. Optimal Contract Design

Next, we will first derive the server's optimal contracts given users' non-IID data under complete and weakly incomplete information scenarios, respectively. Then, we will explain the challenges of the optimal incentive mechanism under strongly incomplete information and leave the detailed analysis as future work.

Denote by $\tilde{\chi}$ the set of user types for which the server will set positive contract items. For the simplicity and tractability of the model, we perform the analysis under the following assumptions.

Assumption 1: Each type-i user truthfully reports his EMD_i as well as his total data size $s_i^{\max}$ to the server, and each user randomly chooses data from his dataset to train the model.

Assumption 2: For user types in $\tilde{\chi}$, the server will let each type-i user use the same fraction α of his total data size for training, i.e., $s_i = \alpha s_i^{\max}, \forall i \in \tilde{\chi}$.[11]

Under these assumptions, the server is able to derive the average EMD of any set of user types:

$$\begin{aligned}\overline{EMD}_{\tilde{\chi}} &= \sum_{i\in\tilde{\chi}} \frac{N_i \alpha s_i^{\max}}{\sum_{i\in\tilde{\chi}} N_i \alpha s_i^{\max}} EMD_i \\ &= \sum_{i\in\tilde{\chi}} \frac{N_i s_i^{\max}}{\sum_{i\in\tilde{\chi}} N_i s_i^{\max}} EMD_i \end{aligned} \tag{13}$$

Then, we analyze the server's optimal contracts under different information scenarios:

1) Complete Information Scenario: In this subsection, we study the server's optimal contract given users' non-IID data in the scenario where the server knows the type of each user.

In the complete information scenario, the server needs to make sure that each user with type in $\tilde{\chi}$ get a non-negative payoff. Then, the optimal contract for users with non-IID data under complete information $\mathcal{C}_{complete}^{opt-non} = (t_{\max}^*, \boldsymbol{\phi}^*)$ is the solution to the following optimization problem:

Problem 5 (Contract Design for Non-IID Users Under Complete Information):

$$\begin{aligned} &\min_{t_{\max},\boldsymbol{\phi}} \tilde{W}(t_{\max}, \boldsymbol{\phi}) \\ &\text{s.t. IR Constaints in (4).} \end{aligned} \tag{16}$$

To characterize the optimal solution to Problem 5, we first illustrate the impact of choosing different sets of user types on the server's cost in the following lemma:

Lemma 6: The server's cost of choosing types in set $\tilde{\chi}$ under complete information is in (14)*, as shown at the bottom of the next page, where*

$$\tilde{\alpha}^{com} \triangleq \frac{\max\{t_i\}_{i\in\tilde{\chi}^*} (\gamma_1 \Pi(\overline{EMD}_{\tilde{\chi}^*}))^{\frac{2}{3}}}{\left(\sum_{i\in\tilde{\chi}^*} N_i s_i^{\max}\right)^{\frac{1}{3}} T \left(2\gamma_2 \sum_{i\in\tilde{\chi}^*} N_i \theta_i s_i^{\max}\right)^{\frac{2}{3}}}.$$

[11]Note that Assumption 2 is not restrictive. It is equivalent to fair and random sampling of all incentivized users' data. The General Data Protection Regulation (GDPR) has also called for the fairness in the use of personal data [25]. GDPR is a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA).

Proof of Lemma 6 is given in Appendix XIV of the technical report [18]. Lemma 6 shows that the server's cost is determined by the chosen users' private information (i.e., θ, t, EMD, and $s^{\max}$). Moreover, Lemma 6 actually transforms users' four-dimensional private information into a one-dimensional criterion, which indicates the server's preference on different sets of user types. Smaller $G^{complete}\left(\{\theta_i\}_{i\in\tilde{\chi}}, \max\{t_i\}_{i\in\tilde{\chi}}, \overline{EMD}_{\tilde{\chi}}, \{s_i^{max}\}_{i\in\tilde{\chi}}\right)$ means higher preference.

Then, we characterize the optimal contract for the server in complete information scenario (i.e., $\mathcal{C}_{complete}^{opt-non} = (t_{\max}^*, \boldsymbol{\phi}^*)$) in Theorem 4.

