Peer Prediction for Learning Agents

Anonymous Author(s)
Affiliation
Address
email

Abstract

Peer prediction refers to a collection of mechanisms for eliciting information from human agents when direct verification of the obtained information is unavailable. They are designed to have a game-theoretic equilibrium where everyone reveals their private information truthfully. This result holds under the assumption that agents are Bayesian and they each adopt a fixed strategy across all tasks. Human agents however are observed in many domains to exhibit learning behavior in sequential settings. In this paper, we explore the dynamics of sequential peer prediction mechanisms when participants are learning agents. We first show that the notion of no regret alone for the agents' learning algorithms cannot guarantee convergence to the truthful strategy. We then focus on a family of learning algorithms where strategy updates only depend on agents' cumulative rewards and prove that agents' strategies in the popular Correlated Agreement (CA) mechanism converge to truthful reporting when they use algorithms from this family. This family of algorithms is not necessarily no-regret, but includes several familiar no-regret learning algorithms (e.g multiplicative weight update and Follow the Perturbed Leader) as special cases. Simulation of several algorithms in this family as well as the ϵ-greedy algorithm, which is outside of this family, shows convergence to the truthful strategy in the CA mechanism.

1 Introduction

A fundamental challenge in many domains is to elicit high-quality information from people when directly verifying the acquired information is *not* feasible, either because the ground truth is not available or because it's too costly to obtain. Notable settings include asking people to label data for machine learning, having students perform peer grading in education, and soliciting customer feedback for products and services.

The peer prediction literature has made impressive progress on this challenge in the past two decades, with many mechanisms that have desirable incentive properties developed for this problem [19, 18, 26, 5, 21, 8, 13, 24, 16, 15, 20, 17, 22, 14].The term peer prediction refers to a collection of reward mechanisms that solicit information from human agents and reward each agent solely based on how the agent's reported information compares with that of the other agents, without having access to the ground truth. Under some assumptions, many peer prediction mechanisms [19, 18, 26, 21, 8, 15, 20] guarantee that every agent truthfully reporting their information is a game-theoretic equilibrium, and the more recent multi-task peer prediction mechanisms [5, 13, 24, 16, 17, 22, 14] further ensure that agents receive the highest expected payoff at the truthful equilibrium, compared with other strategy profiles.

While achieving truthful reporting as a highest-payoff equilibrium is a victory to declare for this challenging without-verification setting, there are however caveats associated with adopting the notion of equilibrium as a solution concept. The equilibrium results rely on the assumption that participants are fully rational Bayesian agents. Equilibrium is a static notion and doesn't address how

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

agents, who act independently, jump to play their equilibrium strategies. Moreover, the equilibrium results of multi-task peer prediction mechanisms heavily depend on a *consistent strategy* assumption, that is, each agent is assumed to adopt a fixed strategy across all tasks that she participates. All together these assumptions exclude the possibility that agents may explore and learn from previous experience, a behavior that's not only commonly observed in practice but also has been modeled in studying other strategic settings [3, 6, 4].

This paper is the first theoretical study on the dynamics of sequential peer prediction mechanisms when participants are learning agents. The main question that we explore is whether and when in sequential peer prediction, learning agents will converge to all playing the truthful reporting strategy. We first consider agents adopting no-regret learning algorithms and prove that the notion of no regret alone cannot guarantee convergence to truthful reporting. We then define a natural family of reward-based learning algorithms where strategy updates only depend on agents' cumulative rewards. While algorithms in this family is not necessarily no-regret (e.g. the Follow the Leader algorithm), this family includes some familiar no-regret learning algorithms, including the Multiplicative Weight Update and the Follow the Perturbed Leader algorithms. Our main result shows that, for the binary-signal setting, agents' strategies in the popular Correlated Agreement (CA) mechanism [5] converge to truthful reporting when agents use any algorithm from this family. We track the process using the agents' accumulated rewards in the CA mechanism, and discover the process has a self-fulfilling property: once Alice and Bob have large accumulated rewards for truth-telling, they are more likely to play truth-telling and resulting in larger accumulated rewards. Theoretically, we carefully partition the process into three stages, bad, intermediate, and good events illustrated in fig. 1, and use tools in martingale theory to argue the progress of the process. Finally, we simulate the strategy dynamics in the CA mechanism for several algorithms in this family as well as for the ϵ-greedy algorithm, which doesn't belong to this family. We observe convergence to truthful reporting for all algorithms considered in our simulation, suggesting an interesting future direction to characterize all learning algorithms that converge to truthful reporting.

