Performance-Based Pricing for Federated Learning via Auction

Zitao Li
Alibaba Group
zitao.l@alibaba-inc.com

Bolin Ding
Alibaba Group
bolin.ding@alibaba-inc.com

Liuyi Yao
Alibaba Group
yly287738@alibaba-inc.com

Yaliang Li
Alibaba Group
yaliang.li@alibaba-inc.com

Xiaokui Xiao
National University of Singapore
xkxiao@nus.edu.sg

Jingren Zhou
Alibaba Group
jingren.zhou@alibaba-inc.com

ABSTRACT

Many machine learning techniques rely on plenty of training data. However, data are often possessed unequally by different entities, with a large proportion of data being held by a small number of data-rich entities. It can be challenging to incentivize data-rich entities to help train models with others via federated learning (FL) if there are no additional benefits. This difficulty arises because these data-rich entities cannot enjoy the revenue increment generated from the improved performances on tasks controlled by data-limited entities. In this paper, we investigate pricing mechanisms through auctions for FL, focusing on auction scenarios with one data seller and some data-limited entities as buyers. The mechanisms aim to account for buyers' performance gains from the FL and provide equitable monetary compensation to the data seller. We first formulate the task as a performance-based auction mechanism design problem and offer a template that can accommodate multiple kinds of auctions with different desiderata. Utilizing this template, we instantiate different truthful strategies with different goals, including maximizing social welfare and maximizing the seller's profit in auctions. In addition, considering the randomness between the model test performance used in the auction and the actual performance in a production environment, we provide theoretical analyses to quantify the impact of the uncertainty on the social welfare or the seller's profit of auction mechanisms. We provide experimental results based on two datasets with synthetic buyers' valuation to illustrate the truthfulness, social welfare, and data sellers' profit.

PVLDB Reference Format:
Zitao Li, Bolin Ding, Liuyi Yao, Yaliang Li, Xiaokui Xiao, and Jingren Zhou. Performance-Based Pricing for Federated Learning via Auction . PVLDB, 17(6): 1269 - 1282, 2024.
doi:10.14778/3648160.3648169

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/ZiTao-Li/fl_auction.

1 INTRODUCTION

Machine learning techniques have demonstrated their great capabilities in accomplishing various tasks. However, data, as the fuel behind those techniques, are distributed unevenly among different entities in practice. For instance, a large hospital in a metropolis probably maintain a larger medical record dataset than those in a rural area. Similarly, a general e-commerce established earlier can gather more customer preference information than a new business in some sub-categories. While data-rich entities can train high-performance machine learning models with their plentiful data, data-limited entities often struggle to obtain models with satisfactory performance due to their limited data. Although federated learning (FL) [28] enables cooperation among different entities with certain data protection guarantees, most data-rich entities have limited interest in such cooperation. An important reason is that the data-rich entities providing data in FL training (i.e., data sellers) cannot obtain satisfactory compensation.

Some existing works provide solutions for incentivizing data-rich parties in a data market setting [1, 8, 10, 27, 32, 40]. Generally, the setting assumes that some buyers want to obtain machine learning models to accomplish some tasks but do not have data to make such predictions. The seller can provide data and training services to produce models sold to those buyers. Within the same setting, the mechanisms of data markets fall into two categories.

The first category enables the data seller to produce multiple versions of a model via quantifiable randomness to control the model performance and lets buyers choose models with their budgets [8, 32, 41]. Nevertheless, these seller-determined prices only reflect the perceived value of data from the seller's perspective, which can pose certain limitations. Firstly, buyers are not guaranteed to truthfully provide information about their valuation for price decisions. Secondly, pricing from the seller's perspective fails to capture the varying values that different buyers place on the same set of data. Thirdly, the noisy model generation can be hard to control and may not be applicable to all models or all tasks.

The other category of data pricing is built on auction mechanisms [1] with the witness to the great success of online advertisement auctions [26]. The most promising aspect of auction mechanisms for pricing is that if the auction is truthful, it can incentivize buyers to submit their true value as bids because this strategy maximizes their profit. Another advantage of truthful auctions is that they provide explicit criteria for winner selection and payment decisions, leading to optimal outcomes, such as maximizing overall social benefit (i.e., social welfare) or maximizing the seller's profit.

Challenges. However, some unique properties of pricing FL as a service differentiate it from conventional auctions for other commodities or the data market. Auction mechanisms need to be redesigned to make them applicable for pricing FL.

Proceedings of the VLDB Endowment, Vol. 17, No. 6 ISSN 2150-8097.
doi:10.14778/3648160.3648169

C1. Difficulty in data valuation from buyers' perspectives. Accurately determining the value of the seller's data is infeasible in FL without first evaluating models trained with input from buyers. (i) This scenario mirrors the data market setting, where it is impractical to expect buyers to have prior knowledge of the true value of the seller's data, even when buyers are well-versed in tasks and potential profitability, because the value of the dataset directly depends on how much the seller's data can help improve the model, which is hard to estimate without actually training a model. Thus, auction mechanisms for other commodities may not be directly employed in this setting. (ii) However, the FL setting is also different from data markets. The sellers or brokers in the data markets can access all data for training and testing [1, 8, 32]. It implies that data valuation in the data market heavily relies on the honesty of sellers or brokers about the model performance. In contrast, sellers and buyers in the FL setting have their own data, which are *heterogeneous* in distribution (horizontal FL) or attributes (vertical FL) and *shall not be shared* directly with other entities. This implies that the model evaluation cannot rely solely on the seller, and such evaluation should be economical and trustworthy for all participants.

C2. Flexibility for varied constraints and desiderata in FL. The number of winners K out of N buyers is preset and decided by the available slots in online advertisement auctions, while the data market setting can have potentially unlimited winners with unlimited buyers. However, different datasets in FL may have different sharing constraints and regulations, even owned by the same organization. Pricing FL in some tasks can be between these two: there can be a *preset* or *variable* number of winners out of N buyers [1]. The limited buyers (and winners) setting is practical when the sold model has timeliness properties (the value of the model diminishes after a certain period) or needs to satisfy privacy constraints (e.g., the total privacy loss [15] because of the sold models needs to be constant). If a slightly changed constraint requires a very different auction mechanism, it may confuse the buyers and consume additional communication and legitimization effort.

C3. Requirement of computation and mechanism efficiency in FL. There is little concern about the computation and communication cost in the auction for traditional commodities. Meanwhile, different from the data market setting where sellers or brokers may access all data and can produce a large number of training trails, a pricing mechanism for FL is unwelcome if it introduces significant computation or communication overhead to the seller or buyers. The desired *computation efficiency* can be: when FL training is abstracted as an oracle, the number of times that the buyers have to call the oracle should be minimized. In terms of *mechanism efficiency*, it refers to the auction mechanism being truthful. Namely, the optimal strategy for all buyers should be to submit their true information (e.g., value) when bidding. This property eliminates the need for extra efforts to explore the optimal bidding strategy, thereby enhancing the overall efficiency of the process.

C4. Desiderata-preserving concerning the model performance differences between the test and production environment. Traditional auctions rely on buyers' direct valuation of the commodities; existing works on pricing data or models rely on test performance with observed labels. However, achieving good performance on a prepared test set cannot directly translate into profit for buyers because the profit of observed data is already realized. The performance in a production environment with unobserved data determines the true financial gain. As such, even if the valuation is based on test performance and certain desiderata are achieved, analyzing the sensitivity of these desiderata when transitioning to the production environment remains a challenge.

Contributions. Our proposed mechanisms in this paper are built on auctions. We consider a basic setting with only two kinds of entities: one data seller and multiple buyers. The seller can be a data holder or a union of data holders acting as multiple participants in an FL training task, with whom each buyer may produce an FL model potentially better than the buyer's locally trained model. The commodities in the auction are the accessibility and authorization to the FL model. This paper aims to overcome the aforementioned challenges, and our main contributions are as follows.

1) Problem formulation for performance-based pricing FL. We characterize the problem of model-performance-based pricing FL with realistic assumptions in Section 3. The assumptions are abstracted based on the existing understanding of FL model performance, privacy and intellectual property protection techniques. We define the terms necessary for performance-oriented auctions. Differences between the auctions in the FL context and the existing auctions are identified and discussed.

2) A flexible performance-based FL auction template. With the assumptions and unique characteristics of the FL auction, we propose a performance-based FL auction template based on our problem formulation and assumptions to price the FL model in Section 3. The template summarizes common information preparation steps for auction winner decisions and payment computation. The template is flexible in the sense that one can customize the allocation (winner decision) and payment decision sub-procedures to instantiate different truthful auction mechanisms for different desiderata, including maximizing social welfare (Section 4) or data seller's utility (Section 5). Meanwhile, our template also ensures computation efficiency, as only one FL training for each buyer is needed.

3) Truthful allocation-payment strategies for both limited and unlimited winners cases. We propose truthful mechanisms for pricing FL when either predefined K out of N buyers are granted the privilege to use the FL models (Section 4), or variable buyers can be granted (Section 5). With the efficiency and provable (approximate) truthfulness of our proposed mechanisms regarding both income-improvement rate and model improvement, our proposed auction designs resolve Challenge C1, C2 and C3.

4) Theoretical and empirical analysis on the impact of test-production performance gap to the desiderata. While the existing works related to pricing data solely rely on the model performance on the prepared test dataset with observed labels, we provide guarantees for the outcomes of auctions with production performance on unseen data. In Section 4 and Section 5, we provide theoretical analyses on how the uncertain test-production performance gap can affect some desiderata of the auction mechanisms, including the utility of buyers, social welfare, and seller's profit. Additionally, we conduct simulations based on two real-world datasets with synthetic bids to verify how our mechanisms perform in different desiderata. This examination is conducted concerning both test and production performance. These results provide theoretical and empirical guarantees for solving Challenge C4.

2 BACKGROUND

In this paper, we consider a single-round auction setting involving N buyers and one seller. Let v_i denote the true value of per unit commodity for buyer $i \in [N]$. For instance, in online keyword advertisement auctions, the commodities refer to the advertising slots, and the true value represents the expected profit that a buyer can attain when a user clicks on the advertisement. The true values v_i constitute *private* information known solely to buyer i, and are not available to the seller or any other buyers.

In the initial stages of auctions, buyer i submits a bid denoted as b_i to the seller. The bid b_i is not necessarily the same as v_i. We represent the vector of bids from all N buyers as $\mathbf{b}$. An auction mechanism $\mathcal{M}$ usually comprises two important functions: *allocation* and *payment*. The allocation function is a procedure in which the seller selects a subset of buyers to be the winners of the auction, denoted by $\mathcal{M}.\mathsf{Allocation}(\mathbf{b})$. These winners are charged with a specific monetary compensation when the commodities are handed over to them, a process referred to as *payment*, represented as $\mathcal{M}.\mathsf{Payment}_i(\mathbf{b})$ for buyer i. Each buyer can generate a profit or income relative to the true value and allocation results, represented as $\mathsf{Income}_i(v_i, \mathcal{M}.\mathsf{Allocation}(\mathbf{b}))$. The utility of buyer i is defined as the net income after payment:

$$U_i(v_i, \mathbf{b}, \mathcal{M}) = \mathsf{Income}_i(v_i, \mathcal{M}.\mathsf{Allocation}(\mathbf{b})) - \mathcal{M}.\mathsf{Payment}_i(\mathbf{b}).$$

The seller usually does not charge buyers who do not win any commodities in the auction, but losers also do not receive the commodities. In essence, if $\mathcal{M}.\mathsf{Allocation}_i(\mathbf{b}) = 0$, then both $\mathsf{Income}_i(v_i, \mathcal{M}.\mathsf{Allocation}(\mathbf{b}))$ and $\mathcal{M}.\mathsf{Payment}_i(\mathbf{b})$ are zero.