Theorem 4: Under complete information scenario, the server's optimal contract for users with non-IID data is

$$\begin{aligned} &\tilde{\chi}^* = \arg\min_{\tilde{\chi}} G^{complete}; \\ &t_{\max}^* = \max\{t_i\}_{i\in\tilde{\chi}^*}; \\ &\phi_i^* = (s_i^*, \theta_i s_i^*), \forall i \in \tilde{\chi}^*, \textit{where } s_i^* = \min\{\tilde{\alpha}^{com} s_i^{\max}, s_i^{\max}\}; \\ &\phi_j^* = \mathbf{0}, \forall j \notin \tilde{\chi}^*. \end{aligned}$$

Proof of Theorem 4 is given in Appendix XV of the technical report [18]. Here are the insights:

- Theorem 4 shows that the server only provides positive contract items for user types in the most preferred type set $\tilde{\chi}^*$ and offers the same zero contract item for all other user types.
- Since the server knows each user's type, he sets rewards to ensure that each user gets 0 payoff (individual rationality).
- The required data size s_i^* set by the server increases in the average EMD and the maximum communication time $t_{\max}^*$, and it decreases in the marginal cost θ_i. A larger non-IID degree increases the accuracy loss, hence the server needs more data to ensure a good accuracy. A longer communication time decreases the global training rounds, which also increases the accuracy loss. Higher marginal costs increase the server's payment to users, which makes the server set a smaller training data size.
- Compared with the IID case where the server only chooses the most preferred type under complete information, the server may choose more than one type in the non-IID case, to avoid a high non-IID degree of training data which leads to a large server cost.

2) Weakly Incomplete Information Scenario: In this subsection, we study the server's optimal contract in the weakly incomplete information scenario. The server does not know each user's type, but knows the specific number of each user type (i.e., $N_i, \forall i \in \mathcal{I}$).

As we have illustrated in the IID case, the server needs to design the contract with the IR and IC constraints under incomplete information. Thus, the optimal contract for non-IID users under weakly incomplete information scenario $\mathcal{C}_{W-incomplete}^{opt-non} = (t_{\max}^*, \boldsymbol{\phi}^*)$ is the solution to Problem 6:

Problem 6 (Contract Design for Non-IID Users Under Weakly Incomplete Information):

$$\begin{aligned} &\min_{t_{\max},\boldsymbol{\phi}} \tilde{W}(t_{\max}, \boldsymbol{\phi}) \\ &\text{s.t. IR Constaints in (4), IC Constaints in (7).} \end{aligned} \tag{17}$$

To present the optimal solution to Problem 6, we first illustrate the impact of choosing different sets of users on the server's cost under weakly incomplete information in the following lemma:

Lemma 7: The server's cost of choosing types in set $\tilde{\chi}$ under weakly incomplete information is in (15)*, as shown at the bottom of the page, where* $\tilde{\alpha}^{W-in} \triangleq \frac{\max\{t_i\}_{i\in\tilde{\chi}^*}(\gamma_1\Pi(\overline{EMD}_{\tilde{\chi}^*}))^{\frac{2}{3}}\left(\sum_{i\in\chi^*}N_i s_i^{\max}\right)^{-\frac{1}{3}}}{T\left(2\gamma_2\sum_{i\in\tilde{\chi}^*}N_i\left(\theta_i s_i^{\max}+\sum_{j=i+1'_{\mathcal{I}}}^{I'_{\mathcal{I}}}(\theta_j-\theta_{j-1})s_j^{\max}\right)\right)^{\frac{2}{3}}}$.

Proof of Lemma 7 is given in Appendix XVI of the technical report [18]. Lemma 7 characterizes how users' multi-dimensional private information influences the server's cost under weakly incomplete information. According to the cost, the server has higher preference on the set of user types with smaller $G^{W-incomplete}\left(\{\theta_i\}_{i\in\tilde{\chi}}, \max\{t_i\}_{i\in\tilde{\chi}}, \overline{EMD}_{\tilde{\chi}}, \{s_i^{\max}\}_{i\in\tilde{\chi}}\right)$ under weakly incomplete information.

Then, the optimal contract for the server in weakly incomplete information scenario (i.e., $\mathcal{C}^{opt-non}_{W-incomplete} = (t^*_{\max}, \boldsymbol{\phi}^*)$) is given in Theorem 5.

Theorem 5: In the weakly incomplete information scenario, the server's optimal contract for users with non-IID data is

$$\begin{aligned}
&\tilde{\chi}^* = \arg\min_{\tilde{\chi}} G^{W-incomplete};\\
&t^*_{\max} = \max\{t_i\}_{i\in\tilde{\chi}^*};\\
&\phi_i^* = (s_i^*, r_i^*), \forall i \in \tilde{\chi}^*, \textit{where } s_i^* = \min\{\tilde{\alpha}^{W-in} s_i^{\max}, s_i^{\max}\}\\
&\qquad \textit{and } r_i^*(\boldsymbol{s}) \textit{ is given in Lemma 4};\\
&\phi_j^* = \mathbf{0}, \forall j \notin \tilde{\chi}^*.
\end{aligned}$$

Proof of Theorem 5 is given in Appendix XVII of the technical report [18]. Theorem 5 shows that the server only provides positive contract items for user types in the most preferred type set of weakly incomplete information scenario, and offers the same zero contract item for all other user types. Moreover, since $G^{W-incomplete}$ is different from $G^{complete}$, the server may choose different types under weakly incomplete information from that under complete information, and thus a different optimal contract. This is different from the IID case where the server always has the same optimal contract in both complete and weakly incomplete information scenarios.

