Incentives for Federated Learning: a Hypothesis Elicitation Approach

Yang Liu [1] **Jiaheng Wei** [1]

Abstract

Federated learning provides a promising paradigm for collecting machine learning models from distributed data sources without compromising users' data privacy. The success of a credible federated learning system builds on the assumption that the decentralized and self-interested users will be willing to participate to contribute their local models in a trustworthy way. However, without proper incentives, users might simply opt out the contribution cycle, or will be mis-incentivized to contribute spam/false information. This paper introduces solutions to incentivize truthful reporting of a local, user-side machine learning model for federated learning. Our results build on the literature of information elicitation, but focus on the questions of *eliciting hypothesis* (rather than eliciting human predictions). We provide a scoring rule based framework that incentivizes truthful reporting of local hypotheses at a Bayesian Nash Equilibrium. We study the market implementation, accuracy as well as robustness properties of our proposed solution too. We verify the effectiveness of our methods using MNIST and CIFAR-10 datasets. Particularly we show that by reporting low-quality hypotheses, users will receive decreasing scores (rewards, or payments).

1. Introduction

When a company relies on distributed users' data to train a machine learning model, federated learning (McMahan et al., 2016; Yang et al., 2019; Kairouz et al., 2019) promotes the idea that users/customers' data should be kept local, and only the locally held/learned hypothesis will be shared/contributed from each user. While federated learning has observed success in keyboard recognition (Hard et al., 2018) and in language modeling (Chen et al., 2019), existing works have made an implicit assumption that participating users will be willing to contribute their local hypotheses to help the central entity to refine the model. Nonetheless, without proper incentives, agents can choose to opt out of the participation, to contribute either uninformative or outdated information, or to even contribute malicious model information. Though being an important question for federated learning (Yang et al., 2019; Liu et al., 2020a; Yu et al., 2020; Yu et al., 2020), this capability of providing adequate incentives for user participation has largely been overlooked. In this paper we ask the questions that: *Can a machine learning hypothesis be incentivized/elicited by a certain form of scoring rules from self-interested agents?* The availability of a scoring rule will help us design a payment for the elicited hypothesis properly to motivate the reporting of high-quality ones. The corresponding solutions complement the literature of federated learning by offering a generic template for incentivizing users' participation.

We address the challenge via providing a scoring framework to elicit hypotheses truthfully from the self-interested agents/users[1]. More concretely, suppose an agent i has a locally observed hypothesis f_i^*. For instance, the hypothesis can come from solving a local problem: $f_i^* = \text{argmin}_{f_i \sim \mathcal{H}_i} \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell_i\left(f_i(X), Y\right)]$ according to a certain hypothesis class $\mathcal{H}_i$, a distribution $\mathcal{D}$, a loss function ℓ_i. The goal is to design a scoring function $S(\cdot)$ that takes a reported hypothesis f_i, and possibly a second input argument (to be defined in the context) such that $\mathbb{E}\left[S(f_i^*, \cdot)\right] \geq \mathbb{E}\left[S(f_i, \cdot)\right], \forall f_i$, where the expectation is w.r.t. agent i's local belief, which is specified in context. If the above can be achieved, $S(\cdot)$ can serve as the basis of a payment system in federated learning such that agents paid by $S(\cdot)$ will be incentivized to contribute their local models truthfully. In this work, we primarily consider two settings, with arguably increasing difficulties in designing our mechanisms:

With ground truth verification (X, Y) We will start with a relatively easier setting where we as the designer has access to a labeled dataset $\{(x_n, y_n)\}_{n=1}^N$. We will demonstrate how this question is similar to the classical

[*] Equal contribution [1] UC Santa Cruz, Santa Cruz, CA, USA. Correspondence to: Yang Liu <yangliu@ucsc.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] Throughout this paper, we will interchange the use of agents and users.

information elicitation problem with strictly proper scoring rule (Gneiting & Raftery, 2007b), calibrated loss functions (Bartlett et al., 2006) and peer prediction (information elicitation without verification) (Miller et al., 2005b).

With only access to features X The second setting is when we only have X but not the ground truth Y. This case is arguably more popular in practice, since collecting label annotation requires a substantial amount of efforts. For instance, a company is interested in eliciting/training a classifier for an image classification problem. While it has access to images, it might not have spent efforts in collecting labels for the images. We will again present a peer prediciton-ish solution for this setting.

Besides establishing the desired incentive properties of the scoring rules, we will look into questions such as when the scoring mechanism is rewarding accurate classifiers, how to build a prediction market-ish solution to elicit improving classifiers, as well as our mechanism's robustness against possible collusion. Our work can be viewed both as a contribution to federated learning via providing incentives for selfish agents to share their hypotheses, as well as a contribution to the literature of information elicitation via studying the problem of hypothesis elicitation. We validate our claims via experiments using the MNIST and CIFAR-10 datasets.

All omitted proofs and experiment details can be found in the supplementary materials.

1.1. Related works

Due to space limit, we only briefly survey the related two lines of works:

Information elicitation Our solution concept relates most closely to the literature of information elicitation (Brier, 1950; Winkler, 1969; Savage, 1971; Matheson & Winkler, 1976; Jose et al., 2006; Gneiting & Raftery, 2007a). Information elicitation primarily focuses on the questions of developing scoring rule to incentivize or to elicite self-interested agents' private probalistic beliefs about a private event (e.g., how likely will COVID-19 death toll reach 100K by May 1?). Relevant to us, (Abernethy & Frongillo, 2011) provides a market treatment to elicit more accurate classifiers but the solution requires the designer to have the ground truth labels and agents to agree on the losses. We provide a more generic solution without above limiations.