1.1 Related Works

This paper relates to two lines of work, information elicitation and mechanisms for learning agents.

Information Elicitation Mechanisms The literature on information elicitation without verification focuses on capturing the strategic aspect of human agents. In multi-task settings, Dasgupta and Ghosh [5] proposed a seminal informed truthful mechanism, the Correlated Agreement (CA) mechanism, for binary positively correlated signals. A series of works then relaxed the binary and positively correlation assumptions [24, 16, 22, 14]. Additionally, Zheng et al. [27] study the limitation of information elicitation in the multi-task setting. However, all of the above works assume agents using consistent strategies that are identical across all tasks. The consistent strategy assumption excludes the possibility of agent learning. Our work removes the consistent strategy assumption and explicitly considers learning agents. We theoretically prove truthful convergence of the CA mechanism when agents using algorithms from a family of reward-based online learning algorithms.

Our work of considering learning agents can be viewed as a way of testing the robustness of information elicitation mechanisms with respect to deviation from the rational Bayesian agent model. From this perspective, Shnayder et al. [25] is closely related to ours. They consider sequential information elicitation and empirically study if agents using replicator dynamics can converge to truth-telling in the CA mechanism and several other mechanisms. Our work theoretically proves that, besides replicator dynamics, learning agents can converge to truth-telling in the CA mechanism when they use a general family of learning algorithms. Additionally, Schoenebeck et al. [23] designed an information elicitation mechanism that was robust against a small fraction of adversarial agents.

Mechanisms for Learning Agents Several works in economics and computer science try to design mechanisms for learning agents, rather than for rational, Bayesian ones. Braverman et al. [3] studied pricing mechanisms for learning agents with no external regrets. Their work was generalized by Deng et al. [6] to consider repeated Stackelberg games in full-information settings. Camara et al. [4] further proposed counterfactual internal regrets (CIR) together with no-CIR assumption, which was proved to be a sufficient behavior assumption for no-regret principal mechanism design in repeated stage games. However, all of these works focus on a single agent or full-information games, while peer prediction is an incomplete-information game with multiple agents. Finally, our goal is slightly

different from that of most sequential mechanism design. Instead of maximizing the mechanism designer's utility, our goal is to incentivize truthful reporting form agents.

2 Sequential Peer Prediction Settings

For simplicity we consider two agents, Alice and Bob, who work on a sequence of tasks indexed by $t \geq 1$. For round t, both agents work on task t, Alice receives a signal $X_t = x_t$ in $\{0, 1\}$, and Bob a signal $Y_t = y_t$ in $\{0, 1\}$. Then Alice and Bob report $\hat{X}_t = \hat{x}_t$ and $\hat{Y}_t = \hat{y}_t$ in $\{0, 1\}$. We define $\mathbf{X}_{\leq t} = \{X_s : 1 \leq s \leq t\} \in \{0, 1\}^t$ and $\hat{\mathbf{X}}_{\leq t} = \{\hat{X}_s : 1 \leq s \leq t\} \in \{0, 1\}^t$ to denote Alice's signal profile and report profiles until t-th round respectively, and define $\mathbf{Y}_{\leq t}$ and $\hat{\mathbf{Y}}_{\leq t}$ for Bob similarly. We use $\mathbf{X}, \hat{\mathbf{X}}, \mathbf{Y}$, and $\hat{\mathbf{Y}}$ for the complete signal and report profiles. Additionally, we consider the signals are generated from some distribution $\mathbb{P}$ that satisfies the following assumptions:

Assumption 2.1 (A priori similar tasks [5])**.** Each pair of signal is identically and independently (i.i.d.) generated: there exists a distribution $P_{X,Y}$ over $\{0, 1\}^2$ such that $(X_t, Y_t) \sim P_{X,Y}$ for any $t \in \mathbb{N}^+$. Moreover, we assume the distribution has full supports, $P_{X,Y}(x, y) > 0$ for all $x, y \in \{0, 1\}$.