3) Strongly Incomplete Information Scenario: In this subsection, we discuss the difficulty of the optimal incentive mechanism design in strongly incomplete information scenario. The server does not know the specific number of each user type, but only knows the total number of users N and the distribution of users' types, i.e., the probability of a user being type i as p_i.

Similar to the analysis in the IID case in Section III-C, the optimal contract given non-IID data under strongly incomplete information is the solution to the following optimization problem:

Problem 7 (Contract Design for Non-IID Users Under Strongly Incomplete Information):

$$\begin{aligned}
&\min_{t_{\max},\boldsymbol{\phi}} \mathbb{E}[\tilde{W}(t_{\max}, \boldsymbol{\phi})]\\
&\qquad = \sum_{(n_1,\ldots,n_I)} P(n_1,\ldots,n_I)\tilde{W}(t_{\max}, \boldsymbol{\phi}; n_1,\ldots,n_I)\\
&\text{s.t. IR Constaints in (4), IC Constaints in (7).}
\end{aligned}$$

First of all, it is very challenging to directly solve Problem 18 analytically. We can show that even after simplifying Problem 7 based on Lemma 4, the new optimization problem is not necessarily convex. Furthermore, even the problem is convex in some special cases, there is no closed-form optimal solution due to the high order polynomial equations in KKT conditions. Second, users' non-IID data make it hard to evaluate the performance of a sub-optimal incentive mechanism. Choosing different types will cause different non-IID degrees of the training data, which directly affects the accuracy. Thus,

$$G^{complete}\left(\{\theta_i\}_{i\in\tilde{\chi}}, \max\{t_i\}_{i\in\tilde{\chi}}, \overline{EMD}_{\tilde{\chi}}, \{s_i^{max}\}_{i\in\tilde{\chi}}\right)$$

$$\triangleq \begin{cases} \dfrac{\gamma_1\Pi(\overline{EMD}_{\tilde{\chi}})\max\{t_i\}_{i\in\tilde{\chi}}}{T} + \left(2^{\frac{1}{3}}+2^{-\frac{2}{3}}\right)\gamma_2{}^{\frac{1}{3}}\left(\gamma_1\Pi(\overline{EMD}_{\tilde{\chi}})\right)^{\frac{2}{3}}\left(\dfrac{\sum_{i\in\tilde{\chi}}N_i\theta_i s_i^{max}}{\sum_{i\in\tilde{\chi}}N_i s_i^{max}}\right)^{\frac{1}{3}}, & \text{if } \tilde{\alpha}^{com}\le 1,\\ \gamma_1\Pi(\overline{EMD}_{\chi})\left(\sqrt{\dfrac{\max\{t_i\}_{i\in\tilde{\chi}}}{T\sum_{i\in\tilde{\chi}}N_i s_i^{\max}}} + \dfrac{\max\{t_i\}_{i\in\tilde{\chi}}}{T}\right) + \gamma_2\dfrac{T}{\max\{t_i\}_{i\in\tilde{\chi}}}\sum_{i\in\tilde{\chi}}N_i\theta_i s_i^{\max}, & \text{if } \tilde{\alpha}^{com}>1. \end{cases} \tag{14}$$

$$G^{W-incomplete}\left(\{\theta_i\}_{i\in\tilde{\chi}}, \max\{t_i\}_{i\in\tilde{\chi}}, \overline{EMD}_{\tilde{\chi}}, \{s_i^{\max}\}_{i\in\tilde{\chi}}\right)$$

$$\triangleq \begin{cases} \dfrac{\gamma_1\Pi(\overline{EMD}_{\tilde{\chi}})\max\{t_i\}_{i\in\tilde{\chi}}}{T} + \left(2^{\frac{1}{3}}+2^{-\frac{2}{3}}\right)\gamma_2{}^{\frac{1}{3}}\left(\gamma_1\Pi(\overline{EMD}_{\tilde{\chi}})\right)^{\frac{2}{3}}\left(\dfrac{\sum_{i\in\tilde{\chi}}N_i\left(\theta_i s_i^{\max}+\sum_{j=i+1'_{\mathcal{I}}}^{I'_{\mathcal{I}}}(\theta_j-\theta_{j-1})s_j^{\max}\right)}{\sum_{i\in\tilde{\chi}}N_i s_i^{\max}}\right)^{\frac{1}{3}}, & \text{if } \tilde{\alpha}^{W-in}\le 1,\\ \dfrac{\gamma_1\Pi(\overline{EMD}_{\tilde{\chi}})\max\{t_i\}_{i\in\tilde{\chi}}}{T} + \dfrac{\gamma_1\Pi(\overline{EMD}_{\chi})\sqrt{\max\{t_i\}_{i\in\tilde{\chi}}}}{\sqrt{T\sum_{i\in\tilde{\chi}}N_i s_i^{\max}}} + \gamma_2\dfrac{T\sum_{i\in\tilde{\chi}}N_i(\theta_i s_i^{\max}+\sum_{j=i+1_{\mathcal{I}'}}^{I'_{\mathcal{I}'}}(\theta_j-\theta_{j-1})s_j^{\max})}{\max\{t_i\}_{i\in\tilde{\chi}}}, & \text{if } \tilde{\alpha}^{W-in}>1. \end{cases} \tag{15}$$