A more challenging setting features an elicitation question while there sans ground truth verification. Peer prediction (Prelec, 2004; Miller et al., 2005a; Witkowski & Parkes, 2012; Radanovic & Faltings, 2013; Witkowski et al., 2013; Dasgupta & Ghosh, 2013; Shnayder et al., 2016; Radanovic et al., 2016; Liu & Chen, 2017; Kong & Schoenebeck, 2019; Liu et al., 2020b) is among the most popular solution concept. The core idea of peer prediction is to score each agent based on another reference report elicited from the rest of agents, and to leverage on the stochastic correlation between different agents' information. Most relevant to us is the Correlated Agreement mechanism (Dasgupta & Ghosh, 2013; Shnayder et al., 2016; Kong & Schoenebeck, 2019). We provide a separate discussion of it in Section 2.1.

Federated learning Federated learning (McMahan et al., 2016; Hard et al., 2018; Yang et al., 2019) arose recently as a promising architecture for learning from massive amounts of users' local information without polling their private data. The existing literature has devoted extensive efforts to make the model sharing process more secure (Chaum, 1988; Philippe Golle, 2004; Rastogi & Nath, 2010; Henry Corrigan Gibbs & Ford, 2013; Goryczka & Xiong, 2015; Bonawitz et al., 2016), more efficient (M. G. Rabbat, 2005; D. Golovin & Young, 2013; Gamal & Lai, 2016; Dan Alistarh & Vojnovic, 2016; Konečný et al., 2016; H. Brendan McMahan & y Arcas, 2017; Smith & Talwalkar, 2017), more robust (L. Lamport & Pease, 1982; D. Alistarh & Li, 2018; Y. Cheng & Ge, 2019; Pillutla et al., 2019) to heterogeneity in the distributed data source, among many other works. For more detailed survey please refer to several thorough ones (Yang et al., 2019; Kairouz et al., 2019).

The incentive issue has been listed as an outstanding problem in federated learning (Yang et al., 2019). There have been several very recent works touching on the challenge of incentive design in federated learning. (Liu et al., 2020a) proposed a currency system for federated learning based on blockchain techniques. (Yu et al., 2020) describes a payoff sharing algorithm that maximizes system designer's utility, but the solution does not consider the agents' strategic behaviors induced by insufficient incentives. (Yu et al., 2020) further added fairness guarantees to an above reward system. We are not aware of a systematic study of the truthfulness in incentiving hypotheses in federated learning, and our work complements above results by providing an incentive-compatible scoring system for building a payment system for federated learning.

2. Formulation

Consider the setting with a set $\mathcal{K} = \{1, 2, ..., K\}$ of agents, each with a hypothesis $f_i^* \in \mathcal{H}_i$ which maps feature space X to label space $Y \in \{1, 2, ..., L\} := [L]$. The hypothesis space $\mathcal{H}_i$ is the space of hypotheses accessible or yet considered by agent i, perhaps as a function of the subsets of X or Y which have been encountered by i or the agent's available computational power. f_i^* is often obtained following a local optimization process. For example, f_i^* can be defined as the function which minimizes a loss function over an agent's hypothesis space.

$$f_i^* = \underset{f_i \sim \mathcal{H}_i}{\operatorname{argmin}} \, \mathbb{E}_{\mathcal{D}_i} \left[\mathbb{1}\Big(f_i(X) \neq Y\Big) \right]$$

where in above $\mathcal{D}_i$ is the local distribution that agent i has access to train and evaluate f_i^*. In the federated learning setting, note that f_i^* can also represent the optimal output from a private training algorithm and $\mathcal{H}_i$ would denote a training hypothesis space that encodes a certainly level of privacy guarantees. In this paper, we do not discuss the specific ways to make a local hypothesis private [2], but rather we focus on developing scoring functions to incentivize/elicit this "private" and ready-to-be shared hypothesis.

Suppose the mechanism designer has access to a dataset D: D can be a standard training set with pairs of features and labels $D := \{(x_n, y_n)\}_{n=1}^N$, or we are in a unsupervised setting where we don't have labels associated with each sample x_i: $D := \{x_n\}_{n=1}^N$.

The *goal* of the mechanism designer is to collect f_i^* truthfully from agent i. Denote the reported/contributed hypothesis from agent i as f_i[3]. Each agent will be scored using a function S that takes all reported hypotheses $f_j, \forall j$ and D as inputs: $S\Big(f_i, \{f_{j\neq i}\}, D\Big)$ such that it is "proper" at a Bayesian Nash Equilibrium:

Definition 1. $S(\cdot)$ *is called inducing truthful reporting at a Bayesian Nash Equilibrium if for every agent i, assuming for all* $j \neq i$, $f_j = f_j^*$ *(i.e., every other agent is willing to report their hypotheses truthfully),*

$$\mathbb{E}\left[S\Big(f_i^*, \{f_{j\neq i}^*\}, D\Big)\right] \geq \mathbb{E}\left[S\Big(f_i, \{f_{j\neq i}^*\}, D\Big)\right], \quad \forall f_i,$$

where the expectation encodes agent i*'s belief about* $\{f_{j\neq i}^*\}$ *and* D.