Two of the most prevalent objectives for auctions are maximizing *social welfare* and maximizing the *seller's profit*. Social welfare represents the total value of all entities created via the auction, $\mathsf{SocWel} = \sum_{i\in[N]} \mathsf{Income}_i(v_i, \mathcal{M}.\mathsf{Allocation}(\mathbf{b}))$. In contrast, the seller's profit is the sum of the buyers' payments, given by $\mathsf{SP} = \sum_{i\in[N]} \mathcal{M}.\mathsf{Payment}_i(\mathbf{b})$.

We use notations with parentheses in the subscript to represent the statistics ranks in non-increasing order. For instance, $b_{(i)}$ denotes the bid ranked at i. Moreover, we use $\mathbf{b}_{-i}$ to represent the vector of bids from all buyers excluding buyer i.

2.1 Truthfulness in Auctions

Truthfulness, also known as incentive compatibility, is among the most desirable properties in many mechanism design problems. This is because truthfulness ensures the efficiency of the mechanism, in the sense that buyers' utilities are maximized when they submit their true values as bids. The variants of truthfulness include universal truthfulness [14], truthfulness with high probability [12], and expectation-based truthfulness [3, 4]. This paper considers the following truthfulness definitions for both deterministic and stochastic mechanisms.

Definition 2.1 (Truthfulness). A deterministic auction is *truthful* if, for each buyer i and any fixed bid values for all other buyers, buyer i's utility is maximized by bidding her true utility value, i.e., $\forall\ \mathbf{b}_{-i},\ v_i \in \arg\max_{b_i} U_i(v_i, \{b_i, \mathbf{b}_{-i}\}, \mathcal{M})$. A mechanism is *truthful in expectation* if a buyer always maximizes her expected utility by bidding truthfully, i.e., $\forall\ \mathbf{b}_{-i},\ v_i \in \arg\max_{b_i} \mathbb{E}\left[\, U_i(v_i, \{b_i, \mathbf{b}_{-i}\}, \mathcal{M}) \,\right]$.

The truthfulness becomes incompatible with other properties in some scenarios. For example, [23] shows that no constant-competitive truthful auction is envy-free. Thus, a relaxed approximate truthfulness notion is also popular.

Definition 2.2 (ω-approximate truthfulness). With a smaller constant ω, the optimal utility in expectation of each buyer i by manipulating her bid is at most ω better than utility of submitting the true value as her bid, i.e., for all $\mathbf{b}_{-i}$,

$$\max_{b_i} \mathbb{E}\left[\, U_i(v_i, \{b_i, \mathbf{b}_{-i}\}, \mathcal{M}) \,\right] \le \mathbb{E}\left[\, U_i(v_i, \{v_i, \mathbf{b}_{-i}\}, \mathcal{M}) \,\right] + \omega$$

There is usually another co-occurring property, individual rationality, to protect the interest of participants in auctions.

Definition 2.3 (Individual Rationality (IR)). A mechanism is individually rational if the expected utility of each buyer is always non-negative, assuming this buyer reports truthfully.

2.2 Classic Auctions

Auctions for pricing FL service share some common properties with the existing auction problems. Thus, we briefly introduce the two most closely related auction games in this section. More details about the mechanism and discussions will be introduced in the context of FL auction in Section 4 and Section 5.

Auctions for search keywords advertisement. In the advertisement keyword auction (or sponsored search), there are two prevalent ranking methods in the auction mechanisms, direct ranking and (estimated) revenue ranking [2, 17], to assign K advertisement slots to buyers and charge them based on different axes.

Although direct ranking is straightforward and has been used in auctions with other commodities, its utility depends on the correlation between willingness to pay and a buyer's relevance to the search keywords [17]. To improve the robustness of this ranking system, revenue ranking sorts the buyers considering their click-through rate (CTR). Let $\mathsf{CTR}_{i,j}$ be the CTR of buyer i at the j-th slot and $\forall j > K, \mathsf{CTR}_{i,j} = 0$. Each buyer i has an associated weight w_i, which can be $w_i = \mathsf{CTR}_{i,1}$ in the simplified Google's implementation. The ranking score of buyer i is the product of the buyer's bid and her weight, namely $w_i b_i$. Buyers are then sorted in non-increasing order based on these ranking scores, and the top K buyers win the corresponding K slots.

Common auction payment rules building on these ranking scores are general second price (GSP), i.e., the buyer ranking at $k \le K$ pays the bid of ranking at $k+1$. However, GSP is not truthful when $K > 1$. Thus, Aggarwal et.al [2] propose a truthful mechanism while maintaining the revenue ranking order. Without loss of generality, we index the buyer according to their ranking score so that $w_i b_i \ge w_{i+1} b_{i+1}$. Its price-per-click rule is derived from Myerson's lemma [36] as $p_i = \sum_{j=i}^{K} \frac{\mathsf{CTR}_{i,j} - \mathsf{CTR}_{i,j+1}}{\mathsf{CTR}_{i,i}} \frac{w_{j+1}}{w_i} b_{j+1}$. Notice that the CTRs are considered as accurate and constant variables known to the seller before the auction.

Digitial goods auction. Another inspiring scenario is that a seller has an unlimited supply of a certain item with negligible duplication cost (e.g., digital copies of movies). In this setting, if such an auction aims to maximize social welfare, the seller should simply assign the item to all buyers at no cost. A more meaningful perspective is to maximize the seller's profit. If the valuations of all buyers are known to the seller, the problem becomes an optimization problem,

looking for the price to maximize the seller's profit, i.e., $\max_{p\in\mathcal{P}} p\times \sum_i \mathbf{1}[v_i \geq p]$. However, the valuation is private information of each buyer, so how to decide the payment is the key focus in this problem. Naively allocating to all buyers regardless of their bids is not the optimal profit for the seller because it can provide a strong incentive for buyers to manipulate bids to maximize their utility. To solve such a problem, two routines have been introduced, random sampling mechanisms [6] and random price selection mechanisms [22, 23].

- Random sampling mechanisms partition buyers into two groups randomly, search the best prices in both groups based on the bids, and apply the price of one group to another. These mechanisms can exclude the impact of a buyer's bid on her own payment, so that manipulating bids cannot bring any advantages to buyers.
- On the other hand, random price selection mechanisms, including the well-known exponential mechanism [35, 38] in differential privacy (DP), transform the price-decision process into stochastic. The probability of a price being selected is positively correlated to the seller's profit that it can bring. However, each buyer has a limited impact on this probability by changing her bids, so the incentive for untruthful reports is limited.

3 GENERAL PERFORMANCE-BASED FL AUCTION PARADIGM

The (approximate) truthful mechanisms for pricing the FL services and their theoretical analyses in this paper are based on the following settings and assumptions.

Auction setting. Similar to the auctions introduced in the previous section, this scenario involves N data-limited buyers and a single seller. Each buyer seeks to enhance her model performance on a specific task using an FL algorithm and the seller's additional data *individually*. As a result, if a buyer wins the auction, she will be granted a unique FL trained model. The true value v_i for a buyer can be interpreted as the rate of increase in income versus the improvement of buyer i's model on a specific metric, such as accuracy or the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). This rate for valuation is referred to hereafter as the *income-improvement rate*. The true rate is considered a constant private value to the buyer because buyers usually know the value of their model's per-unit improvement. For instance, based on her business history, a buyer has records of her average income if the model successfully identifies a new customer who wants to buy her products. Thus, if the final model improvement in the production environment is $\bar{\alpha}_i$ for buyer i, then buyer i's additional gain from the model is $\bar{\alpha}_i v_i$. Analogically, a bid $b_i > 0$ represents the monetary amount buyer i is willing to pay for each unit of the improvement.

Assumption 1. *All buyers agree on a common evaluation metric to make the auction a fair game. The buyers also have high confidence in their private true value v_i, and v_i is a constant in the auction process. As in other auction settings, buyers do not collude with each other.*

We summarize the notations used in this paper in Table 1 and compare these concepts in the context of online search keyword advertisement auctions and our new FL auctions.

FL setting. Our auction can be applied with either *horizontal FL* (HFL) or *vertical FL* (VFL). In the HFL, the seller helps a buyer improve model performance by joining FL and providing extra data with the same attributes but collected from different users. Symmetrically, the seller enhances a buyer's model by introducing extra attributes on the same set of user samples in VFL. Information exchange occurs only between the seller and each buyer, with no information sharing among the buyers. Notice that the seller in the auction can be either a single data holder or a union of multiple ones. The data holder(s) and the buyer participate in a FL course as clients. However, we make the following assumptions on the accessibility and cost of the FL model and the evaluation results.

Assumption 2. *Buyers are granted to use the trained FL models in production only if they are winners in the auction.*

Developing the FL algorithms satisfying the above assumptions is beyond the scope of this paper. However, it is important to note that these assumptions are practical. The accessibility assumptions can be fulfilled by solutions with security tools, including training tree-based models in VFL [48], linear models in HFL [53], or adopting the FL watermarking strategies [42] to ensure the model authorization.

- *Data market v.s. pricing for FL.* A significant difference is that the data market is usually assumed to have little constraints regarding data accessibility for training models [1, 8, 32]. The sellers or brokers in the data market can access all training data without concerns about computation and communication complexity or data privacy. Data market can conduct more intricate data selection and auction mechanisms, such as model versioning and combinatorial auction at per-sample or per-attribute levels, but need to assume the sellers or brokers are *honest about the model performance*. However, the data are owned partially by sellers and buyers in FL. They are *heterogeneous* in either distribution or attributes but cannot be *shared with others directly*. Thus, the model evaluation must be conducted by or involved buyers, and this process must economically support the truthfulness of the mechanism. Besides, while the data market setting focuses on maximizing the seller's profit exclusively with an unlimited number of buyers, pricing for FL may need to accommodate more diverse constraints and desiderata.
- *Classic auctions v.s. pricing for FL.* The existing digital goods auctions [6, 35, 38] do not consider performance-based bidding, and the buyers can value the commodities before the auction. Although both search keywords auctions [2] and our design for FL focus on performance-based, the accessibility of the FL models is uniform and independent of rankings among winners (i.e., each winner will be granted her unique FL models, which has only one version), while the advertisement slot awarded to winners in search keywords auctions differ by rankings (i.e., a higher spot has a higher CTR: $\mathsf{CTR}_{i,j} \geq \mathsf{CTR}_{i,j+1}$). This difference requires re-designing the auction mechanisms for FL to ensure truthfulness.

Performance assumption. In real-world applications, FL models are evaluated on test sets, which are subsets of the observed data before or during the auction. This evaluation introduces two problems. 1) A client's FL model performance can be no better or even poorer than the one of her locally trained model when the seller's data and the task do not match. 2) The profits or incomes from the models typically depend on the model performance realized on a much larger, unseen dataset (e.g., unexplored customers) in a production environment, which is unknown to all before an FL model is handed over to the winners. The model performances in the production environment will not be exactly the same as the ones on the test set and are unknown to all participants.

Table 1: Comparison between search keyword auction and FL auction. $\hat{W}$ denotes the set of winners in an auction.