even if we propose a sub-optimal contact, unless after an exhaustive search, we do not know which set of types can minimize the server's cost by using this proposed contact. Without the minimum cost that the proposed contact can achieve, it is difficult to evaluate its performance. Therefore, we will perform further study on this problem as the future work.

V. Numerical Experiments

In this section, we perform numerical experiments to evaluate the performance of the proposed contracts and validate our analytical results. Specifically, the experiments for IID case is given in Section V-A and that for non-IID case is given in Section V-B.

A. Experiments for IID Case

In this subsection, we first present the good performance of the contracts in three information scenarios, compared with an optimal uniform contract benchmark defined as follows (Fig. 5). Then, we show that the server does not always choose user types with higher preference under strongly incomplete information (Fig. 6). Finally, we train a federated learning model with users' multiple local updates, to verify the robustness of our contracts' performance (Fig. 7).

Regarding the system parameters, we choose $T = 10$, $\gamma_1 = 6751.269$, $\gamma_2 = 1$, and $C = 6.2$. There are five user types with parameters $(\theta_A^{\frac{1}{3}}, t_A) = (2.6, 1.5)$, $(\theta_B^{\frac{1}{3}}, t_B) = (2.1, 4)$, $(\theta_C^{\frac{1}{3}}, t_C) = (7.1, 1.3)$, $(\theta_D^{\frac{1}{3}}, t_D) = (3.6, 7.5)$, and $(\theta_E^{\frac{1}{3}}, t_E) = (5.6, 8.5)$.[12] The preference order is $A \succ B \succ C \succ D \succ E$. The fractions (distribution, respectively) of each type in complete and weakly incomplete (strongly incomplete, respectively) information scenario are $p_A = p_B = p_C = p_D = p_E = 0.2$.

We consider an optimal *uniform contract* for IID case as the benchmark, which contains a single uniform contract item for all users. Specifically, $t^*_{\max} = t_E$, $\phi^* = (\frac{1}{N\frac{T}{t_E}[\frac{2\gamma_2\theta_C}{\gamma_1}]^{\frac{2}{3}}}, \frac{\theta_C}{N\frac{T}{t_E}[\frac{2\gamma_2\theta_C}{\gamma_1}]^{\frac{2}{3}}})$.

In Fig. 5, we compare the server's cost under three different information scenarios and the uniform contract: 1) comparing with complete information, weakly incomplete information does not increase server cost, but strongly incomplete information does; 2) the performance of the optimal TPU contract is very close to that of the optimal contract under strongly incomplete information, especially when the number of users is large; 3) all designed contracts in the three information scenarios achieve up to 73.72% cost reduction of uniform contract when N is large; 4) When the number of users increases, the server's cost decreases under strongly incomplete information and finally approaches to that under complete information.

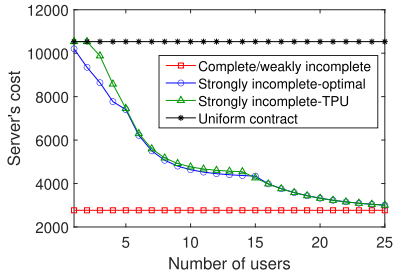


Fig. 5. Cost comparison.

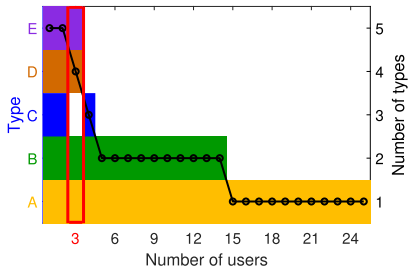


Fig. 6. Server's type set χ^* in strongly incomplete information scenario.

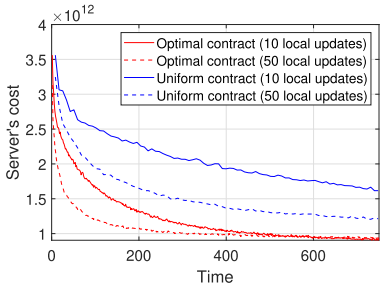


Fig. 7. Server's cost when users make multiple updates in each global round under complete/weakly incomplete information.