2.1. Peer prediction

Peer prediction is a technique developed to truthfully elicit information when there is no ground truth verification. Suppose we are interested in eliciting private observations about a categorical event $y \in [L]$ generated according to a random variable Y (in the context of a machine learning task, Y can be thought of as labels). Each of the $K \geq 2$ agents holds a noisy observation of y, denoted as $y_i \in [L]$, $i \in [K]$. Again the goal of the mechanism designer is to elicit the y_is, but they are private and we do not have access to the ground truth Y to perform an evaluation. The scoring function S is designed so that truth-telling is a strict Bayesian Nash Equilibrium (implying other agents truthfully report their y_j), that is, $\forall i$, $\mathbb{E}_{y_j}\left[S\left(y_i, y_j\right)|y_i\right] > \mathbb{E}_{y_j}\left[S\left(r_i, y_j\right)|y_i\right], \forall r_i \neq y_i$.

Correlated Agreement Correlated Agreement (CA) (Dasgupta & Ghosh, 2013; Shnayder et al., 2016) is a recently established peer prediction mechanism for a multi-task setting. CA is also the core and the focus of our subsequent sections. This mechanism builds on a Δ matrix that captures the stochastic correlation between the two sources of predictions y_i and y_j. For $k, l \in [L]$, $\Delta \in \mathbb{R}^{L\times L}$ is then defined as a squared matrix with its entries defined as follows:

$$\Delta(k,l) = \mathbb{P}\big(y_i = k, y_j = l\big) - \mathbb{P}\big(y_i = k\big)\mathbb{P}\big(y_j = l\big).$$

The intuition of above Δ matrix is that each (i, j) entry of Δ captures the marginal correlation between the two predictions. $Sgn(\Delta)$ denotes the sign matrix of Δ: where $Sgn(x) = 1, x > 0$; $Sgn(x) = 0$, o.w.

CA requires each agent i to perform multiple tasks: denote agent i's observations for the N tasks as $y_{i,1}, ..., y_{i,N}$. Ultimately the scoring function $S(\cdot)$ for each task k that is shared between i, j is defined as follows: randomly draw two other tasks k_1, k_2 , $k_1 \neq k_2 \neq k$,

$$S\big(y_{i,k}, y_{j,k}\big) := Sgn\big(\Delta(y_{i,k}, y_{j,k})\big) - Sgn\big(\Delta(y_{i,k_1}, y_{j,k_2})\big),$$

It was established in (Shnayder et al., 2016) that CA is truthful and proper (Theorem 5.2, (Shnayder et al., 2016)) [4]. $\mathbb{P}(y_j = y'|y_i = y) < \mathbb{P}(y_j = y'), \forall i, j \in [K]$, $y' \neq y$ then $S(\cdot)$ is strictly truthful (Theorem 4.4, (Shnayder et al., 2016)).

3. Elicitation with verification

We start by considering the setting where the mechanism designer has access to ground truth labels, i.e., $D = \{(x_n, y_n)\}_{n=1}^N$.

3.1. A warm-up case: eliciting Bayes optimal classifier

As a warm-up, we start with the question of eliciting the Bayes optimal classifier:

$$f_i^* = \operatorname*{argmin}_{f_i} \mathbb{E}_{(X,Y)}\left[\mathbb{1}\Big(f_i(X) \neq Y\Big)\right].$$

It is straightforward to observe that, by definition using $-\mathbb{1}(\cdot)$ (negative sign changes a loss to a reward (score)) and any affine transformation of it $a\mathbb{1}(\cdot) + b, a < 0$ will be sufficient to incentivize truthful reporting of hypothesis. Next we are going to show that any classification-calibrated loss function (Bartlett et al., 2006) can serve as a proper scoring function for eliciting hypothesis.[5]

[2] There exists a variety of definitions of privacy and their corresponding solutions for achieving so. Notable solutions include *output perturbation* (Chaudhuri et al., 2011) or *output sampling* (Bassily et al., 2014) to preserve privacy when differential privacy (Dwork, 2006) is adopted to quantify the preserved privacy level.

[3] f_i can be none if users chose to not contribute.

[4] To be precise, it is an informed truthfulness. We refer interested readers to (Shnayder et al., 2016) for the detailed differences.

[5] We provide details of the calibration in the proof. Classical examples include cross-entropy loss, squared loss, etc.

Theorem 1. *Any classification calibrated loss function $\ell(\cdot)$ (paying agents $-\ell(f_i(X), Y)$) induces truthful reporting of the Bayes optimal classifier.*

3.2. Eliciting "any-optimal" classifier: a peer prediction approach

Now consider the case that an agent does not hold an absolute Bayes optimal classifier. Instead, in practice, agent's local hypothesis will depend on the local observations they have, the privacy level he desired, the hypothesis space and training method he is using. Consider agent i holds the following hypothesis f_i^*, according to a loss function ℓ_i, and a hypothesis space $\mathcal{H}_i$: $f_i^* = \text{argmin}_{f_i \sim \mathcal{H}_i} \mathbb{E}\left[\ell_i\Big(f_i(X), Y\Big)\right]$.

By definition, each specific ℓ_i will be sufficient to incentivize a hypothesis. However, it is unclear how f_i^* trained using ℓ_i would necessarily be optimal according to a universal metric/score. We aim for a more generic approach to elicit different f_i^*s that are returned from different training procedure and hypothesis classes. In the following sections, we provide a peer prediction approach to do so.