Auction concepts	**Search keyword auction** (with [2])	**FL auction** (with $\mathcal{M} \in$ {K-FLA, EM, RSM})
Commodities	K advertisement slots $j \in [K]$	(Limited or unlimited) improved models after FL with seller
Performance	$\text{CTR}_{i,j}$: CTR of buyer i on j-th slot	$\hat{\alpha}_i$ / $\bar{\alpha}_i$: test/production performance improvement after FL
True value	v_i: income for buyer i given a click happen	v_i: income-improvement rate
Ranking score	$w_i b_i$, where w_i depends on $\text{CTR}_{i,1}$	$b_i \hat{\alpha}_i$
Seller's profit	(In expectation) $\sum_{i \in \hat{W}} p_i \text{CTR}_{i,\text{rank}(i)}$	$\sum_{i \in \hat{W}} \mathcal{M}.payment_i(\hat{\boldsymbol{\alpha}}, \mathbf{b})$
Buyer's utility	(In expectation) $\forall i \in \hat{W}, (v_i - p_i) \times \text{CTR}_{i,\text{rank}(i)}$	$\text{Income}_i(v_i, \bar{\alpha}_i, \mathcal{M}(\hat{\boldsymbol{\alpha}}, \mathbf{b})) - \mathcal{M}.\text{Payment}_i(\hat{\boldsymbol{\alpha}}, \mathbf{b})$
Uncertainty	Both seller's profit and buyer's utility are in expectation because of uncertainty of clicks.	Buyer's utility is uncertain because of the uncertainty of $\hat{\alpha}_i$ to $\bar{\alpha}_i$, seller's profit is decided in the auction.

In our performance-based FL auction, we name the performance improvement (subtracting the performance of the locally trained model in a metric from the one of the FL model) evaluated on test sets as *test improvement* and results in the production environment as *production improvement*. Notably, these two improvement evaluation results may have subtle differences. The test-production improvement gap may consist of multiple sources of randomness, including but not limited to the randomness of the model empirical error to its generalization error.

Because we can only use test improvements in the auction, but the actual values of models depend on production improvements, the randomness of test to production improvement can potentially influence the desired guarantees of our mechanisms. Consequently, it becomes essential to investigate the extent to which the truthfulness and other desiderata of our mechanism are sensitive to variations in the test-production improvement gap. Despite the complexity of the sources of the gap, we only make the following general assumption, regardless of the source of randomness.

Assumption 3. *Let $\hat{\alpha}$ be the model improvement evaluated on a test set, and $\bar{\alpha}$ be the improvement observed in the production environment. We assume that the evaluation on the test set is representative in the sense that $\hat{\alpha} = \bar{\alpha} + \eta$, where η is a zero-mean sub-Gaussian random noise with variance proxy parameter σ^2. Notice that $\hat{\alpha}$ and $\bar{\alpha}$ are not necessarily always positive.*

Under the assumption of zero-mean sub-Gaussian random noise, we can use the bounds on the tail as $\Pr[\eta < -\beta] \leq e^{-\beta/(2\sigma^2)}$ or $\Pr[\eta > \beta] \leq e^{-\beta^2/(2\sigma^2)}$. In practice, if the samples in the test set follow i.i.d. distribution as close as possible to the production environment, the σ may converge to 0 as the number of samples in the test set increases towards infinity. This assumption echoes many studies on the generalization error [7, 24, 49] where empirical error is assumed or demonstrated following sub-Gaussian distribution in different machine learning tasks. If $\hat{\alpha}$ or $\bar{\alpha} > 0$, the FL service helps the client positively. The non-positive possibility is introduced to cover the fact that collaboration in federated learning does not always guarantee performance improvement for any one of the participants in practice.

3.1 FL Auction Template

Considering the settings and assumptions outlined earlier, we propose a mechanism template for the FL auctions. The template guarantees the following properties: (1) valuation from the client's perspective and based on the performance improvement of FL models; (2) charging payments before the model is granted; (3) serving N clients in an auction and ensuring truthfulness and different desiderata with different allocation and payment methods for specific scenarios (e.g., a preset or a variable number of winners).

Mechanism 1 General performance-based FL auction template

1: **Bids submission:** All N buyers submit their bids $b_1, \ldots, b_N$, for the income-improvement rate.
2: **FL training & evaluation:** The seller conducts FL training with each of the N buyers. The improvements of these N models $\hat{\alpha}_1, \ldots, \hat{\alpha}_N$ are evaluated on test sets.
3: **Score computation & ranking:** The seller calculates and sorts with $\forall i \in [N], \hat{\phi}_i = b_i \times \hat{\alpha}_i$.
4: **Allocation & Payment imposition:** The seller applies $\mathcal{M}$.Allocation and $\mathcal{M}$.Payment sub-mechanisms to decide winners and their payments.

FL auction template. The template Mechanism 1 consists of four stages: 1) bids submission, 2) FL training and evaluation, 3) score computation and ranking and 4) allocation and payment imposing.

1) Bid submission as participation intention. In the first stage, buyers submit their bids indicating their interest in the auction. By design, this bid $b_i > 0$ represents the valuation of the income-improvement rate if the buyer i is truthful.

2) Training and evaluation before deciding the winners and payments. In the second stage, the seller conducts FL training with each buyer, but the buyers do not share information with each other. The seller follows the training protocols decided by the buyers, including the algorithms, initial models, and hyper-parameters, but does not provide guarantees that the FL models are better than the buyers' local models. After all models are trained and evaluated, the seller can obtain model improvement $\hat{\alpha}_i$ corresponding to each buyer $i \in [N]$ on a test set for the ranking. This information is used in allocation and payment imposing in the following stage. The seller conducts FL with all buyers at this stage because obtaining a relatively accurate estimate for performance improvement via FL without actual training is hard, because performance improvement highly depends on the correlation of the seller's and buyer's data and the suitability for the task.

3) Ranking score computation. In the third stage, a buyer's ranking score is computed as the product of her model improvements after

FL training and submitted bid. The intuition of the performance-based ranking score definition is that with a performance improvement on the test dataset, the seller can estimate the true value of the FL via this improvement and the buyer's truthful bid. If the buyer submits the bid truthfully, $b_i = v_i$, then the ranking score is an estimate of income the buyer can make with the FL model on the test set. This design of ranking score can reduce the risk to buyers of overestimating/underestimating the value of FL and encourage truthful bidding.

4) Allocation and payment imposition. The fourth stage, including the sub-mechanism of allocation strategies and payment strategies, are the key components to ensure the auction mechanism is truthful. We will introduce different mechanisms $\mathcal{M} \in \{\text{K-FLA}, \text{EM}, \text{RSM}\}$ in Section 4 and 5, depending on whether there is a restriction on the number of reveal models.

Remark 1. The auction involves one seller and N buyers. However, the seller is a concept for auction. It is not necessarily restricted to a single data holder in FL. Our mechanism can extend to more general FL tasks in two different aspects. (i) The seller is a proxy of multiple data holders with a consensus of maximizing their union's desiderata. The payment received from buyers can be further split to data holders by metrics like Shapley value [1, 40]. (ii) Our mechanism can be invoked repeatedly with multiple sellers sequentially. For example, a buyer can participate in multiple auctions selling data from different data holders. This approach can be considered as a greedy approximation to the combinatorial auction [34].

Adapted notations. To make the following discussion more concise and precise, we overwrite some of the notations as the following. Because the allocation and payment of a mechanism depend on all buyers' model test or production performance improvements and bids, the allocation and payment results based on test improvement are denoted as $\mathcal{M}.\text{Allocation}_i(\hat{\boldsymbol{\alpha}}, \mathbf{b})$ and $\mathcal{M}.\text{Payment}_i(\hat{\boldsymbol{\alpha}}, \mathbf{b})$. Both functions return 0 if buyer i loses the auction or the auction is aborted, but the allocation function returns 1 if wins and the payment will be a non-negative real number. With the above notations, the real incomes are based on the *production* improvement and auction allocation according to the *test* improvement:

$$\text{Income}_i(v_i, \bar{\boldsymbol{\alpha}}, \mathcal{M}(\hat{\boldsymbol{\alpha}}, \mathbf{b})) = (v_i \bar{\alpha}_i) \cdot \mathcal{M}.\text{Allocation}_i(\hat{\boldsymbol{\alpha}}, \mathbf{b}),$$

but $\text{Income}_i(v_i, \hat{\boldsymbol{\alpha}}, \mathcal{M}(\hat{\boldsymbol{\alpha}}, \mathbf{b}))$ can be understood as an estimate of the buyer's income based on the test improvement. Following the above definition, buyer i's utility is denoted as

$$U_i(v_i, \cdot, \mathcal{M}(\hat{\boldsymbol{\alpha}}, \mathbf{b})) = \text{Income}_i(v_i, \cdot, \mathcal{M}(\hat{\boldsymbol{\alpha}}, \mathbf{b})) - \mathcal{M}.\text{Payment}_i(\hat{\boldsymbol{\alpha}}, \mathbf{b}),$$

where the placeholder $\cdot$ can be either $\bar{\boldsymbol{\alpha}}$ or $\hat{\boldsymbol{\alpha}}$ for production/test utility analysis. Similarly, social welfare is denoted as $\text{SocWel} = \sum_{i\in[N]} \text{Income}_i(v_i, \bar{\boldsymbol{\alpha}}, \mathcal{M}(\hat{\boldsymbol{\alpha}}, \mathbf{b}))$ and seller's profit $\text{SP} = \sum_{i\in[N]} \mathcal{M}.\text{Payment}_i(\hat{\boldsymbol{\alpha}}, \mathbf{b})$. In some analysis, we may switch between $\hat{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\alpha}}$ in the above notions.

We use $\widehat{\text{OPT}}(\cdot)$ to denote the optimal of a variable (e.g., social welfare) on the test improvement, but assuming all private values are known; we denote $\overline{\text{OPT}}(\cdot)$ as the optimal of a variable assuming both production improvements and private true values are known.

4 TRUTHFUL AUCTION RELEASING MODELS TO LIMITED WINNERS

We first consider the scenario where the seller in the FL auction is constrained to only grant trained models to K out of N buyers in an auction. Limiting the granted models can be due to different reasons, including privacy. For instance, if differential privacy [15] is considered as the privacy notion. Each revealed model consumes a portion of the predefined privacy budget for a dataset. The privacy budget for a model cannot be too small; otherwise, the valuable data information will be dominated by the added noise in training, and the model's utility is hard to guarantee.

While following the first three stages in Mechanism 1, the Allocation and Payment methods must be adapted as Sub-mechanism 2 to ensure the mechanism is truthful.

Sub-mechanism 2 K Winner FL Auction (K-FLA)

Sort the buyers according to $\hat{\phi}_i$. If $\hat{\phi}_{(K+1)} \le 0$, abort the acution.
Allocation rule: select at th top K buyers in the ranking according the ranking scores as winners : $\hat{W} = \{i | \hat{\phi}_i > \hat{\phi}_{(K+1)}\}$.
Payment rule: each winner i pays $\text{Payment}_i = \hat{\phi}_{(K+1)}$.

Theorem 4.1. *Sub-mechanism 2 is truthful and IR for buyer i given any $(\hat{\boldsymbol{\alpha}}_{-i}, \mathbf{b}_{-i})$ regarding the test datasets, i.e.,*

$$(\hat{\alpha}_i, v_i) \in \arg\max_{\alpha_i, b_i} U_i(v_i, \hat{\boldsymbol{\alpha}}, \textit{K-FLA}(\{\alpha_i, \hat{\boldsymbol{\alpha}}_{-i}\}, \{b_i, \mathbf{b}_{-i}\}))$$

and $U_i(v_i, \hat{\boldsymbol{\alpha}}, \textit{K-FLA}(\{\hat{\alpha}_i, \hat{\boldsymbol{\alpha}}_{-i}\}, \{v_i, \mathbf{b}_{-i}\}))) \ge 0$.

Proof. To prove IR with test improvement, we can see that any winner in $\hat{W}$ has $\hat{\phi}_i \ge \text{Payment}_i$ by the allocation rule. Thus, the incomes of the winners regarding the test improvements $\hat{\boldsymbol{\alpha}}$ are always at least as much as their payment.