Assumption 2.2 (Positively correlated signals)**.** The distribution $P_{X,Y}$ is positively correlated, $\min\{P_{X,Y}(1, 1), P_{X,Y}(0, 0)\} > \max\{P_{X,Y}(1, 0), P_{X,Y}(0, 1)\}$.

A *sequential information elicitation mechanism* is a sequence of payment functions $\mathcal{M} = \{M_t : t \geq 1\}$ where $M_t : \{0, 1\}^{2\times t} \to [-1, 1]$ for all t. After Alice and Bob reporting $\hat{\mathbf{x}}_{\leq t}$ and $\hat{\mathbf{y}}_{\leq t}$ in round t, the mechanism computes $(r_t, s_t) := M_t(\hat{\mathbf{x}}_{\leq t}, \hat{\mathbf{y}}_{\leq t})$ and pay r_t to Alice and s_t to Bob. Here we assume M_t can only depends on a constant k round of reports so that $M_t(\hat{\mathbf{x}}_{\leq t}, \hat{\mathbf{y}}_{\leq t}) = M_t(\hat{x}_{t-k+1}, \hat{x}_{t-k+1}, \ldots, \hat{x}_t, \hat{y}_{t-k+1}, \hat{y}_{t-k+1}, \ldots, \hat{y}_t)$ for all t, $\hat{x}_{\leq t}$, and $\hat{y}_{\leq t}$, and we call such $\mathcal{M}$ *rank k mechanism* . For instance, previous work [5, 24, 25] study *correlated agreement mechanism* (CA mechanism)[1] that can be adopted as a sequential rank 2 information elicitation mechanism: At round t, the payment is

$$M_t^{agree}(\hat{\mathbf{x}}_{\leq t}, \hat{\mathbf{y}}_{\leq t}) = (\mathbb{I}[\hat{x}_t = \hat{y}_t] - \mathbb{I}[\hat{x}_t = \hat{y}_{t-1}], \mathbb{I}[\hat{y}_t = \hat{x}_t] - \mathbb{I}[\hat{y}_t = \hat{x}_{t-1}]) \tag{1}$$

We let $\hat{x}_0$ and $\hat{y}_0$ to be both 0 for the first round. Note that *multi-task information elicitation mechanisms* on T task is a payment function that maps agents' reports on all T tasks to payment, e.g., the payment of CA mechanism is M^T when the number of tasks is T. Thus, multi-task information elicitation mechanisms can be seen as a special case of the sequential one.

Now we formally define agents' actions and strategies. Due to symmetry, we introduce notation for Alice and omit Bob's. Given an information elicitation mechanism $\mathcal{M}$, at round t, Alice can choose an action observes her signal x_t and reports $\hat{x}_t$. Thus, Alice has four *options (actions)*: 1) opt_1: report the private signal truthfully, 2) opt_2: flip the private signal, 3) opt_3: report 1 regardless of the signal, and 4) opt_4: report 0 regardless of the signal. We call opt_3 and opt_4 as uninformative actions. We use opt_t^X to denote Alice's actions, and r_t for their payoff at round r. At each round t, Alice knows her previous signals $\mathbf{x}_{\leq t} \in \{0, 1\}^t$, her actions $\text{opt}_1^X, \ldots, \text{opt}_{t-1}^X$, and Bob's reports $\hat{\mathbf{y}}_{\leq t-1}$, so we use $\mathcal{F}_t = \{\mathbf{x}_{\leq t}, \text{opt}_{\leq t-1}^X, \hat{\mathbf{y}}_{\leq t-1}\} \subset \{0, 1\}^t \times \{\text{opt}_1, \text{opt}_2, \text{opt}_3, \text{opt}_4\}^{t-1} \times \{0, 1\}^{t-1}$ to denote Alice knowledge at round t. Thus, Alice's strategy at round t is a stochastic mapping σ_t^X so that $\sigma_t^X(\mathcal{F}_t) \in \{\text{opt}_1, \text{opt}_2, \text{opt}_3, \text{opt}_4\}$, and a learning algorithm of Alice is a mapping from an information elicitation mechanism $\mathcal{M}$ to her strategy.