In Fig. 6, we verify our insights in Proposition 2 about the types in the optimal type set χ^* under strongly incomplete information. Interestingly, when the number of users is 3, the server chooses type A, B, D, E instead of A, B, C, D, though A, B, C, D rank top four in the order of preference.[13] Moreover, when the total number of users decreases, the number of chosen user types increases. The server needs to ensure a high enough existence probability of chosen user types to avoid the cost of no training data, especially when N is small.

In Fig. 7, we show that the performance of our contract is robust when each user executes multiple local updates per round. Specifically, we train a federated learning model on CIFAR-10 dataset in complete/weakly incomplete information

[12]The values of γ_1 and γ_2 are set by the server based on his needs or interests. Different values of γ_1 and γ_2 will influence the server's optimal strategies and costs. When γ_1 increases, the server attaches more importance on accuracy loss. In this case, the server prefers users with a smaller communication time (and a smaller average EMD) as indicated by the server's preference criteria in the paper. Also, the server will require more data from incentivized users as shown by the contracts. When γ_2 increases, the server pays more attention on the payment to users. In this case, he prefers users with smaller data-usage costs, and he will require less data from incentivized users to reduce the payment. This paper provides a general framework that applies to all possible combinations of (γ_1, γ_2) for different cases. In the numerical experiments, we just give specific values to γ_1 and γ_2 as an example, and they could take other values. The specific value choices will be up to different applications and are outside the scope of this paper.

[13]Note that the number of types that the users may belong to can be larger than the number of users. For example, there are three users, each with a 0.2 probability of belonging to one out of five types.

scenario[14] with $T = 750$, $\gamma_1 = 4.394 \times 10^{12}$, and $N = 500$. Other parameters remain unchanged as before. Our convolutional neural network (CNN) model consists of six 3×3 convolution layers (with 64, 64, 128, 128, 256, 256 channels, respectively, and every two followed with 2×2 max pooling), a Drop-out layer (0.5), a fully-connected layer with 10 units and ReLU activation, and a final softmax output layer. The server's cost in Fig. 7 consists of the accuracy loss in experiment and the total payment to users. Even if users perform multiple updates in each global iteration, the proposed contracts have a better performance than that of the uniform contract, up to 45.69% (37.37%, respectively) cost reduction for 10 (50, respectively) local updates per global iteration.

B. Experiments for Non-IID Case

In this subsection, we evaluate the performance of our proposed incentive mechanism under different non-IID degrees by comparing with three benchmarks in literature.

1) Experiment Setup: Regarding the experiment setup, we consider three non-IID cases:

- Non-IID(1): each user is randomly assigned 100 data with one label.
- Non-IID(2): each user is randomly assigned 2 labels, each label with 50 data.
- Non-IID(5): each user is randomly assigned 5 labels, each label with 20 data.

That means, all users have the same EMD_i in each non-IID case and the same $s_i^{\max} = 100$. We still use the CIFAR-10 dataset for experiments. According to the empirical results in Fig. 4, the values of $\Pi(\overline{EMD})$ are 3.1343, 1.9380, and 1.3201 for the non-IID(1), non-IID(2), and non-IID(5) cases, respectively. Other experiment parameters are the same as that in IID case.

2) Benchmark 1 - Optimal Uniform Contract (OUC): OUC contains a single uniform contract item for all users, i.e., $t_{\max}^* = t_E$, $\phi^* = \left(\frac{1}{N\frac{T}{t_E}[\frac{2\gamma_2\theta_C}{\gamma_1\Pi(\overline{EMD})}]^{\frac{2}{3}}}, \frac{\theta_C}{N\frac{T}{t_E}[\frac{2\gamma_2\theta_C}{\gamma_1\Pi(\overline{EMD})}]^{\frac{2}{3}}} \right)$.

3) Benchmark 2 - Incentive Mechanism in [9] (RMA): The authors proposed an auction called RMA for non-IID data under weakly incomplete information, but they did not derive the optimal required data size for each user or consider each user's communication delay.

4) Benchmark 3: Incentive Mechanism in [6] (SBG): The authors formulated a stackelberg game between the server and users in the complete information scenario but did not distinguish between IID data and non-IID data.

5) Experiment Results: We denote our proposed contract by ALC, the acronym of accuracy loss coefficient which is a key feature of our non-IID model. Fig. 8 shows server's cost comparison among ALC, RMA [9], OUC, and SBG [6] in three non-IID cases, respectively. Note that, for ALC, although the server may have different optimal contracts under complete and weakly incomplete information, under our experiment setup where all users have the same EMD_i and $s_i^{\max}$, the server has the same optimal contract and thus the same cost in two information scenarios. The simulation results show that, the proposed ALC has a better performance than the three benchmarks. The maximum cost reduction of ALC compared with other benchmarks is summarized in Table II.

[14]The performance of strongly incomplete information is almost the same, because the number of users N is very large in this case (Proposition 1).