We first state the hypothesis elicitation problem as a standard peer prediction problem. The connection is made by firstly rephrasing the two data sources, the classifiers and the labels, from agents' perspective. Let's re-interpret the ground truth label Y as an "optimal" agent who holds a hypothesis $f^*(X) = Y$. We denote this agent as A^*. Each local hypothesis f_i^* agent i holds can be interpreted as the agent that observes $f_i^*(x_1), ..., f_i^*(x_N)$ for a set of randomly drawn feature vectors $x_1, ..., x_N$: $f_i^*(x_n) \sim \mathsf{A}_i(X)$. Then a peer prediction mechanism induces truthful reporting if: $\mathbb{E}\left[S(f_i^*(X), f^*(X))\right] \geq \mathbb{E}\left[S(f(X), f^*(X))\right], \forall f$.

Correlated Agreement for hypothesis elicitation To be more concrete, consider a specific implementation of peer prediction mechanism, the Correlated Agreement (CA). Recall that the mechanism builds on a correlation matrix $\Delta(f_i^*(X), f^*(X))$ defined as follows:

$$\begin{aligned}\Delta^*(k,l) =& \mathbb{P}\big(f_i^*(X) = k, f^*(X) = l\big) \\ &- \mathbb{P}\big(f_i^*(X) = k\big)\mathbb{P}\big(f^*(X) = l\big),\ k, l \in [L].\end{aligned}$$

Then the CA for hypothesis elicitation is summarized in Algorithm 1.

We reproduce the incentive guarantees and required conditions:

Theorem 2. *CA mechanism induces truthful reporting of a hypothesis at a Bayesian Nash Equilibrium.*

Knowledge requirement of Δ^* We'd like to note that knowing the sign of Δ^* matrix between f_i^* and f^* is a relatively weak assumption to have to run the mechanism. For example, for a binary classification task $L = 2$, define the following accuracy measure,

Algorithm 1 CA for Hypothesis Elicitation

```
For each sample $x_n$, randomly sample two other tasks $x_{p_1} \neq x_{p_2} \neq x_n$ to pair with.
Pay a reported hypothesis $f(\cdot)$ for $x_n$ according to
    $S(f(x_n), f^*(x_n)) := Sgn\left(\Delta^*(f(x_n), f^*(x_n))\right) - Sgn\left(\Delta^*(f(x_{p_1}), f^*(x_{p_2}))\right)$   (1)
Total payment to a reported hypothesis $f$:
    $S(f, f^*) := \sum_{n=1}^{N} S(f(x_n), f^*(x_n)).$
```

$$\mathsf{FNR}(f) := \mathbb{P}(f(X) = 2 | Y = 1),$$

$$\mathsf{FPR}(f) := \mathbb{P}(f(X) = 1 | Y = 2).$$

We offer the following:

Lemma 1. *For binary classification ($L = 2$), if $\mathsf{FNR}(f_i^*) + \mathsf{FPR}(f_i^*) < 1$, $Sgn(\Delta^*)$ is an identify matrix.*

$\mathsf{FNR}(f_i^*) + \mathsf{FPR}(f_i^*) < 1$ is stating that f_i^* is informative about the ground truth label Y (Liu & Chen, 2017). Similar conditions can be derived for $L > 2$ to guarantee an identify $Sgn(\Delta^*)$. With identifying a simple structure of $Sgn(\Delta^*)$, the CA mechanism for hypothesis elicitation runs in a rather simple manner.

When do we reward accuracy The elegance of the above CA mechanism leverages the correlation between a classifier and the ground truth label. Ideally we'd like a mechanism that rewards the accuracy of the contributed classifier. Consider the binary label case:

Theorem 3. *When $\mathbb{P}(Y = 1) = 0.5$ (uniform prior), and let $Sgn(\Delta^*) = I_{2\times 2}$ be the identity matrix, the more accurate classifier within each $\mathcal{H}_i$ receives a higher score.*

Note that the above result does not conflict with our incentive claims. In an equal prior case, misreporting can only reduce a believed optimal classifier's accuracy instead of the other way. It remains an interesting question to understand a more generic set of conditions under which CA will be able to incentivize contributions of more accurate classifiers.

A market implementation The above scoring mechanism leads to a market implementation (Hanson, 2007) that incentivizes improving classifiers. In particular, suppose agents come and participate at discrete time step t. Denote the hypothesis agent contributed at time step t as f_t^* (and his report f_t). Agent at time t will be paid according to $S(f_t(X), Y) - S(f_{t-1}(X), Y),$ where $S(\cdot)$ is an incentive-compatible scoring function that elicits f_t truthfully using Y. The incentive-compatibility of the market payment is immediate due to $S(\cdot)$. The above market implementation incentivizes improving classifiers with bounded budget [6].

Calibrated CA scores When $Sgn(\Delta^*)$ is the identity matrix, the CA mechanism reduces to:

$$\begin{aligned} S(f(x_n), f^*(x_n)) :=& \mathbb{1}\left(f(x_n) = f^*(x_n)\right) \\ &- \mathbb{1}\left(f(x_{p_1}) = f^*(x_{p_2})\right) \end{aligned}$$

That is the reward structure of CA builds on 0-1 loss function. We ask the question of can we extend the CA to a calibrated one? We define the following loss-calibrated scoring function for CA:

$$\begin{aligned} \textsf{Calibrated CA:}\quad & S_\ell(f(x_n), f^*(x_n)) \\ &= -\ell(f(x_n), f^*(x_n))) \\ &\quad - (-\ell(f(x_{p_1}), f^*(x_{p_2}))). \end{aligned}$$

Here again we negate the loss ℓ to make it a reward (agent will seek to maximize it instead of minimizing it). If this extension is possible, not only we will be able to include more scoring functions, but also we are allowed to score/verify non-binary classifiers directly. Due to space limit, we provide positive answers and detailed results in Appendix, while we will present empirical results on the calibrated scores of CA in Section 5.