To prove truthfulness of $(\hat{\alpha}_i, v_i)$, we want to show that for any buyer c, she cannot benefit more by manipulating her ranking score $\hat{\phi}_i = \hat{\alpha}_i v_i$. When $\hat{\phi}_{(K+1)} > 0$, assume all other buyers' bids and test improvements are fixed arbitrarily except for the ones of buyer c. Suppose buyer c ranks at r_c if she submits her bid truthfully, $b_c = v_c$ and $\alpha_c = \hat{\alpha}_c$. By manipulating buyer c's bid $b_c \ne v_c$ and/or test improvement $\alpha_c \ne \hat{\alpha}_c$, buyer c can change her rank to r. (1) When $r > K$ and $r_c > K$, or $r \le K$ and $r_c \le K$, the test income of buyer c does not change. (2) When $r > K \ge r_c$, buyer c could have won the auction but it does not because of manipulating the rank. Recall that winning the auction always brings non-negative income. So the utility of ranking at r_c can always be no less than the ranking at r. (3) When $r \le K < r_c$, it means buyer c wins the auction, but the payment outweighs its true income on test improvement, i.e., $\alpha_c b_c \ge \hat{\phi}_{(K+1)} \ge v_c \hat{\alpha}_c$. When $\hat{\phi}_{(K+1)} \le 0$, increasing any $\hat{\phi}_i$ can either keep the $\hat{\phi}_{(K+1)} \le 0$ or $\hat{\phi}_{(K+1)} > 0$. The former does not change the utilities of all clients, while the latter can introduce negative utility to the buyer manipulating the score. □

The bids b_i or test improvements $\hat{\alpha}_i$ hold truthfulness guarantees alone as an implication from Theorem 4.1.

Corollary 4.2. *Sub-mechanism 2 also ensures truthfulness for buyer i's bid and test improvement given any $(\hat{\boldsymbol{\alpha}}_{-i}, \mathbf{b}_{-i})$, i.e.,*

$$v_i \in \arg\max_{b_i} U_i(v_i, \hat{\boldsymbol{\alpha}}, \textit{K-FLA}(\{\hat{\alpha}_i, \hat{\boldsymbol{\alpha}}_{-i}\}, \{b_i, \mathbf{b}_{-i}\})),$$

$\hat{\alpha}_i \in \arg\max_{\alpha_i} U_i(v_i, \hat{\boldsymbol{\alpha}}, \textit{K-FLA}(\{\alpha_i, \hat{\boldsymbol{\alpha}}_{-i}\}, \{v_i, \mathbf{b}_{-i}\}))$.

Remark 2. Notice that $\hat{\alpha}_i$ is not known yet when buyer i submits her bid b_i, and the payment $\hat{\phi}_{(K+1)}$ is independent of $\hat{\phi}_i$ if buyer i wins the auction. Therefore, if the buyer i cheats on either her bid b_i or her test performance improvement α_i, the best result for her is *spending extra computation costs but obtaining the same utility* regarding her actual test improvement.

4.1 Impact of the Difference Between Test and Production Improvements

In the industrial machine learning model operational management (MLOps), a standard setting is that the evaluation in the production environment can only be revealed days or even weeks after the model is deployed, after which the production improvement can be calculated. Since there are unavoidable gaps between the evaluations on the test set used in auctions and the ones in the production environment, one may be interested in how sensitive FL auctions are to this random gap regarding truthfulness and social welfare. We provide analyses for the client utility and social welfare when the auction is not aborted.

Impact on the client utility. The optimal utility of buyer i is $U_i^* = \max\{\bar{\alpha}_i v_i - \bar{\phi}_{(K+1)}, 0\}$ if production improvements are known. However, a buyer's utility can be negative without knowing the exact $\bar{\alpha}_i$ as in practice.

Theorem 4.3. *If Assumption 3 holds, the mechanism K-FLA, when it is not aborted, ensures that with probability at least* $1-\delta$*, the utility concerning production improvements can be bounded as*

$$|U_i(v_i, \hat{\boldsymbol{\alpha}}, \textit{K-FLA}(\hat{\boldsymbol{\alpha}}, \mathbf{b})) - U_i(v_i, \bar{\boldsymbol{\alpha}}, \textit{K-FLA}(\hat{\boldsymbol{\alpha}}, \mathbf{b}))| \le v_i \sigma_i \sqrt{2\ln(1/\delta)}.$$

Proof. The cases in which truthful bids are not optimal for the buyer are either $\hat{\alpha}_i v_i > \hat{\phi}_{(K+1)} > \bar{\alpha}_i v_i$ (real income is less than the payment) or $\hat{\alpha}_i v_i < \hat{\phi}_{(K+1)} < \bar{\alpha}_i v_i$ (real income is larger than the required payment, but the buyer loses the auction). We consider the case $\hat{\alpha}_i v_i > \hat{\phi}_{(K+1)} > \bar{\alpha}_i v_i$, while the other one is symmetric. When Assumption 3 holds, the probability of $v_i(\hat{\alpha}_i - \bar{\alpha}_i) \ge \beta$ by the property of the sub-Gaussian randomness is $\Pr[v_i(\hat{\alpha}_i - \bar{\alpha}_i) \ge \beta] \le e^{-\beta^2/2v_i^2\sigma_i^2}$. Therefore, with probability $1-\delta$, the buyer cannot gain more than $v_i\sigma_i\sqrt{2\ln(1/\delta)}$. Following the same idea, with probability $1-\delta$, the buyer's utility will be at least $-v_i\sigma_i\sqrt{2\ln(1/\delta)}$. □

Impact on social welfare. The differences between test and production improvements can lead to a consequence that the winners selected with test improvements (i.e., Allocation$(\hat{\boldsymbol{\alpha}}, \mathbf{b})$) are different from those that should have been selected if production improvements were known (i.e., Allocation$(\bar{\boldsymbol{\alpha}}, \mathbf{b})$). The differences in winners will lead to differences in social welfare.

Let $\bar{W}$ be the set of top K buyers selected via K-FLA.Allocation$(\bar{\boldsymbol{\alpha}}, \mathbf{b})$, and $\hat{W}$ as the top K buyers by K-FLA.Allocation$(\hat{\boldsymbol{\alpha}}, \mathbf{b})$. We demonstrate the stability of social welfare by bounding $\sum_{i\in\bar{W}} v_i\bar{\alpha}_i - \sum_{j\in\hat{W}} v_j\bar{\alpha}_j$. For simplicity, we follow Assumption 3 and further assume that all η_i have the same σ^2 and $\forall i \in [N], v_i = 1$. But our results can be easily extended to different buyers with different σ_i^2 and true values.

We start by analyzing a special case where there is only one winner $K = 1$, then generalize the result for arbitrary K.

Lemma 4.4. *When there is only one winner* ($K = 1$) *and Assumption 3 holds, with probability at least* $1 - \frac{1}{N}$*, the selected buyer's social welfare* $v_\ell\bar{\alpha}_\ell$ *for some* ℓ *is at most* $\beta < 4\sigma\sqrt{\ln N}$ *worse than the optimal social welfare* $\max\{v_i\bar{\alpha}_i | i \in [N]\}$.

Proof. By Assumption 3, $\mathbb{E}[v_i\hat{\alpha}_i] = v_i\bar{\alpha}_i$, without loss of generality, we further index the buyers so that $i < j$, then $v_i\bar{\alpha}_i \ge v_j\bar{\alpha}_j$.

To help the proof, we define a function as $\gamma(i,\beta) = \min\{j | v_i\bar{\alpha}_i - v_j\bar{\alpha}_j \ge \beta\}$, which returns the smallest index of those buyers with scores at least β smaller than buyer i's. By assuming $K = 1$, the problem can be reduced to a sub-problem bounding the probability $\Pr\left[\gamma(i,\beta) = \arg\max_{j\in[N]} v_j\hat{\alpha}_j\right]$.

$$\begin{aligned}\Pr\left[\gamma(i,\beta) = \arg\max_{j\in[N]} v_j\hat{\alpha}_j\right] &= \Pr\left[\cap_{x\in[N], x\neq\gamma(i,\beta)} v_x\hat{\alpha}_x \le v_{\gamma(i,\beta)}\hat{\alpha}_{\gamma(i,\beta)}\right] \\ &\le \Pr\left[v_1\hat{\alpha}_1 \le v_{\gamma(i,\beta)}\hat{\alpha}_{\gamma(i,\beta)}\right]\end{aligned}$$

By the tail bound of sub-Gaussian and the union bound for all buyers with indices larger than $\gamma(1,\beta)$,

$$\begin{aligned}\Pr[v_1\bar{\alpha}_1 - v_\ell\bar{\alpha}_\ell > \beta] &\le (N - \gamma(1,\beta))\Pr\left[v_1\hat{\alpha}_1 \le v_{\gamma(i,\beta)}\hat{\alpha}_{\gamma(i,\beta)}\right] \\ &\le (N-\gamma(1,\beta))e^{-\frac{\beta^2}{8\sigma^2}} \le Ne^{-\frac{\beta^2}{8\sigma^2}}\end{aligned}$$

The second inequality comes from the probability bound of the random variable $v_1\hat{\alpha}_1 - v_{\gamma(i,\beta)}\hat{\alpha}_{\gamma(i,\beta)}$ smaller than 0 with variance proxy 2σ. By setting $\beta = 4\sigma\sqrt{\ln N}$, the probability is at most $\frac{1}{N}$. □

More general case: $K > 1$. When the number of winners is set to be larger than 1, we can prove the social welfare of the output buyers is close to the one of the optimal set of buyers with high probability.

Theorem 4.5. *When there are K winners, two sets of winners* $\bar{W} =$ *K-FLA.Allocation*$(\bar{\boldsymbol{\alpha}}, \mathbf{b})$ *and* $\hat{W} =$ *K-FLA.Allocation*$(\hat{\boldsymbol{\alpha}}, \mathbf{b})$*, based on production and test improvements respectively, introduce difference in social welfare in production environment* $\sum_{i\in\bar{W}} v_i\bar{\alpha}_i - \sum_{j\in\hat{W}} v_j\bar{\alpha}_j \le 2K\sigma\sqrt{2\ln NK(N-K)}$ *with probability at least* $1 - \frac{1}{N}$.

Proof. The key idea of the proof inherit the proof of Lemma 4.4.

$$\Pr\left[\sum_{i\in\bar{W}} v_i\bar{\alpha}_i - \sum_{j\in\hat{W}} v_j\bar{\alpha}_j > \beta\right] \le \Pr\left[\exists i \in \bar{W}, j \in \hat{W}, v_i\bar{\alpha}_i - v_j\bar{\alpha}_j \ge \frac{\beta}{K}\right]$$

$$\le K(N-K)\Pr\left[i \in \bar{W}, v_i\hat{\alpha}_i < v_{\gamma(i,\frac{\beta}{K})}\hat{\alpha}_{\gamma(i,\frac{\beta}{K})}\right] \le K(N-K)e^{-\frac{\beta^2}{8K^2\sigma^2}}$$

The first inequality comes from the union bound and the second comes from a similar idea in the proof of Lemma 4.4. By setting $\beta = 2K\sigma\sqrt{2\ln NK(N-K)}$, the above probability is bounded by $\frac{1}{N}$ □

5 OPTIMIZED SELLER'S PROFIT WITH VARIABLE WINNERS

In the previous section, we discuss how to decide the winners when the number of opportunities for granted models, K, is predefined and fixed. However, there may be scenarios where the seller is primarily interested in maximizing its revenue without restrictions on the number of released models. Notice that revealing FL models to all participants is not always the most profitable strategy because no payment method can ensure truthfulness and maximize the seller's profit with such an allocation strategy. Before discussing

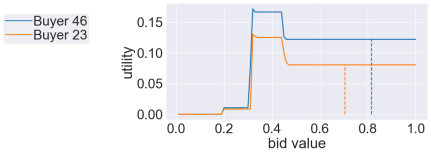


Figure 1: The solid lines are utilities based on test improvement $U_i(v_i, \hat{\boldsymbol{\alpha}}, \text{Naive}(\hat{\boldsymbol{\alpha}}, \{b_i, \mathbf{v}_{-i}\}))$, the dashed lines are the true values of buyers. Untruthful bids can increase buyers' utilities in Naive.

the solutions, we first formulate the goal of this setting as the optimal price selection problem:

$$\max_{p\in\mathcal{P}} p \times \sum_i \mathbf{1}\left[\hat{\alpha}_i b_i \geq p\right] \tag{1}$$

A strawman mechanism: Naive. Assuming all buyers truthfully report their bids, a Naive mechanism can maximize the seller's profit by following steps. Naive.Payment outputs the price by solving Equation (1) to get the optimal price $\hat{p}_{\text{Naive}}$ and Naive.Allocation selects buyers with $\hat{\alpha}_i b_i \geq \hat{p}_{\text{Naive}}$ as winners.