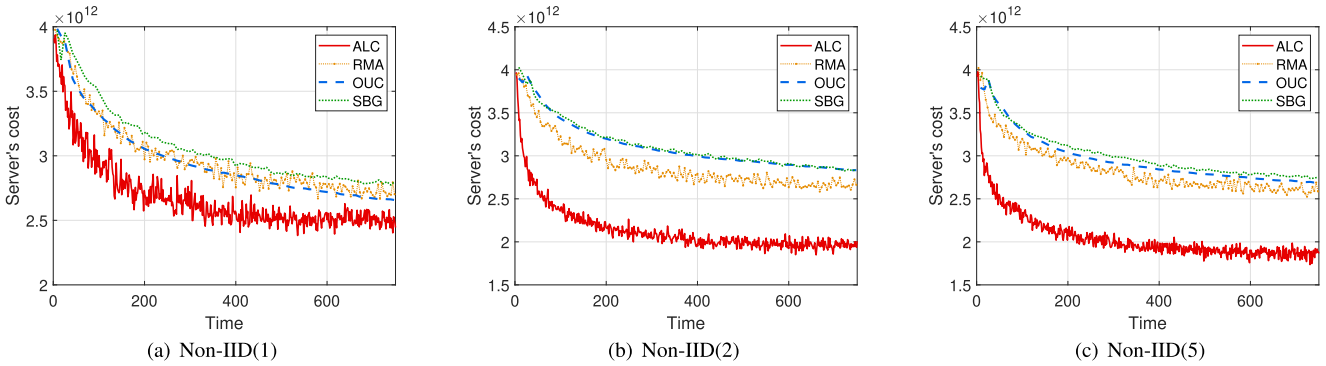
TABLE II
SERVER COST REDUCTION OF ALC COMPARED WITH THREE BENCHMARKS

Case \ Benchmark	RMA	OUC	SBG
Non-IID(1)	16.79%	16.45%	18.88%
Non-IID(2)	30.64%	35.42%	35.69%
Non-IID(5)	33.62%	36.17%	37.01%

Moreover, we compare the rewards paid to users as summarized in Table III. The server's payment to users under the proposed mechanism ALC is medium compared with other benchmarks. The payment of ALC is 29.37, 21.77, and 37.65 times higher than RMA in three Non-IID cases, respectively, and 84.29%, 88.21%, and 78.26%, lower than OUC in three Non-IID cases, respectively. The high payment of OUC is due to the fact that OUC offers an optimal uniform contract item for all users including those with high costs. Although a larger payment can attract more users' participation and improve the accuracy, the accuracy cannot be significantly improved when the number of participating users reaches a certain threshold. Thus, a mechanism with a very high payment to users (e.g., OUC) leads to a large cost for the server. The low payment in RMA is due to the fact that user selection in RMA is mainly based on users' costs. A mechanism with a very low payment to users (e.g., RMA) does not provide enough incentives for users' participation, leading to a poor accuracy. Our proposed ALC mechanism incentivizes users by properly selecting users based on their multi-dimensional heterogeneity, which leads to a medium payment but minimizes the server's cost (on both accuracy loss and payment).

C. Discussion on Implementation

Let's consider a concrete example of a mobile phone keyboard such as Gboard (Google Keyboard). A large amount of local data will be generated when users use the keyboard app on their mobile devices. Suppose that Google server wants to train a next-word prediction model based on users' data. The server can announce the learning project to users through the app and encourage their participation. If a user wants to know more about this project, the app will display an interface to show the specific contract items for users to choose (similar as the request for permission that we commonly see on Android devices). If the user thinks one contract item is optimal and beneficial to him, he will choose the contract item and sign the contract with the server. Once enough users decide to participate, the server will start the training process by broadcasting an initial global model to all participating users. On behalf of the user, the app will download this global model and upload the model updates generated by the training on the user's local data. Algorithm 1 captures the detailed model training process. After finishing the model training project,



(a) Non-IID(1) (b) Non-IID(2) (c) Non-IID(5)

Fig. 8. Server' costs in three non-IID cases under complete and weakly incomplete information.

TABLE III
COMPARISON OF SERVER'S PAYMENT TO USERS IN EACH ROUND

Case \ Mechanism	ALC	RMA	OUC	SBG
Non-IID(1)	2,812,160	92,610	17,895,550	3,397,204
Non-IID(2)	2,109,120	92,610	17,895,550	3,397,204
Non-IID(5)	3,579,110	92,610	16,463,906	3,397,204

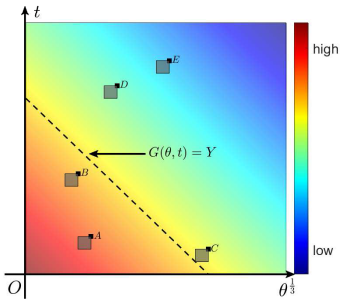


Fig. 9. User distribution.