2.1 Strongly Truthful for Rational and Bayesian agents

Previous works on multi-task information elicitation try to ensure truth-telling opt_1 is the best action for rational and strategic agents. In particular, a mechanism is *strongly truthful* if Alice and Bob get strictly higher payment than any other non-permutation Bayesian Nash equilibrium. We present the formal definitions in the appendix. However, previous mechanisms not only require assumption 2.1 but further assume that agents use *consistent strategy*. Specifically, Alice uses a consistent strategy if there is a fixed distribution μ^X on $\{\text{opt}_1, \text{opt}_2, \text{opt}_3, \text{opt}_4\}$ so that her action at round t is generated from μ^X that is independent of her private signals on other tasks. For instance, when Alice and Bob

[1] In this paper, we call the mechanism eq. (1) as the CA mechanism. However, CA mechanism can be used on non binary setting. [24] Under assumption 2.2, the CA mechanism reduces to eq. (1) that is first proposed in [5].

are Bayesian and use consistent strategies under assumption 2.2 and 2.1, Dasgupta and Ghosh [5] show CA mechanism in eq. (1) is strongly truthful. Intuitively, positive correlation assumption 2.2 guarantees that truthful reporting can maximize the chance of agreeing with the peer on the same task while avoiding agreeing on reports on other tasks. Furthermore, in appendix B we show CA mechanism merely has three Bayesian Nash equilibria which are 1) truth-telling opt_1, 2) flip the signal opt_2, or 3) uninformative reports (mixture between $\text{opt}_3, \text{opt}_4$).

2.2 Truthful convergence for learning agents

However, as we consider agents use a family of online learning to update their actions, standard solution concepts like Bayesian Nash equilibrium no longer apply. Additionally, online linear algorithms often conduct exploration, so we cannot hope agents always use telling the truth. For learning agents, our goal is to design mechanism $\mathcal{M}$ so that agents will *converge* to truthful reporting when they deploy certain learning behavior that goes beyond obliviously consistent strategies.

We now formalize the convergence of algorithms to truthful reporting. Because we want to elicit information without verification, it is information-theoretically impossible for us to separate permutation equilibrium from truthful equilibrium without any additional information[16]. Analogous to the strongly truthful guarantee, we ask if the limit of opt_t^X and opt_t^Y can be both truth-telling (opt_1) or the same permutation (opt_2).

Definition 2.3. An information elicitation mechanism $\mathcal{M}$ achieves *truthful convergence* for agents using algorithms A_1, A_2 if and only if the sequence of actions converges to both truth-telling or both flipping the reports.

$$\Pr\left\{\lim_{t\to+\infty} \text{opt}_t^X = \lim_{t\to+\infty} \text{opt}_t^Y = \text{opt}_1 \vee \lim_{t\to+\infty} \text{opt}_t^X = \lim_{t\to+\infty} \text{opt}_t^Y = \text{opt}_2\right\} = 1.$$

Note that definition 2.3 requires almost surely convergence which is very strong convergence result. Additionally, by using a single bit of information, e.g., whether the prior of 0 is larger than 1, we can remove the permutation and convert the agents' report back to their private signal regardless of whether they using the truth-telling opt_1 or flipping action opt_2.

3 Online Learning Algorithms

In this section, we explore candidates to model agents' learning behavior to replace Bayesian agents' oblivious consistent strategies in the literature.

We first show the conventional no-regret assumption is not enough in section 3.1. In particular, theorem 3.1 shows that for any sequential information elicitation mechanism with bounded rank, there exist no regret agents that collide at the beginning of the game and decide a sequence of uninformative actions.

Then in section 3.2, we introduce our reward-based online learning algorithm to model agents' learning behavior. Informally, a reward-based online learning algorithm uses the accumulated reward to decide each round's action using a fixed update function that is closely related to mean-based algorithms by [3]. We show the family of reward-based online learning algorithms contains several common no regret algorithms as special cases in theorem 3.5.

3.1 No regret online learning algorithms

One natural family of algorithms for Alice and Bob to use are no regret algorithms. Given a sequential information elicitation mechanism $\mathcal{M}$, we define $r_{i,t}$ be the payoff when Alice uses action opt_i and Bob's choices are unchanged. Then Alice's regret is $Reg^X(T) = \sum_{t\leq T} r_t - \max_i \sum_{t\leq T} r_{i,t}$, and we say her online learning algorithm is no regret if $\mathbb{E}[Reg^X(T)] = o(T)$.