4. Elicitation without verification

Now we move on to a more challenging setting where we do not have ground truth label Y to verify the accuracy, or the informativeness of $f(X)$, i.e., the mechanism designer only has access to a $D = \{x_n\}_{n=1}^N$. The main idea of our solution from this section follows straight-forwardly from the previous section, but instead of having a ground truth agent f^*, for each classifier f_i^* we only have a reference agent f_j^* drawn from the rest agents $j \neq i$ to score with. The corresponding scoring rule takes the form of $S(f_i(X), f_j(X))$, and similarly the goal is to achieve the following: $\mathbb{E}\left[S(f_i^*(X), f_j^*(X))\right] \geq \mathbb{E}\left[S(f(X), f_j^*(X))\right], \forall f$.

As argued before, if we treat f_i and f_j as two agents A_i and A_j holding private information, a properly defined peer prediction scoring function that elicits A_i using A_j will suffice to elicit f_i using f_j. Again we will focus on using Correlated Agreement as a running example. Recall that the mechanism builds on a correlation matrix $\Delta^*(f_i^*(X), f_j^*(X))$.

$$\begin{aligned} \Delta^*(k,l) =& \mathbb{P}\big(f_i^*(X) = k, f_j^*(X) = l\big) \\ &- \mathbb{P}\big(f_i^*(X) = k\big)\mathbb{P}\big(f_j^*(X) = l\big),\ k,l \in [L] \end{aligned}$$

The mechanism then operates as follows: For each task x_n, randomly sample two other tasks x_{p_1}, x_{p_2}. Then pay a reported hypothesis according to

$$\begin{aligned} S(f_i, f_j) := \sum_{n=1}^N & Sgn\left(\Delta^*(f_i(x_n), f_j(x_n))\right) \\ &- Sgn\left(\Delta^*(f_i(x_{p_1}), f_j(x_{p_2}))\right) \end{aligned} \tag{2}$$

We reproduce the incentive guarantees:

Theorem 4. *CA mechanism induces truthful reporting of a hypothesis at a Bayesian Nash Equilibrium.*

The proof is similar to the proof of Theorem 2 so we will not repeat the details in the Appendix.

To enable a clean presentation of analysis, the rest of this section will focus on using/applying CA for the binary case $L = 2$. First, as an extension to Lemma 1, we have:

Lemma 2. *If f_i^* and f_j^* are conditionally independent given Y, $\textsf{FNR}(f_i^*) + \textsf{FPR}(f_i^*) < 1$ and $\textsf{FNR}(f_j^*) + \textsf{FPR}(f_j^*) < 1$, then $Sgn(\Delta^*)$ is an identify matrix.*

When do we reward accuracy As mentioned earlier that in general peer prediction mechanisms do not incentivize accuracy. Nonetheless we provide conditions under which they do. The result below holds for binary classifications.

Theorem 5. *When (i) $\mathbb{P}(Y = 1) = 0.5$, (ii) $Sgn(\Delta^*) = I_{2\times 2}$, and (iii) $f_i^*(X)$ and $f_j^*(X)$ are conditional independent of Y, the more accurate classifier within each $\mathcal{H}_i$ receives a higher score in expectation.*

4.1. Peer Prediction market

Implementing the above peer prediction setting in a market setting is hard, due to again the challenge of no ground truth verification. The use of reference answers collected from other peers to similarly close a market will create incentives for further manipulations.

Our first attempt is to crowdsource to obtain an independent survey answer and use the survey answer to close the market. Denote the survey hypothesis as f' and use f' to close the market:

$$S(f_t(x), f'(x)) - S(f_{t-1}(X), f'(x)) \tag{3}$$

Theorem 6. *When the survey hypothesis $f'(x)$ is (i) conditionally independent from the market contributions, and (ii) Bayesian informative, then closing the market using the crowdsourcing survey hypothesis is incentive compatible.*

[6] Telescoping returns: $\sum_{t=1}^T \left(S(f_t(X), Y) - S(f_{t-1}(X), Y)\right) = S(f_T(X), Y) - S(f_0(X), Y)$.

The above mechanism is manipulable in several aspects. Particularly, the crowdsourcing process needs to be independent from the market, which implies that the survey participant will need to stay away from participating in the market - but it is unclear whether this will be the case.

Our idea of making a robust extension is to pair the market with a separate "survey" elicitation process that elicits redundant C hypotheses:

Algorithm 2 Market for Hypothesis Elicitation

```
Pay crowdsourcing survey participants using surveys and the CA mechanism;
Randomly draw a hypothesis from the surveys (from the $C$ collected ones) to close the market according to Eqn. (3), up to a scaling factor $\lambda > 0$.
```

Assuming the survey hypotheses are conditional independent w.r.t. the ground truth Y. Denote by f' a randomly drawn hypothesis from the surveys, and

$$\alpha' := \mathbb{P}(f'(X) = 2|Y = 1), \;\; \beta' := \mathbb{P}(f'(X) = 1|Y = 2)$$

For an agent who participated both in survey and market will receive the following:

$$S(f_i, f'_{-i}) + \lambda(S(f_t, f') - S(f_{t-1}, f'))$$

Below we establish its incentive property:

Theorem 7. *For any δ that:*

$$\delta := \frac{\lambda}{(1-\alpha'-\beta')\cdot(C+\lambda(C-1))} \underbrace{\longrightarrow}_{C\to\infty} 0,$$

agents have incentives to report a hypothesis that is at most δ less accurate than the truthful one.