However, the truthfulness assumption on buyers is unrealistic. For example, Figure 1 shows the relations between utility and bids assuming a buyer can change her bid arbitrarily while the other buyer's bids are fixed as true values[1]. The vertical dashed lines indicate their utility regarding test improvement if submitting true values and test improvements. The Buyer 46 and 23 are the two buyers with the highest ranking scores if they bid truthfully. But they can further improve their utility by submitting a bid lower than their true value, which violates the definition of truthfulness.

Thus, to design truthful auctions (approximately) solving Equation (1), we need to borrow the wisdom from the traditional *digital goods auction*, e.g., exponential mechanism [35] and random sampling mechanism [6]. Following the previous literature[35], we assume that the range of legitimate bids is predefined as $[b_{\min}, b_{\max}]$; the improvement on a metric is in the range $[\alpha_{\min}, \alpha_{\max}]$ for both the test and production performance improvement, which can be obtained by definition of metric used in the auction. As the candidate prices are defined by the seller, the price (or payment for buyers) is in the range $[\max\{b_{\min}\alpha_{\min}, p_{\min}\}, b_{\max}\alpha_{\max}]$, where $p_{\min}$ is the lowest price the seller can accept.

5.1 Approximate Truthfulness via Exponential Mechanism

Following the digital goods auction routine, the auction process is described in Sub-mechanism 3. The selection of the price becomes stochastic. The key idea of EM is that the price introducing a higher seller's profit is more likely to be selected. But any buyer can only have a limited impact on changing the probability of a price p being selected. Formally, the probability EM outputs $\hat{p}$ as the price is

$$\Pr\left[\text{EM.Payment}^{\mathcal{P},\epsilon}(\hat{\boldsymbol{\alpha}}, \mathbf{b}) = \hat{p}\right] = \frac{\exp\left(\frac{\epsilon u(\hat{\boldsymbol{\alpha}}, \mathbf{b}, \hat{p})}{2\Delta}\right)}{\sum_{p\in\mathcal{P}} \exp\left(\frac{\epsilon u(\hat{\boldsymbol{\alpha}}, \mathbf{b}, p)}{2\Delta}\right)}, \tag{2}$$

where $\Delta = b_{\max}\alpha_{\max} - \max\{b_{\min}\alpha_{\min}, p_{\min}\}$ and $u(\hat{\boldsymbol{\alpha}}, \mathbf{b}, p) = p\sum_{i\in[N]} \mathbf{1}[\hat{\alpha}_i b_i \geq p]$ as EM's utility functions for the price p.

Notice that we follow the [35] assuming the seller can impose the winners to pay for the model. It differs from [38], where buyers can rescind to pay and give up being allocated. Namely, we assume EM itself is with imposition. The approximate truthful regarding the ranking scores $\hat{\phi}_i = \hat{\alpha}_i b_i$ of EM can be directly derived from [35].

[1]Details about the experiment settings are deferred to Section 6.

Sub-mechanism 3 Exponential mechanism (EM)

Given a set of candidate prices $\mathcal{P} = \{p_0, \ldots, p_{|\mathcal{P}|-1}\}$:
Seller samples a $\hat{p} \in \mathcal{P}$ with probability as Equation (2).
Allocation rule: release models to the winners $\hat{W} = \{i | \hat{\alpha}_i b_i > \hat{p}\}$.
Payment rule: charge the winner with price $\hat{p}$.

Theorem 5.1 (From [35, 38]). *The* EM *is* $(e^\epsilon - 1)\Delta$*-approximate truthful, namely*

$$\max_{\alpha_i, b_i} \mathbb{E}\left[U_i(v_i, \hat{\boldsymbol{\alpha}}, \text{EM}^{\mathcal{P},\epsilon}(\{\alpha_i, \hat{\boldsymbol{\alpha}}_{-i}\}, \{b_i, \mathbf{b}_{-i}\}))\right]$$
$$\leq \mathbb{E}\left[U_i(v_i, \hat{\boldsymbol{\alpha}}, \text{EM}^{\mathcal{P},\epsilon}(\hat{\boldsymbol{\alpha}}, \{v_i, \mathbf{b}_{-i}\}))\right] + (e^\epsilon - 1)\Delta$$

As stated in [38], the above result is meaningful when $\epsilon \in (0, 1]$ and $\Delta \leq 1$, so $e^\epsilon - 1 \leq 2\epsilon$ and EM becomes $2\epsilon\Delta$-approximate truthful. Similar to Corollary 4.2, the truthfulness can be extended to either the bids and the test improvements alone. On the other hand, it has been shown that EM can also provide prices that is close to the optimal (regarding the test improvement $\widehat{\text{OPT}}(\text{SP})$) with high probability. The proof follows [35, 38].

Theorem 5.2 (Adapted from [35]). *With probability at least* $1 - \delta$*, the* EM *can guarantee its seller's profit* SP^{EM}

$$\widehat{\text{OPT}}(\text{SP}) - \text{SP}^{\text{EM}} \leq \frac{2\Delta}{\epsilon}\ln(|\mathcal{P}|/\delta)$$

Difference between SP^{EM} **and** $\overline{\text{OPT}}(\text{SP})$**.** $\overline{\text{OPT}}(\text{SP})$ is the optimal seller's profit if the production improvements and true values of all buyers are known, i.e., $\overline{\text{OPT}}(\text{SP}) = \max_{p\in\mathcal{P}} p\sum_i \mathbf{1}[v_i\tilde{\alpha}_i \geq p]$. Similarly, the optimal seller profit based on test improvements and known true values is $\widehat{\text{OPT}}(\text{SP}) = \max_{p\in\mathcal{P}} p\sum_i \mathbf{1}[v_i\hat{\alpha}_i \geq p]$. Since the desiderata in this setting is to maximize the seller's profit, a desirable mechanism should guarantee that SP_{EM} is not significantly worse than $\overline{\text{OPT}}(\text{SP})$ with high probability.

As we assume there are lower bounds for both bids and model improvements, we set the smallest candidate price as p_0. For convenience, we denote $\bar{c}_j = \sum_i \mathbf{1}[b_i\tilde{\alpha}_i \geq p_j]$ and $\hat{c}_j = \sum_i \mathbf{1}[b_i\hat{\alpha}_i \geq p_j]$ as the counts for the number of buyers whose production/test ranking scores are greater than the j-th price (sorted in increasing order) in the candidate set.

We set the set of candidate prices as $\mathcal{P} = \{p_0, p_1, \ldots, p_{|\mathcal{P}|-1}\}$, such that either of the following statements holds: Ⓐ $p_j = (1+\rho)p_{j-1}$ or Ⓑ $p_j = p_{j-1} + \tau$, where $\rho > 0$ and $\tau > 0$. Then we can have the following lemma.

Lemma 5.3. *Assuming* $\forall i \in [N], b_i = 1$ *and* $\sigma_i = \sigma$ *with Assumption 3. With probability* $1 - \delta$*,*

$$\overline{\text{OPT}}(\text{SP}) - \widehat{\text{OPT}}(\text{SP}) \leq \max\{\sigma N\sqrt{2\ln N/\delta}, \lambda N\},$$

where $\lambda = \rho p_0$ *with* Ⓐ *or* $\lambda = \tau$ *with* Ⓑ*.*

Proof. Our proof consists of the following possible cases.

Case 1: If $p_0 = \arg\max_{p\in\mathcal{P}} p\sum_i \mathbf{1}[b_i\bar{\alpha}_i \geq p]$, then $\overline{\mathrm{OPT}}(\mathrm{SP}) = p_0 N$. The profit based on test improvements $\widehat{\mathrm{OPT}}(\mathrm{SP})$ cannot be worse than $p_0 N$ with Ⓐ or Ⓑ.

Case 2: If $p_1 = \arg\max_{p\in\mathcal{P}} p\sum_i \mathbf{1}[b_i\bar{\alpha}_i \geq p]$, then $\overline{\mathrm{OPT}}(\mathrm{SP}) = p_1\bar{c}_1$. In contrast, $\widehat{\mathrm{OPT}}(\mathrm{SP}) = \max_{p\in\mathcal{P}} p\sum_i \mathbf{1}[b_i\hat{\alpha}_i \geq p] \geq p_0 N$.

Thus, setting Ⓐ leads to $\overline{\mathrm{OPT}}(\mathrm{SP}) - \widehat{\mathrm{OPT}}(\mathrm{SP}) \leq p_1\bar{c}_1 - p_0 N \leq \rho p_0 N$; setting Ⓑ leads to $\overline{\mathrm{OPT}}(\mathrm{SP}) - \widehat{\mathrm{OPT}}(\mathrm{SP}) \leq p_1\bar{c}_1 - p_0 N \leq \tau N$.

Case 3: If $p_x = \arg\max_{p\in\mathcal{P}} p\sum_i \mathbf{1}[b_i\bar{\alpha}_i \geq p]$ for some $x \geq 2$. An observation is that if $\overline{\mathrm{OPT}}(\mathrm{SP}) - \widehat{\mathrm{OPT}}(\mathrm{SP}) = p_x\bar{c}_x - p_{x-1}\hat{c}_{x-1} > (p_x - p_{x-1})\bar{c}_x$, then it must hold that $\left\{i|\bar{\phi}_i \geq p_x\right\} \cap \left\{i|\hat{\phi}_i \leq p_{x-1}\right\} \neq \emptyset$. More general, if $\overline{\mathrm{OPT}}(\mathrm{SP}) - \widehat{\mathrm{OPT}}(\mathrm{SP}) = p_x\bar{c}_x - p_{x-1}\hat{c}_{x-j} > (p_x - p_{x-j})\bar{c}_x$, then there must be $\left\{i|\bar{\phi}_i \geq p_x\right\} \cap \left\{i|\hat{\phi}_i \leq p_{x-j}\right\} \neq \emptyset$. Therefore, the following inequality holds with union bound:

$$\Pr\left[\overline{\mathrm{OPT}}(\mathrm{SP}) - \widehat{\mathrm{OPT}}(\mathrm{SP}) \geq (p_x - p_{x-j})\bar{c}_x\right]$$

$$\leq \Pr\left[\exists i, \bar{\phi}_i - \hat{\phi}_i \geq p_x - p_{x-j}\right] \leq \bar{c}_x e^{-\frac{(p_x - p_{x-j})^2}{2\sigma^2}}$$

Let $\beta = (p_x - p_{x-j})\bar{c}_x$, then

$$\Pr\left[\overline{\mathrm{OPT}}(\mathrm{SP}) - \widehat{\mathrm{OPT}}(\mathrm{SP}) \geq \beta\right] \leq \bar{c}_x e^{-\frac{\beta^2}{2\sigma^2\bar{c}_x^2}} \leq N e^{-\frac{\beta^2}{2\sigma^2 N^2}} = \delta$$

Substituting β with δ, we have $\beta = \sigma N\sqrt{2\ln(N/\delta)}$ as the result. □

With the above result, we can conclude the seller's profit guarantee of EM on the production improvement.