In previous literature, one may use some no-regret as the behavior assumption for agents and show that a mechanism can achieve good results for agents ruled by this behavior assumption. However, the following theorem shows that we cannot have an information elicitation mechanism that achieves truthful convergence for no regret agents.

Theorem 3.1. *For any rank k information elicitation mechanism $\mathcal{M}$, there exist no-regret algorithms for Alice and Bob so that $\mathcal{M}$ does not have truthful convergence on the algorithms.*

The main idea of the proof is that no regret assumption cannot prevent Alice and Bob from colluding. In our counterexample, Alice and Bob decide on a no-regret sequence of reports regardless of their signals once the mechanism is announced. Standard multi-task information elicitation settings exclude such collusion by requiring agents to use consistent strategy.

Technically, we use probabilistic method to show the existence of a deterministic and no regret sequence of actions $(\text{opt}_t^X, \text{opt}_t^Y)_{t\geq 1}$ that consists of reporting 1 and 0 regardless of private signal $\text{opt}_3, \text{opt}_4$. The formal proof is in appendix C.

3.2 Reward-based Online Learning Algorithms

As we show in theorem 3.1, no regret assumption is not enough to tame agents' unrealistic behavior, and they may collude in sequential information elicitation mechanisms. In this paper, we introduce a general family of online learning algorithms that include several popular no regret algorithms as special cases (theorem 3.5).

In this section, we introduce our learning algorithms under the general full feedback bandit setting, and we will apply these algorithms in sequential information elicitation mechanisms later. For simplicity, we introduce a family $\mathcal{A}$ of learning algorithms on four actions $\{\text{opt}_1, \text{opt}_2, \text{opt}_3, \text{opt}_4\}$ which can be extended easily. Recall that the payoff of choosing option $\text{opt}_i, i \in [4]$ at round t is $r_{i,j}$ when others' choices are unchanged. We denote the accumulated payoffs of these four options as $R_{i,t} = \sum_{j=1}^{t} r_{i,j}$ for $i \in [4]$. Symmetrically, we use $S_{i,t}$ and $s_{i,t}$ to represent accumulated payoffs and the payoff of turning to choose option opt_i in the tth round for Bob. For example, for our specific peer prediction game using CA mechanism in eq. (1), $r_{1,t} = \mathbb{I}[x_t = \hat{y}_t] - \mathbb{I}[x_t = \hat{y}_{t-1}]$ and therefore, $R_{1,t} = \sum_{j=1}^{t}(\mathbb{I}[x_t = \hat{y}_t] - \mathbb{I}[x_t = \hat{y}_{t-1}])$ for Alice.

We consider a family of *reward-based online learning algorithms* $\mathcal{A}$ that use a *update function* $f : R^4 \to \triangle^3$, and choose opt_i with probability $f_i(R_1^t, R_2^t, R_3^t, R_4^t)$ for $i \in [4]$, where f_i is the ith coordinate of f. A such mechanism based on accumulated payoffs is denoted by A_f. We have three assumptions for the update function f, which are all very natural. The first two require that f is exchangeable[2] and preserves ordering.

Assumption 3.2 (Exchangeability of f)**.** For any $R_1, R_2, R_3, R_4 \in \mathbb{R}$ and an arbitrary permutation of them $R_{i_1}, R_{i_2}, R_{i_3}, R_{i_4}$, it holds for f that

$$f_{i_j}(R_1, R_2, R_3, R_4) = f_j(R_{i_1}, R_{i_2}, R_{i_3}, R_{i_4}), j \in [4].$$

Assumption 3.3 (Order preservation of f)**.** For any $R_1, R_2, R_3, R_4 \in \mathbb{R}$ and suppose that $R_{i_1}, R_{i_2}, R_{i_3}, R_{i_4}$ is a non-increasing order of them, for f we have

$$f_{i_1}(R_1, R_2, R_3, R_4) \geq f_{i_2}(R_1, R_2, R_3, R_4) \geq f_{i_3}(R_1, R_2, R_3, R_4) \geq f_{i_4}(R_1, R_2, R_3, R_4).$$

Finally we consider the update function f will choose an action when the accumulated payoff is much higher than other actions' accumulated payoffs.