Remark Before we conclude this section, we remark that the above solution for the *without verification* setting also points to an hybrid solution when the designer has access to both sample points with and without ground truth labels. The introduction of the pure peer assessment solution helps reduce the variance of payment.

4.2. Robust elicitation

Running a peer prediction mechanism with verifications coming only from peer agents is vulnerable when facing collusion. In this section we answer the question of how robust our mechanisms are when facing a γ-fraction of adversary in the participating population. To instantiate our discussion, consider the following setting

- There are $1-\gamma$ fraction of agents who will act truthfully if incentivized properly. Denote the randomly drawn classifier from this $1-\gamma$ population as $f^*_{1-\gamma}$.
- There are γ fraction of agents are adversary, whose reported hypotheses can be arbitrary and are purely adversarial.

Denote the following quantifies $\alpha := \mathbb{P}(f^*_{1-\gamma}(X) = 2|Y = 1)$, $\beta := \mathbb{P}(f^*_{1-\gamma}(X) = 1|Y = 2)$ $\alpha^* := \mathbb{P}(f^*(X) = 2|Y = 1)$, $\beta^* := \mathbb{P}(f^*(X) = 1|Y = 2)$, that is α, β are the error rates for the eliciting classifier $f^*_{1-\gamma}$ while α^*, β^* are the error rates for the Bayes optimal classifier. We prove the following

Theorem 8. *CA is truthful in eliciting hypothesis when facing γ-fraction of adversary when γ satisfies: $\frac{1-\gamma}{\gamma} > \frac{1-\alpha^*-\beta^*}{1-\alpha-\beta}$.*

When the agent believes that the classifier the $1-\gamma$ crowd holds is as accurate as the Bayes optimal classifier we have $\frac{1-\alpha^*-\beta^*}{1-\alpha-\beta} = \frac{1-\alpha^*-\beta^*}{1-\alpha^*-\beta^*} = 1$, then a sufficient condition for eliciting truthful reporting is $\gamma < 50\%$, that is our mechanism is robust up to half of the population manipulating. Clearly the more accurate the reference classifier is, the more robust our mechanism is.

5. Experiments

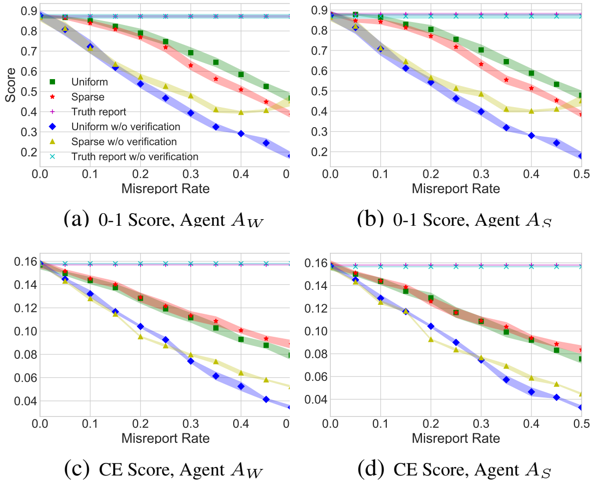
In this section, we implement two reward structures of CA: 0-1 score and Cross-Entropy (CE) score as mentioned at the end of Section 3.2. We experiment on two image classification tasks: MNIST (LeCun & Haffner, 1998a) and CIFAR-10 (Krizhevsky, 2009) in our experiments. For agent A_W (weak agent), we choose LeNet (LeCun & Haffner, 1998a) and ResNet34 (He, 2016) for MNIST and CIFAR-10 respectively. For A_S (strong agent), we use a 13-layer CNN architecture for both datasets.

Either of them is trained on random sampled 25000 images from each image classification training task. After the training process, agent A_W reaches 99.37% and 62.46% test accuracy if he truthfully reports the prediction on MNIST and CIFAR-10 test data. Agent A_S is able to reach 99.74% and 76.89% test accuracy if the prediction on MNIST and CIFAR-10 test data is truthfully reported.

A_W and A_S receive hypothesis scores based on the test data X_{test} (10000 test images) of MNIST or CIFAR-10. For elicitation with verification, we use ground truth labels to calculate the hypothesis score. For elicitation without verification, we replace the ground truth labels with the other agent's prediction - A_W will serve as A_S's peer reference hypothesis and vice versa.