Theorem 5.4. *With probability $1-\delta$, the price selected by* EM *mechanism can guarantee the seller's profit with respect to production improvement is at least*

$$\overline{\mathrm{OPT}}(\mathrm{SP}) - (2\Delta/\epsilon)\ln(2|\mathcal{P}|/\delta) - \sigma N\sqrt{2\ln(2N/\delta)}$$

This can be proved by combining Theorem 5.2 and Lemma 5.3.

Discussion of randomness of test improvement. The upper bound of the profit gap is $O(N\sqrt{\ln N})$ and it has a lower bound of at least $\Theta(N)$ because of Case 2 in our analysis. But notice that $N \leq \frac{\overline{\mathrm{OPT}}}{p_0}$, which indicates the gap can be small if σ/p_0 is small. Besides, such a worst case happens when all buyers' ranking scores are very close to each other and close to the prices in $\mathcal{P}$. In practice, the randomness of test improvement can be alleviated if either 1) the ranking scores are distributed "separably enough" in the domain or 2) the test improvements are close to production improvement.

5.2 Truthful Dual Price via Random Sampling Mechanism

If the environment does not have the constraint that all winners must enjoy the same price but wants strictly truthful, we can borrow the results from [6] about the truthfulness and utility guarantee, as shown in Sub-mechanism 4. Because the winners are decided by the ranking scores, the truthfulness can be extended to the income-improvement rate and the test improvement as Corollary 4.2.

With RSM, a buyer can never affect the price in her own group. However, she will be forced to pay if her ranking score is higher than the threshold in her group, so reporting a higher bid/improvement increases the risk of paying more than the best utility. On the other hand, it will increase the probability of losing the auction she could have won if her bid/improvement is lower than the true value. Thus, RSM is *truthful* on the test improvement. Using the utility guarantee proved in Theorem 6 of [6] and our Lemma 5.3, we can give the guarantee on the seller's profit brought by $\mathrm{RSM}(\hat{\boldsymbol{\alpha}}, \mathbf{b})$.

Sub-mechanism 4 Random sampling mechanism (RSM)

Given a set of candidate prices $\mathcal{P} = \{p_0, \ldots, p_{|\mathcal{P}|-1}\}$:
Randomly partition users into two subsets S_1 and S_2.
Obtain $\hat{p}_1 = \arg\max_{p\in\mathcal{P}} p\sum_{i\in S_1}\mathbf{1}[\hat{\phi}_i \geq p]$ and $\hat{p}_2 = \arg\max_{p\in\mathcal{P}} p\sum_{i\in S_2}\mathbf{1}[\hat{\phi}_i \geq p]$;
Allocation rule: release models to the winners in $W = W_1 \cup W_2$ where $W_1 = \{i \in S_1 | \hat{\phi}_i > \hat{p}_2\}$ and $W_2 = \{i \in S_2 | \hat{\phi}_i > \hat{p}_1\}$;
Payment rule: charge the W_1 with $\hat{p}_2$ and W_2 with $\hat{p}_1$.

Corollary 5.5. *Under the same assumptions as Lemma 5.3 and setting the candidate price set as Ⓐ, with probability $1-\delta$, using* RSM *can obtain profit at least*

$$\overline{\mathrm{OPT}}(\mathrm{SP}) - 8\sqrt{\overline{\mathrm{OPT}}(\mathrm{SP})\log(2/\rho\delta)} - \sigma N\sqrt{2\ln(2N/\delta)}.$$

6 EXPERIMENTS

To better understand the desiderata and properties of mechanisms discussed in the previous section, we have developed simulations based on semi-real-world datasets. [2]

Datasets. Due to the relatively early stage of pricing for FL research and the confidential commercial information typically involved, there are no public datasets available for this task. Therefore, we simulate real-world scenarios by combining real-world datasets with synthetic true values of buyers. For the FL datasets, we employ the public Criteo dataset [13] and a private dataset related to CTR prediction tasks (referred to as "Business")[3]. As is common in many mechanism design studies, we assume that both the true values and bids fall within the domain $(0, 1)$. Therefore, we employ a method of sampling values from a uniform distribution in $(0, 1)$.

• *Train, test and production data simulation.* We split the datasets into training, testing and production sets. The training subset, denoted as (X^{tr}, Y^{tr}), is used for model training, while the testing subset, denoted as (X^{te}, Y^{te}), is employed to evaluate the test performance of FL model for auctions. The remaining is used to simulate production performance, represented as (X^{prod}, Y^{prod}). The train/test/production size ratio is 5:2:3. The sizes of production sets are larger than the test sets because the models are expected to serve a more extensive range of users in the production environment.

• *Vertical federated learning simulation.* In our simulation, we emulate a vertical federated learning setting, where the data seller possesses the datasets of half of the attributes, denoted as X_s^{tr}, X_s^{te}, and X_s^{prod}. Each buyer's local dataset comprises 5 attributes randomly selected from the remaining attribute set, along with the label. These are represented as (X_i^{tr}, Y_i^{tr}), (X_i^{te}, Y_i^{te}), and (X_i^{prod}, Y_i^{prod}). The training datasets are respectively assigned to the data seller and the buyers. We assume all samples are aligned between buyers' and the data seller's datasets by record ids.

[2]The simulation code and a full manuscript with more results in appendix can be found on https://github.com/ZiTao-Li/fl_auction.

[3]The private data is not available to the public at this stage.

• *Buyers' true values simulation.* In our experiments, we consider the number of buyers in the auctions as $N = 100$ or 500. For each buyer, we sample her private true value from a uniform distribution in the range $(0, 1)$. These values are fixed and retained across all experiments conducted with the same settings.
More details about the datasets used can be found in Table 2.

FL training and evaluation simulation. For our simulations, we use the logistic regression (LR) model from the Scikit-learn package [39]. For Criteo dataset, we preprocess the dataset so that all the categorical attributes are transformed to one-hot encoding. Since our interest lies primarily in verifying the properties of our auction mechanisms, we simplify the process as follows. For each buyer $i \in [N]$, we train an LR model exclusively with local data (denoted as $\mathcal{F}^{\text{LC}}$) and obtain the local performance on the test set and production set as $\hat{g}_i^{\text{LC}} = \text{Eval}(\mathcal{F}^{\text{LC}}, X_i^{te}, Y_i^{te})$ and $\bar{g}_i^{\text{LC}} = \text{Eval}(\mathcal{F}^{\text{LC}}, X_i^{prod}, Y_i^{prod})$, respectively. Subsequently, we merge the data seller's attributes with the buyer's dataset and train a different model (denoted as $\mathcal{F}^{\text{FL}}$). This setup is often considered a baseline in VFL literature. We evaluate these pseudo-VFL models of buyers on the testing set and on the production set, yielding performances as $\hat{g}_i^{\text{FL}} = \text{Eval}(\mathcal{F}^{\text{LC}}, [X_i^{te}, X_s^{te}], Y_i^{te})$ and $\bar{g}_i^{\text{FL}} = \text{Eval}(\mathcal{F}^{\text{LC}}, [X_i^{prod}, X_s^{prod}], Y_i^{G})$.

We then compute the model improvements as $\hat{\alpha}_i = \hat{g}_i^{\text{FL}} - \hat{g}_i^{\text{LC}}$ and $\bar{\alpha}_i = \bar{g}_i^{\text{FL}} - \bar{g}_i^{\text{LC}}$. As the CTR prediction task often has unbalanced label distribution, we employ the AUC-ROC at the metric. As we observe, both $\hat{\alpha}_i$ and $\bar{\alpha}_i \in (0, 0.35)$ in all experiments. Thus, we set $\alpha_{\min} = 0$ and $\alpha_{\max} = 0.35$ for all experiments. We replicate training with each buyers 20 times with different train/test splits but keep the production set the same to simulate the inherent randomness comparing test improvements to production improvements.

Efficiency of the auction mechanisms. Since the end-to-end efficiency of the auction depends on many factors, including the federated learning algorithms and frameworks, we focus on the efficiency of the auctions when all the evaluation results are ready. Our experiments are conducted on a server with Intel(R) Xeon(R) Platinum 8163 CPU @ 2.50GHz CPUs. The K-FLA is very efficient in the sense that the auction can be finished in 0.15/1.14 ms on average with 100/500 buyers, as its computation complexity is dominated by sorting ($O(N \log N)$) to identify the top K winners. The EM takes 0.18/0.26 ms per execution with 100/500 buyers and is dominated by the profit calculation ($O(|\mathcal{P}|N)$) and sampling function, while every execution of RSM elapses 0.17/0.23 ms. Notice that the complexity is based on our straightforward implementation and can be improved by advanced algorithms.

6.1 Truthfulness and Buyers' Utilities

To validate the truthfulness, we conduct simulations following these steps. We initially fix all buyers' ranking scores/bids/test improvements equal to their true ranking scores/values/test improvements but vary those variables of the Buyer i within the domains. If truthfulness to test improvement holds, we can observe that there would be no other potential dishonest report that could provide Buyer i with a strictly higher utility than simply reporting the truth. Results are shown in Figure 2 and 3. The solid lines are utilities with test improvement $U_i(v_i, \hat{\boldsymbol{\alpha}}, \mathcal{M}(\{\alpha_i, \hat{\boldsymbol{\alpha}}_{-i}\}, \{b_i, \mathbf{v}_{-i}\}))$, and the dotted line is $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathcal{M}(\{\alpha_i, \bar{\boldsymbol{\alpha}}_{-i}\}, \{b_i, \mathbf{v}_{-i}\}))$, where the Buyer i can modify α_i, b_i or both to affect her ranking scores. The scatter points are $U_i(v_i, \overline{\boldsymbol{\alpha}}, \text{K-FLA}(\bar{\boldsymbol{\alpha}}, \mathbf{v}))$ in Figure 2 and $U_i(v_i, \overline{\boldsymbol{\alpha}}, \text{Naive}(\bar{\boldsymbol{\alpha}}, \mathbf{v}))$ in Figure 3. The numbers in the parentheses in the legend are the ranking of the buyers with $\overline{\alpha}_i v_i$. The shadow of the line means the range of utility based on test improvement on different splits.

Truthfulness and buyer utility of K-FLA. In Figure 2, we report the utility of the top-2 winners (e.g., Buyer 90 and 91 in Criteo) based on the true value and *production* improvements, as well as for those buyers situated around the decision boundaries (10 or 20). Notice that manipulating the ranking score is equivalent to manipulating the bid and test performance together or separately.

As observed, the utility of the top winners increases with a rise in K values, pertinent to both test and production improvement (e.g., Figure 2(a) to Figure 2(d)). This can be attributed to the decrease in payments $\text{Payment}_i = \hat{\phi}_{(K+1)}$ to maintain truthfulness, as the number of winners K increases.