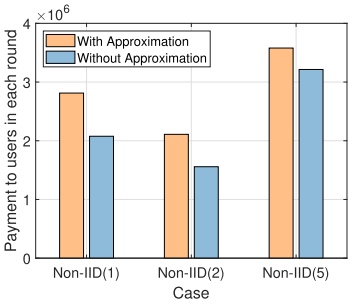


Fig. 10. Server's payment to users in each round with/without the approximation of user types.

the server will give users rewards (e.g., money) based on the contract that they have signed.

For simplicity in reality, the server may divide users with different data-usage costs and communication time into several groups based on market research, and each group can be approximated as a super-type. In this case, the server only needs to design one contract item for each group (instead of each type), and users only need to choose the optimal contract item from a few options.

Next, we perform experiments to validate that the approximation of users to several super-types will not significantly increase the server's payment. For the experiment setup, we consider 500 users with different data-usage cost θ_i and communication time t_i from each other (i.e., there are 500 types). The users with (θ_i, t_i) uniformly distributed over $[2.1, 2.6] \times [1, 1.5] \cup [1.6, 2.1] \times [3.5, 4] \cup [6.6, 7.1] \times [0.8, 1.3] \cup [3.1, 3.6] \times [7, 7.5] \cup [5.1, 5.6] \times [8, 8.5]$ (i.e., the gray squares in Fig. 9). According to such a distribution, we divide users into five groups: group 1 users with $2.1 \leq \theta_i^{\frac{1}{3}} \leq 2.6$ and $1 \leq t_i \leq 1.5$, group 2 users with $1.6 \leq \theta_i^{\frac{1}{3}} \leq 2.1$ and $3.5 \leq t_i \leq 4$, group 3 users with $6.6 \leq \theta_i^{\frac{1}{3}} \leq 7.1$ and $0.8 \leq t_i \leq 1.3$, group 4 users with $3.1 \leq \theta_i^{\frac{1}{3}} \leq 3.6$ and $7 \leq t_i \leq 7.5$, and group 5 users with $5.1 \leq \theta_i^{\frac{1}{3}} \leq 5.6$ and $8 \leq t_i \leq 8.5$. Each group has 100 users due to the uniform distribution. Other parameter settings are the same as Section V-B. We approximate each group as a super-type, with $(\theta_A^{\frac{1}{3}}, t_A) = (2.6, 1.5)$ for group 1 users, $(\theta_B^{\frac{1}{3}}, t_B) = (2.1, 4)$ for group 2 users, $(\theta_C^{\frac{1}{3}}, t_C) = (7.1, 1.3)$ for group 3 users, $(\theta_D^{\frac{1}{3}}, t_D) = (3.6, 7.5)$ for group 4 users, and $(\theta_E^{\frac{1}{3}}, t_E) = (5.6, 8.5)$ for group 5 users.

For the user type approximation in the experiment setup, the approximated data-usage cost could be 126.10% higher than the original data-usage cost, and the approximated communication time could be 62.50% higher than the real value. However, as shown in Fig. 10, the approximation only increases the server's payment by 35.43% in the Non-IID(1) and Non-IID(2) cases and 11.35% in the Non-IID(5) case. Therefore, the approximation of users to several super-types will not significantly increase the server's payment.

VI. CONCLUSION

This paper has focused on the important issue of incentive mechanism design in federated learning. To the best of our knowledge, this is one of the first papers that deal with multi-dimensional private information for federated learning,

considering different levels of information asymmetry as well as IID/non-IID training data. One of our key contributions is to identify a way to summarize users' multi-dimensional private information with a one-dimensional metric. We have also revealed the effect of information asymmetry levels. The experiments demonstrate the good performance of our proposed contracts in all three information scenarios, for both IID and non-IID data. For the future work, we will consider the contract design for competing servers when multiple servers are interested in using the data from the same pool of users to train similar types of machine learning models.

REFERENCES

[1] N. Ding, Z. Fang, and J. Huang, "Incentive mechanism design for federated learning with multi-dimensional private information," in *Proc. 18th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, Jun. 2020, pp. 1–8.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[3] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[4] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE INFOCOM*, Apr. 2019, pp. 1387–1395.

[5] Y. Liu, J. J. Q. Yu, J. Kang, D. Niyato, and S. Zhang, "Privacy-preserving traffic flow prediction: A federated learning approach," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7751–7763, Aug. 2020.

[6] S. Feng, D. Niyato, P. Wang, D. I. Kim, and Y.-C. Liang, "Joint service pricing and cooperative relay communication for federated learning," in *Proc. Int. Conf. Internet Things (iThings)*, Jul. 2019, pp. 815–820.