5.1. Results

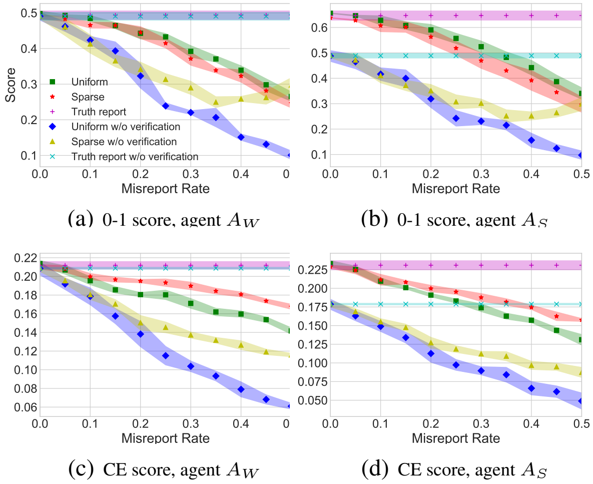
Statistically, an agent i's mis-reported hypothesis can be expressed by a misreport transition matrix T. Each element $T_{j,k}$ represents the probability of flipping the truthfully re-



(a) 0-1 Score, Agent A_W (b) 0-1 Score, Agent A_S

(c) CE Score, Agent A_W (d) CE Score, Agent A_S

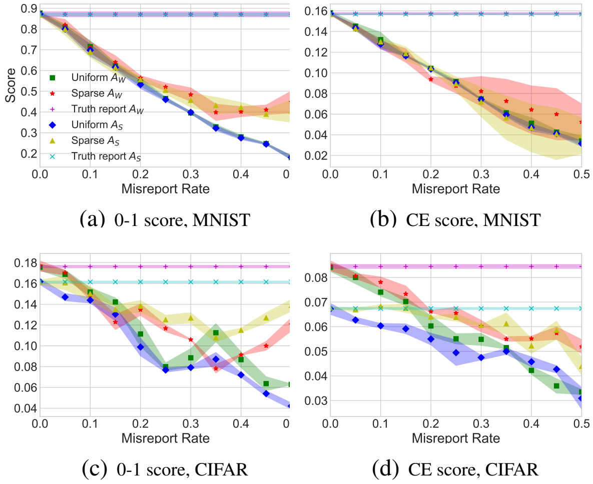
Figure 1. Hypothesis scores versus misreport rate on MNIST dataset.



(a) 0-1 score, agent A_W (b) 0-1 score, agent A_S

(c) CE score, agent A_W (d) CE score, agent A_S

Figure 2. Hypothesis scores versus misreport rate on CIFAR-10 dataset.



(a) 0-1 score, MNIST (b) CE score, MNIST

(c) 0-1 score, CIFAR (d) CE score, CIFAR

Figure 3. Hypothesis scores versus misreport rate (with adversarial attack).

ported label $f_i^*(x) = j$ to the misreported label $\tilde{f}_i(x) = k$: $T_{j,k} = \mathbb{P}(\tilde{f}_i(X) = k | f_i^*(X) = j)$. Random flipping predictions will degrade the quality of a classifier. When there is no adversary attack, we focus on two kinds of misreport transition matrix: a uniform matrix or a sparse matrix. For the uniform matrix, we assume the probability of flipping from a given class into other classes to be the same: $T_{i,j} = T_{i,k} = e, \forall i \neq j \neq k$. e changes gradually from 0 to 0.56 after 10 increases, which results in a 0%–50% misreport rate. The sparse matrix focuses on particular 5 pairs of classes which are easily mistaken between each pair. Denote the corresponding transition matrix elements of class pair (i, j) to be: $(T_{ij}, T_{ji}), i \neq j$, we assume that $T_{ij} = T_{ji} = e, \forall (i, j)$. e changes gradually from 0 to 0.5 after 10 increases, which results in a 0%–50% misreport rate.

Every setting is simulated 5 times. The line in each figure consists of the median score of 5 runs as well as the corresponding "deviation interval", which is the maximum absolute score deviation. The y axis symbolizes the averaged score of all test images.

As shown in Figure 1, 2, in most situations, 0-1 score and CE score of both A_W and A_S keep on decreasing while the misreport rate is increasing. As for 0-1 score without ground truth verification, the score of either agent begins to fluctuate more when the misreport rate in sparse misreport model is $> 35\%$. Our results conclude that both the 0-1 score and CE score induce truthful reporting of a hypothesis and will penalize misreported agents whether there is ground truth for verification or not.

5.2. Elicitation with adversarial attack

We test the robustness of our mechanism when facing a 0.3-fraction of adversary in the participating population. We introduce an adversarial agent, LinfPGDAttack, introduced in AdverTorch (Ding et al., 2019) to influence the labels for verification when there is no ground truth. In Figure 3, both the 0-1 score and CE score induce truthful reporting of a hypothesis for MNIST.

However, for CIFAR-10, with the increasing of misreport rate, the decreasing tendency fluctuates more often. Two factors attribute to this phenomenon: the agents' abilities as well as the quality of generated "ground truth" labels. When the misreport rate is large and generated labels are of low quality, the probability of successfully matching the misreported label to an incorrect generated label can be much higher than usual. But in general, these two scoring structures incentivize agents to truthfully report their results.

6. Concluding remarks

This paper provides an elicitation framework to incentivize contribution of truthful hypotheses in federated learning. We have offered a scoring rule based solution template which we name as hypothesis elicitation. We establish the incentive property of the proposed scoring mechanisms and have tested their performance with real-world datasets extensively. We have also looked into the accuracy, robustness of the scoring rules, as well as market approaches for implementing them.

References

Abernethy, J. D. and Frongillo, R. M. A collaborative mechanism for crowdsourcing prediction problems. In *Advances in Neural Information Processing Systems*, pp. 2600–2608, 2011.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Chaum, D. The dining cryptographers problem: unconditional sender and recipient untraceability. In *Journal of Cryptology*, pp. 65–75. 1988.