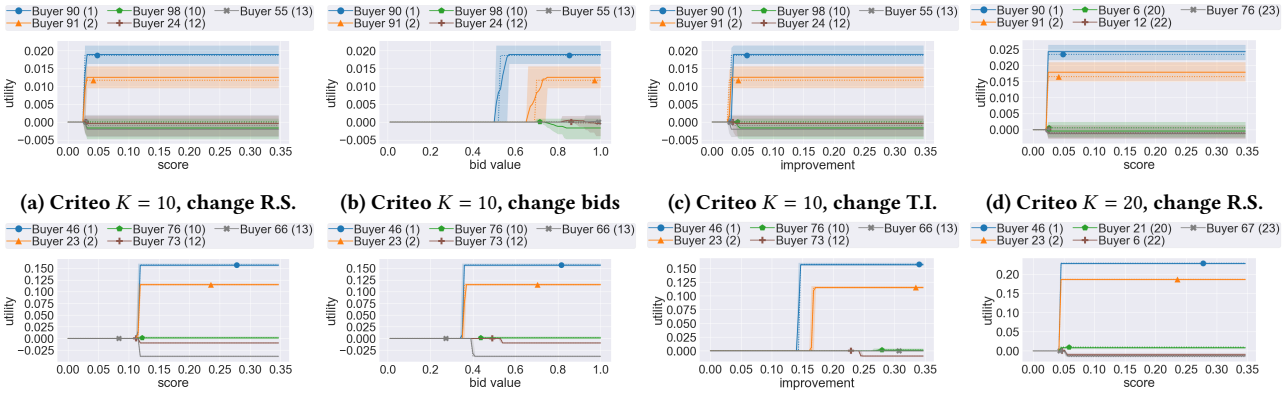
Through our experiments, we observe that those who should have won the auction if bidding and reporting their test improvement truthfully, indeed maximize their profit when they actually bid and report truthfully. Submitting a lower score/bid/improvement can often result in the buyer losing the auction. For example, if Buyer 90 manipulates her ranking score to 0.03 (Figure 2(a)), or submits a lower bid around 0.5 (Figure 2(b)), or similarly lowers her test improvement to 0.03 (Figure 2(c)), the risk of obtaining sub-optimal utility (i.e., losing the auction) significantly increases. Conversely, for those buyers who should not have won the auction, bidding higher than their truth could lead to negative utility. An example is Buyer 66 in Figure 2(e)-2(g), where she obtains 0 utility when truthfully reports but experiences negative utility when exaggerating her score/bid/test improvement to be a winner.

Upon comparing (averaged) utilities based on test improvement ($\hat{\boldsymbol{\alpha}}$) and production improvements ($\bar{\boldsymbol{\alpha}}$), the gaps between them are tiny. It means the utilities based on test improvement can be reliably good estimates of the outcomes when the production improvements are revealed and final true values are calculated. Notably, higher-ranking buyers tend to suffer less from this disparity. However, for buyers with ranking scores around the thresholds, there exists a risk of experiencing negative utility.

Truthfulness and Buyer utility of EM. The shadows in Figure 3 mean the range of 0.25 and 0.75 quantiles, representing a divergence from the EM results. Upon comparing with the fixed-K setting, it is evident that the variance of utility primarily arises from the randomness of EM. According to our results, untruthful ranking score, bids or improvement does not offer substantial benefits on test/production utility to buyers in expectation with EM. It is because the payment decision process is designed with inherent randomness, which means a single buyer is unlikely to influence the probability of a particular price being selected without significantly deviating from the truth. However, significantly deviating from the truth can only bring large losses to the buyers. For example, untruthful ranking scores and test improvements (Figure 3(a) and 3(c)) can lead to large negative utility because these two variables are significantly smaller than 0.35. The utilities based on test and production improvements are very close, because the randomness of EM will dominate in this setting.

Table 2: Dataset information.

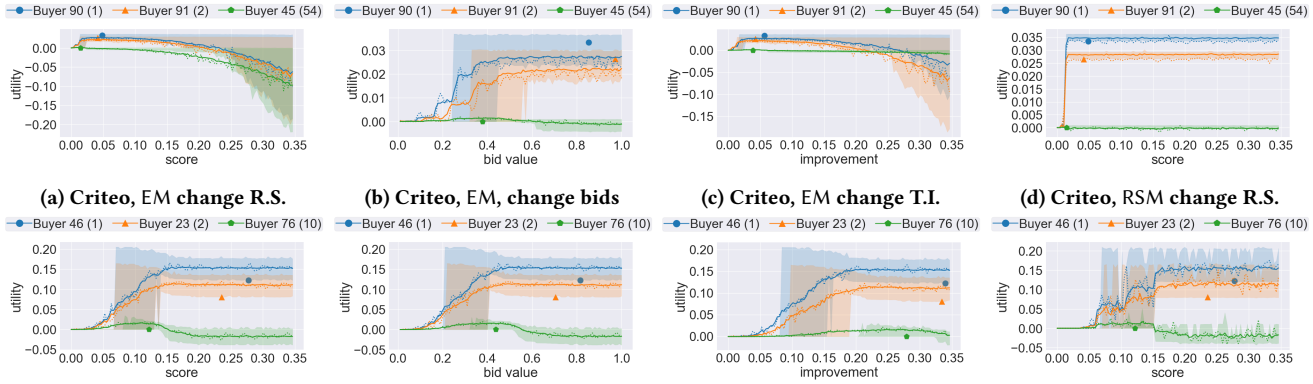
Dataset	# seller's attributes	# buyer's attributes	Training size	Testing size	production simulation size
Criteo	19	5 out of 20	378277	151310	226967
Business	26	5 out of 27	1042442	446762	638230



(a) Criteo $K = 10$, change R.S. **(b) Criteo $K = 10$, change bids** **(c) Criteo $K = 10$, change T.I.** **(d) Criteo $K = 20$, change R.S.**

(e) Business $K = 10$, change R.S. **(f) Business $K = 10$, change bids** **(g) Business $K = 10$, change T.I.** **(h) Business 1 $K = 20$, change R.S.**

Figure 2: Utility of buyers with K-FLA and different ranking scores (R.S.) / bids / test improvements (T.I.).



(a) Criteo, EM **change R.S.** **(b) Criteo,** EM**, change bids** **(c) Criteo,** EM **change T.I.** **(d) Criteo,** RSM **change R.S.**

(e) Business, EM **change R.S.** **(f) Business,** EM **change bids** **(g) Business,** EM **change T.I.** **(h) Business,** RSM **change R.S.**

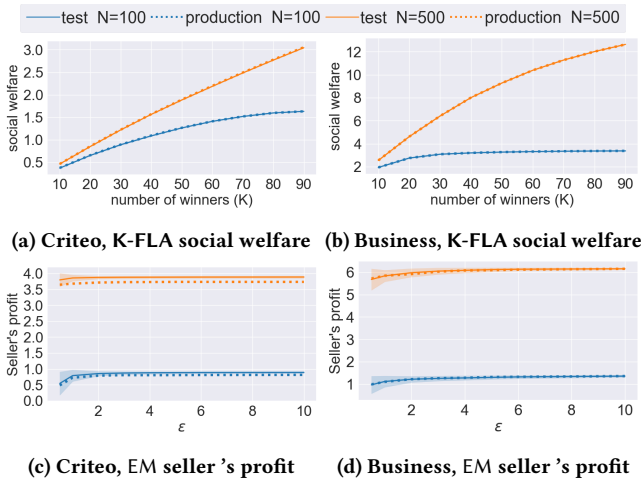
Figure 3: Utility with $\mathcal{M} \in \{\mathsf{EM}(\epsilon = 2), \mathsf{RSM}\}$ and different ranking scores (R.S.) / bids / test improvements (T.I.).

Notice that a phenomenon different from K-FLA experiments that the marker points, $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathrm{Naive}(\bar{\boldsymbol{\alpha}}, \mathbf{v}))$, may be higher or lower to the line of $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathsf{EM}^{\mathcal{P},\epsilon}(\bar{\boldsymbol{\alpha}}, \{b_i, \mathbf{v}_{-i}\}))$. This is because EM is not unbiased on the selected price. The price from $\mathrm{Naive}(\bar{\boldsymbol{\alpha}}, \mathbf{v})$ may be different from $\mathsf{EM}^{\mathcal{P},\epsilon}(\bar{\boldsymbol{\alpha}}, \mathbf{v})$ in expectation. But as ϵ increases, EM has higher probability selecting the optimal price same as $\mathrm{Naive}(\bar{\boldsymbol{\alpha}}, \mathbf{v})$, and thus, $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathsf{EM}^{\mathcal{P},\epsilon}(\bar{\boldsymbol{\alpha}}, \{b_i, \mathbf{v}_{-i}\}))$ becomes closer to $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathrm{Naive}(\bar{\boldsymbol{\alpha}}, \mathbf{v}))$.

Truthfulness of RSM**.** RSM shows similar phenomena as for EM in Figure 3(d) and 3(h), but no matter how a buyer manipulates her ranking score, the expected utilities will not surpass the one when she truthfully submits the private value as her bid. However, we can see that the utility may depend on the randomness of the test improvements. For the Criteo dataset, the variances of the utilities are very small; however, for Business, RSM shows a larger variance. Also, there will not be large negative utilities on the Criteo dataset because the prices applied to Buyer i are generated from another group, which consists of benign buyers. Other properties can be similar to the results of EM, including the relation between $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathrm{Naive}(\bar{\boldsymbol{\alpha}}, \mathbf{v}))$ and $U_i(v_i, \bar{\boldsymbol{\alpha}}, \mathsf{RSM}(\bar{\boldsymbol{\alpha}}, \{b_i, \mathbf{v}_{-i}\}))$.

6.2 Stability with Respect to Production Data

Stability of social welfare with K-FLA. Figure 4 compares the results of $\sum_{i \in \bar{W}} v_i \bar{\alpha}_i$ and $\sum_{j \in \hat{W}} v_j \bar{\alpha}_j$ with both 100 and 500 buyers. As we can see from the figure, the $\sum_{i \in \bar{W}} v_i \bar{\alpha}_i$ is always very close to $\sum_{j \in \hat{W}} v_j \bar{\alpha}_j$. There are two reasons. The first is that the test performances are usually very close to production performances in our experiments, making $\hat{\phi}_i \approx \bar{\phi}_i$ (as shown in Figure 2). A second



(a) Criteo, K-FLA social welfare **(b) Business, K-FLA social welfare**

(c) Criteo, EM seller 's profit **(d) Business, EM seller 's profit**

Figure 4: Comparing of social welfare and seller 's profit

reason is that there are many buyers with ranking scores very close to the $\hat{\phi}_{(K+1)}$ or $\bar{\phi}_{(K+1)}$ (as shown in Figure 2(a) and 2(e)). Any of those buyers selected does not introduce huge differences between $\hat{\phi}_{(K+1)}$ and $\bar{\phi}_{(K+1)}$. Thus, our theorem in Section 4 can serve as the worst-case guarantee. Besides, the social welfare with 500 buyers are larger than the ones of 100 buyers. This is because in the simulation, the top K randomly sampled values from 500 samples are expected to be larger than the top K from 100 samples.

Sensitivity of seller's profit via EM. As demonstrated in Figure 4(c) and 4(d), the seller's profit can exceed or fall short of the value based on production improvement. The reason is that the randomness of EM and the payments are calculated according to test improvement, which can either outperform or under-perform compared to production improvement. However, the expectation of the seller's profit based on both test and production improvement remains relatively small. Unsurprisingly, as the parameter ϵ increases, both the seller's profit based on test improvement and production improvement rise. The reason is that the optimal price stands a higher chance of being selected when ϵ increases. The difference between the seller's profit of 500 and 100 buyers can be explained by more buyers leading to more high-valued FL models, which eventually lead to more winners even with the same price.

7 RELATED WORK

There is a comprehensive survey on pricing the data [40].