[7] Y. Sarikaya and O. Ercetin, "Motivating workers in federated learning: A stackelberg game perspective," *IEEE Netw. Lett.*, vol. 2, no. 1, pp. 23–27, Mar. 2020.

[8] Y. Zhan, P. Li, Z. Qu, D. Zeng, and S. Guo, "A learning-based incentive mechanism for federated learning," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6360–6368, Jul. 2020.

[9] Y. Jiao, P. Wang, D. Niyato, B. Lin, and D. I. Kim, "Toward an automated auction framework for wireless federated learning services market," *IEEE Trans. Mobile Comput.*, early access, May 14, 2020, doi: 10.1109/TMC.2020.2994639.

[10] Z. Wang, L. Gao, and J. Huang, "Multi-dimensional contract design for mobile data plan with time flexibility," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2018, pp. 51–60.

[11] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*. [Online]. Available: http://arxiv.org/abs/1806.00582

[12] J. Ren, G. Yu, and G. Ding, "Accelerating DNN training in wireless federated edge learning systems," 2019, *arXiv:1905.09712*. [Online]. Available: http://arxiv.org/abs/1905.09712

[13] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," 2018, *arXiv:1808.04866*. [Online]. Available: http://arxiv.org/abs/1808.04866

[14] M. Hao, H. Li, G. Xu, S. Liu, and H. Yang, "Towards efficient and privacy-preserving federated deep learning," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[15] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*. [Online]. Available: http://arxiv.org/abs/1912.04977

[16] *Gboard*. Accessed: May 15, 2020. [Online]. Available: https://apps.apple.com/us/app/gboard-the-google-keyboard/id1091700242

[17] W. Y. B. Lim *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," 2019, *arXiv:1909.11875*. [Online]. Available: http://arxiv.org/abs/1909.11875

[18] *Online Technical Report*. Accessed: Sep. 25, 2020. [Online]. Available: https://www.dropbox.com/s/802w3vr74pgimku/appendix.pdf?dl=0

[19] Z. Zhao, C. Feng, H. H. Yang, and X. Luo, "Federated-Learning-Enabled intelligent fog radio access networks: Fundamental theory, key techniques, and future trends," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 22–28, Apr. 2020.

[20] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo, "Analyzing federated learning through an adversarial lens," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 634–643.

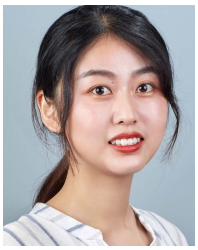
[21] D. Conway-Jones, T. Tuor, S. Wang, and K. K. Leung, "Demonstration of federated learning in a resource-constrained networked environment," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2019, pp. 484–486.

[22] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," 2019, *arXiv:1910.06378*. [Online]. Available: http://arxiv.org/abs/1910.06378

[23] M. Li, T. Zhang, Y. Chen, and A. J. Smola, "Efficient mini-batch training for stochastic optimization," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 661–670.

[24] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches," *J. Mach. Learn. Res.*, vol. 13, pp. 165–202, Jan. 2012.

[25] *General Data Protection Regulation*. Accessed: Jun. 18, 2020. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/



Ningning Ding received the B.S. degree from Southeast University, Nanjing, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. Her research interests include the field of network economics and optimization, with current emphasis on pricing and mechanism design for networked systems.

Zhixuan Fang received the B.S. degree in physics from Peking University, China, in 2013, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2018. He is currently a tenure-track Assistant Professor with the Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University. His main research interests include the design and analysis of multi-agent systems and networked systems.

Jianwei Huang (Fellow, IEEE) received the Ph.D. degree from Northwestern University in 2005. He worked as a Post-Doctoral Research Associate with Princeton University from 2005 to 2007. He is currently a Presidential Chair Professor and the Associate Dean of the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen. He is also the Vice President of the Shenzhen Institute of Artificial Intelligence and Robotics for Society. He has published more than 280 papers in leading international journals and conferences in the area of network optimization and economics, with a total Google Scholar citations of more than 12 700 and an H-index of 57. He has coauthored seven books, including the textbook on *Wireless Network Pricing*. He has been a Distinguished Lecturer of IEEE Communications Society and a Clarivate Analytics Highly Cited Researcher in Computer Science. He is the coauthor of nine Best Paper awards, including the IEEE Marconi Prize Paper Award in Wireless Communications in 2011. He received the CUHK Young Researcher Award in 2014 and IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award in 2009. He was a recipient of the IEEE ComSoc Multimedia Communications Technical Committee Distinguished Service Award in 2015 and the IEEE GLOBECOM Outstanding Service Award in 2010. He has served as the Chair for IEEE ComSoc Cognitive Network Technical Committee and Multimedia Communications Technical Committee. He has served as an Associate Editor for IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS - Cognitive Radio Series, and IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is the Associate Editor-in-Chief of IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.