Chen, M., Suresh, A. T., Mathews, R., Wong, A., Allauzen, C., Beaufays, F., and Riley, M. Federated learning of n-gram language models. *arXiv preprint arXiv:1910.03432*, 2019.

D. Alistarh, Z. A.-Z. and Li, J. Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 31, pp. 4618–4628. 2018.

D. Golovin, D. Sculley, H. B. M. and Young, M. Large-scale learning with less ram via randomization. In *30th International Conference on Machine Learning*, pp. 325–333, 2013.

Dan Alistarh, Jerry Li, R. T. and Vojnovic, M. Qsgd: Randomized quantization for communication-optimal stochastic gradient descent. *arXiv preprint arXiv:1610.02132*, 2016.

Dasgupta, A. and Ghosh, A. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 319–330. International World Wide Web Conferences Steering Committee, 2013.

Ding, G. W., Wang, L., and Jin, X. AdverTorch v0.1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*, 2019.

Dwork, C. Differential privacy. In *Automata, languages and programming*, pp. 1–12. Springer, 2006.

Gamal, M. E. and Lai, L. On randomized distributed coordinate descent with quantized updates. *arXiv:1609.05539*, 2016.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007a.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007b.

Goryczka, S. and Xiong, L. A comprehensive comparison of multiparty secure additions with differential privacy. In *IEEE Transactions on Dependable and Secure Computing*, volume 14, pp. 463–477, 2015.

H. Brendan McMahan, Eider Moore, D. R. S. H. and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Hanson, R. Logarithmic markets coring rules for modular combinatorial information aggregation. *The Journal of Prediction Markets*, 1(1):3–15, 2007.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

He, K., Z. X. R. S. S. J. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Henry Corrigan Gibbs, D. I. W. and Ford, B. Proactively accountable anonymous messaging in verdict. In *22nd USENIX International Conference on security*, pp. 147–162. USENIX Association, 2013.

Jose, V. R., Nau, R. F., and Winkler, R. L. Scoring rules, generalized entropy and utility maximization. Working Paper, Fuqua School of Business, Duke University, 2006.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016. URL https://arxiv.org/abs/1610.05492.

Kong, Y. and Schoenebeck, G. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)*, 7(1):2, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 2009.

L. Lamport, R. E. S. and Pease, M. C. The byzantine generals problem. In *ACM Trans. Program. Lang. Syst.*, volume 4, pp. 382–401. 1982.

LeCun, Y., B. L. B. Y. and Haffner, P. Gradient-based learning applied to document recognition. In *IEEE*, volume 86, pp. 2278–2324, 1998a.

Liu, Y. and Chen, Y. Machine Learning aided Peer Prediction. *ACM EC*, June 2017.

Liu, Y., Sun, S., Ai, Z., Zhang, S., Liu, Z., and Yu, H. Fedcoin: A peer-to-peer payment system for federated learning, 2020a.

Liu, Y., Wang, J., and Chen, Y. Surrogate scoring rules. *ACM EC*, 2020b.

M. G. Rabbat, R. D. N. Quantized incremental algorithms for distributed optimization. In *IEEE Journal on Selected Areas in Communications*, volume 23, pp. 798–808. 2005.

Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management Science*, 22 (10):1087–1096, 1976.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

Miller, N., Resnick, P., and Zeckhauser, R. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359 –1373, 2005a.

Miller, N., Resnick, P., and Zeckhauser, R. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005b.

Philippe Golle, A. J. Dining cryptographers revisited. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 456–473. Springer, 2004.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

Prelec, D. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.

Radanovic, G. and Faltings, B. A robust bayesian truth serum for non-binary signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, AAAI '13, 2013.

Radanovic, G., Faltings, B., and Jurca, R. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):48, 2016.

Rastogi, V. and Nath, S. Differentially private aggregation of distributed time-series with transformation and encryption. In *2010 ACM SIGMOD International Conference on Management of data*, pp. 735–746, 2010.

Savage, L. J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Shnayder, V., Agarwal, A., Frongillo, R., and Parkes, D. C. Informed Truthfulness in Multi-Task Peer Prediction. *ACM EC*, March 2016.

Shnayder, V., Agarwal, A., Frongillo, R., and Parkes, D. C. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 179–196. ACM, 2016.

Smith, V., C. C.-K. S. m. and Talwalkar, A. S. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, pp. 4424–4434. 2017.

Winkler, R. L. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 64(327):1073–1078, 1969.

Witkowski, J. and Parkes, D. A robust bayesian truth serum for small populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, AAAI '12, 2012.

Witkowski, J., Bachrach, Y., Key, P., and Parkes, D. C. Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*, 2013.

Y. Cheng, I. D. and Ge, R. High-dimensional robust mean estimation in nearly-linear time. In *ACM-SIAM Symposium on Discrete Algorithms*, pp. 2755–2771. 2019.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Yu, H., Liu, Z., Liu, Y., Chen, T., Cong, M., Weng, X., Niyato, D., and Yang, Q. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 393–399, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375840. URL https://doi.org/10.1145/3375627.3375840.

Yu, H., Liu, Z., Liu, Y., Chen, T., Cong, M., Weng, X., Niyato, D., and Yang, Q. A sustainable incentive scheme for federated learning. *IEEE Intelligent Systems*, pp. 1–1, 2020.