Data market. Pricing the data is closely related to pricing federated learning. A vision paper [18] talks about the desirable properties of data market platforms, from system design to mechanism properties. Agarwal et al. [1] proposes a framework for the FL cooperation marketplace, from designing allocation function to control model quality to payment decision and division functions. Their setting assumes the buyers come one by one. Their revenue maximization property depends on a pricing function based on multi-armed bandit optimization. However, their revenue function and pricing update function need to explore *all* possible bids lower than the current one, which can be computationally expensive. There are other work building data marketplace platforms, but focusing more on using differential privacy as a knob to control the data quality and the price [21, 31, 32, 41]. Besides, some other studies consider Shapley value to evaluate data in central setting [20, 29] and in FL setting [33, 46]. Those research results are orthogonal to our auctions because Sharply value is a metric of profit division, where the total revenue is known, and the focus is on how to allocate the revenue fairly; the auctions are designed for pricing objectively. In the decentralized setting, [47] proposes an incentive-aware mechanism based on reinforcement learning and allows participants to achieve Nash equilibrium with their reward functions. However, the prices (costs) of pulling parameters are decided without a competitive pricing mechanism. [50] focuses on how to prevent the data seller cheating by replicating data.

Auction in FL. Compared to our approach, most of the existing work on FL focuses on auction mechanisms for selecting buyers/devices to participate in FL [9, 11, 30, 44, 51]. In this case, the server announces the learning task, and buyers submit their prices, which represent their cost of conducting the task. But these problems are formulated as *reverse auction problems*. In this setting, truthfulness is defined as buyers being unable to increase their utility by bidding higher than their learning *cost*. The existing works vary in their objective of allocation and payment design. One representative work is FAIR [11], which formulates the objective as maximizing the learning quality of each task subject to (1) the total payment being within budget and (2) the rewards/payment to each buyer being greater than the bids. The authors demonstrate the truthfulness of the mechanism using Myerson's theorem [36]. In [30], the authors propose a Vickrey-Clarke-Groves (VCG) auction mechanism with the aim of identifying the most suitable buyers and determining the appropriate payment. The mechanism was subsequently refined in [9] by prioritizing the maximization of social welfare and minimizing the imbalances within the federation.

Auction in Advertisement. GSP and VCG auctions are well-known auction mechanisms that have been extensively researched and implemented in diverse advertising systems [5, 16, 19, 37, 45, 52]. The most related paper is [2], in which the authors discuss the truthfulness of the auction in advertisement with (weighted) bids. In [45], the authors discuss the attractive properties and potential drawbacks of VCG ad auctions. They argue that in VCG ad auction has the attractive property that bidding the true value is a dominant strategy for all players. In [43], the authors consider GSP variants that are revenue-equivalent to the truthful equilibrium of corresponding dominant-strategy mechanisms with the same allocation rules. Later, a series of GPS variants that aims to optimize multiple objectives in ads auctions and meanwhile ensure truthfulness are proposed, including [5, 19, 52] In more general auction settings, some existing results consider guaranteeing truthfulness via the exponential mechanism of DP [25, 35, 38].

8 CONCLUSIONS

In this paper, we provide solutions for pricing FL via auction mechanisms and account for performance improvement. We propose a template and present different instantiated auctions for different constraints and desiderata, with theoretical guarantee and empirical simulation supports. To bring pricing FL closer to practice, some problems can be further explored, such as adapting computationally expensive auctions (e.g., combinatorial auctions) in FL.

REFERENCES

[1] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 701–726.

[2] Gagan Aggarwal, Ashish Goel, and Rajeev Motwani. 2006. Truthful auctions for pricing search keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce*. 1–7.

[3] Aaron Archer, Christos Papadimitriou, Kunal Talwar, and Éva Tardos. 2004. An approximate truthful mechanism for combinatorial auctions with single parameter agents. *Internet Mathematics* 1, 2 (2004), 129–150.

[4] Aaron Archer and Éva Tardos. 2001. Truthful mechanisms for one-parameter agents. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*. IEEE, 482–491.

[5] Yoram Bachrach, Sofia Ceppi, Ian A Kash, Peter Key, and David Kurokawa. 2014. Optimising trade-offs among stakeholders in ad auctions. In *Proceedings of the fifteenth ACM conference on Economics and computation*. 75–92.

[6] M-F Balcan, Avrim Blum, Jason D Hartline, and Yishay Mansour. 2005. Mechanism design via machine learning. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*. IEEE, 605–614.

[7] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. 2020. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory* 1, 1 (2020), 121–130.

[8] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. 2019. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the 2019 International Conference on Management of Data*. 1535–1552.

[9] Mingshu Cong, Han Yu, Xi Weng, Jiabao Qu, Yang Liu, and Siu Ming Yiu. 2020. A VCG-based fair incentive mechanism for federated learning. *arXiv preprint arXiv:2008.06680* (2020).

[10] Zicun Cong, Xuan Luo, Jian Pei, Feida Zhu, and Yong Zhang. 2022. Data pricing in machine learning pipelines. *Knowledge and Information Systems* 64, 6 (2022), 1417–1455.

[11] Yongheng Deng, Feng Lyu, Ju Ren, Yi-Chao Chen, Peng Yang, Yuezhi Zhou, and Yaoxue Zhang. 2021. FAIR: Quality-Aware Federated Learning with Precise User Incentive and Model Aggregation. In *40th IEEE Conference on Computer Communications, INFOCOM 2021, Vancouver, BC, Canada, May 10-13, 2021*. IEEE, 1–10.

[12] Nikhil R Devanur and Sham M Kakade. 2009. The price of truthfulness for pay-per-click auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*. 99–106.

[13] Diemert Eustache, Meynet Julien, Pierre Galland, and Damien Lefortier. 2017. Attribution Modeling Increases Efficiency of Bidding in Display Advertising. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, Halifax, NS, Canada, August, 14, 2017*. ACM, To appear.

[14] Shahar Dobzinski and Shaddin Dughmi. 2013. On the power of randomization in algorithmic mechanism design. *SIAM J. Comput.* 42, 6 (2013), 2287–2304.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[16] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review* 97, 1 (2007), 242–259.

[17] Juan Feng, Hemant K Bhargava, and David M Pennock. 2007. Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms. *INFORMS Journal on Computing* 19, 1 (2007), 137–148.

[18] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2013. Data Market Platforms: Trading Data Assets to Solve Data Problems. *Proceedings of the VLDB Endowment* 13, 11 (2013).

[19] Sahin Cem Geyik, Sergey Faleev, Jianqiang Shen, Sean O'Donnell, and Santanu Kolay. 2016. Joint optimization of multiple performance metrics in online video advertising. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 471–480.

[20] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*. PMLR, 2242–2251.

[21] Arpita Ghosh and Aaron Roth. 2011. Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce*. 199–208.

[22] Andrew V Goldberg and Jason D Hartline. 2003. Competitiveness via consensus.. In *SODA*, Vol. 3. 215–222.

[23] Andrew V Goldberg and Jason D Hartline. 2003. Envy-free auctions for digital goods. In *Proceedings of the 4th ACM conference on Electronic commerce*. 29–35.

[24] Jean Honorio and Tommi Jaakkola. 2014. A unified framework for consistency of regularized loss minimizers. In *International Conference on Machine Learning*. PMLR, 136–144.

[25] Zhiyi Huang and Sampath Kannan. 2012. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, 140–149.

[26] Bernard J Jansen and Tracy Mullen. 2008. Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business* 6, 2 (2008), 114–131.

[27] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. 2019. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619* (2019).

[28] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.

[29] Yongchan Kwon, Manuel A Rivas, and James Zou. 2021. Efficient computation and analysis of distributional shapley values. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 793–801.

[30] Tra Huong Thi Le, Nguyen Hoang Tran, Yan Kyaw Tun, Zhu Han, and Choong Seon Hong. 2020. Auction based Incentive Design for Efficient Federated Learning in Cellular Wireless Networks. In *2020 IEEE Wireless Communications and Networking Conference, WCNC 2020, Seoul, Korea (South), May 25-28, 2020*. IEEE, 1–6.

[31] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. 2014. A theory of pricing private data. *ACM Transactions on Database Systems (TODS)* 39, 4 (2014), 1–28.

[32] Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. 2021. Dealer: an end-to-end model marketplace with differential privacy. *Proceedings of the VLDB Endowment* 14, 6 (2021).

[33] Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. 2022. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 4 (2022), 1–21.

[34] Brendan Lucier and Allan Borodin. 2010. Price of anarchy for greedy auctions. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 537–553.

[35] Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 94–103.

[36] Roger B Myerson. 1981. Optimal auction design. *Mathematics of operations research* 6, 1 (1981), 58–73.

[37] Noam Nisan and Amir Ronen. 2007. Computationally feasible VCG mechanisms. *Journal of Artificial Intelligence Research* 29 (2007), 19–47.

[38] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. 2012. Approximately optimal mechanism design via differential privacy. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 203–213.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[40] Jian Pei. 2020. A survey on data pricing: from economics to data science. *IEEE Transactions on knowledge and Data Engineering* 34, 10 (2020), 4586–4608.

[41] Peng Sun, Xu Chen, Guocheng Liao, and Jianwei Huang. 2022. A profit-maximizing model marketplace with differentially private federated learning. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1439–1448.

[42] Buse G. A. Tekgul, Yuxi Xia, Samuel Marchal, and N. Asokan. 2021. WAFFLE: Watermarking in Federated Learning. In *40th International Symposium on Reliable Distributed Systems, SRDS 2021, Chicago, IL, USA, September 20-23, 2021*. IEEE, 310–320.

[43] David R. M. Thompson and Kevin Leyton-Brown. 2013. Revenue optimization in the generalized second-price auction. In *Proceedings of the fourteenth ACM Conference on Electronic Commerce, EC 2013, Philadelphia, PA, USA, June 16-20, 2013*, Michael J. Kearns, R. Preston McAfee, and Éva Tardos (Eds.). ACM, 837–852.

[44] Xuezhen Tu, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Yang Zhang, and Juan Li. 2022. Incentive Mechanisms for Federated Learning: From Economic and Game Theoretic Perspective. *IEEE Trans. Cogn. Commun. Netw.* 8, 3 (2022), 1566–1593.

[45] Hal R Varian and Christopher Harris. 2014. The VCG auction in theory and practice. *American Economic Review* 104, 5 (2014), 442–445.

[46] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. 2020. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive* (2020), 153–167.

[47] Yatong Wang, Yuncheng Wu, Xincheng Chen, Gang Feng, and Beng Chin Ooi. 2023. Incentive-Aware Decentralized Data Collaboration. *Proceedings of the ACM on Management of Data* 1, 2 (2023), 1–27.

[48] Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, and Beng Chin Ooi. 2020. Privacy Preserving Vertical Federated Learning for Tree-based Models. *Proceedings of the VLDB Endowment* 13, 11 (2020).

[49] Aolin Xu and Maxim Raginsky. 2017. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems* 30 (2017).

[50] Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2021. Validation free and replication robust volume-based data valuation. *Advances in Neural Information Processing Systems* 34 (2021), 10837–10848.

[51] Rongfei Zeng, Shixun Zhang, Jiaqi Wang, and Xiaowen Chu. 2020. FMore: An Incentive Scheme of Multi-dimensional Auction for Federated Learning in MEC. In *40th IEEE International Conference on Distributed Computing Systems, ICDCS 2020, Singapore, November 29 - December 1, 2020*. IEEE, 278–288.

[52] Zhilin Zhang, Xiangyu Liu, Zhenzhe Zheng, Chenrui Zhang, Miao Xu, Junwei Pan, Chuan Yu, Fan Wu, Jian Xu, and Kun Gai. 2021. Optimizing multiple performance metrics with deep GSP auctions for e-commerce advertising. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 993–1001.

[53] Wenting Zheng, Raluca Ada Popa, Joseph E Gonzalez, and Ion Stoica. 2019. Helen: Maliciously secure coopetitive learning for linear models. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 724–738.