Assumption 3.4 (Full exploitation of f)**.**

$$\lim_{R_1 - \max\{R_2, R_3, R_4\} \to +\infty} f_1(R_1, R_2, R_3, R_4) = 1.$$

In Appendix D.2, we prove that assumption 3.4 is a necessary condition for reward-based online learning algorithms to achieve no regret for any online decision problem.

Now we show the family of reward-based online learning algorithms $\mathcal{A}$ that use f satisfying assumptions 3.2 to 3.4 contains several classical no regret online learning algorithms [10, 12].

Theorem 3.5. *$\mathcal{A}$ contains FPL algorithm, Multiplicative Weights algorithm and replicator dynamics as special cases. Corresponding f's for them are listed as below:*

Multiplicative Weight algorithm $f_i^{hedge\ 1}(R_1, R_2, R_3, R_4) = \frac{e^{\beta R_i}}{\sum_{j\in[4]} e^{\beta R_j}}$ *for* $i \in [4]$

[2]This is also known as neurality in voting theory

FPL algorithm *Given a noise distribution* $\mathcal{N}^X$ *on scalars,* $f_i^{FPL}(R_1, R_2, R_3, R_4) = \Pr\left\{R_i + p_i = \max_{j\in[4]}\{R_j + p_j\} \middle| p_j \overset{\text{iid}}{\sim} \mathcal{N}^X, j \in [4]\right\}$ *for* $i \in [4]$.

If we focus on the binary peer prediction problem using CA mechanism, we can furthermore prove that replicator dynamics and linear updating multiplicative weight algorithm [2] are both in $\mathcal{A}$ in appendix E.

Remark 3.6. Our reward based online learning is very similar to the mean-based algorithm in [3] and both use the accumulated rewards to characterize the algorithm's choice. Additionally, like the mean-based algorithms, our reward based online algorithm may contain algorithms with regret, e.g., follow the leader.

However, the reward based online algorithm uses the identical update function across all rounds. Thus some no regret algorithm are not a reward based online algorithm, e.g., ϵ-greedy with time decreasing ϵ. Finally, if a mean-based algorithm uses the same update function in each round, then the update function needs to satisfy assumption 3.4.

4 Truthful Convergence of CA Mechanism on Reward-based Algorithms

Now we present our main result. We will show that CA mechanism can achieve truthful convergence if both agents use reward-based online learning algorithms $\mathcal{A}$. This convergence result suggests that the classical CA is robust even when agents deviate from Bayesian rational behavior and use a general family of online learning algorithms.

Theorem 4.1. *For a binary signal repeated peer prediction game under assumptions 2.1 and 2.2 and update functions* f *and* g *satisfying assumptions 3.2 to 3.4, CA mechanism in eq.* (1) *achieves truthful convergence when agents that use algorithms* A_f *and* A_g *that are in* $\mathcal{A}$.

Note that given agents' online learning algorithm and the payment function, the sequence of accumulated reward vector $(R_{1,t}, R_{2,t}, R_{3,t}, R_{4,t}, S_{1,t}, S_{2,t}, S_{3,t}, S_{4,t})_{t\geq 1}$ forms a stochastic process and we define $\mathcal{H}_t$ as the game history of the rewards and private signals in the first t round. Additionally, if the accumulated reward of truth-telling $R_{1,t}, S_{1,t}$ is much larger than the others', we can show Alice and Bob converge to the truth-telling by assumption 3.3. Thus, it is sufficient for us to track the evolution of accumulated reward vector. Though the process of accumulated reward vector is not a Markov chain because the payment function eq. (1) depends on reports in two rounds, we can still use ideas from semi-martingale to track the process.

Before proving our main result, we firstly present two properties of CA mechanism for our binary signal peer prediction problem. First Lemma 4.2 shows that the accumulated payoffs of uninformative actions $\text{opt}_3, \text{opt}_4$ $R_{3,t}, R_{4,t}, S_{3,t}$, and $S_{4,t}$ are always bounded.

Lemma 4.2. *Given the game defined in theorem 4.1, for any round* t*, the accumulated payoffs* $R_{3,t}, R_{4,t}, S_{3,t}$*, and* $S_{4,t}$ *for two agents are bounded by* $[-1, 1]$.

The second one, Lemma 4.3, tells us that the summations of accumulated payoffs of opt_1 and opt_2 is always fixed and is equal to the summations of accumulated payoffs of uninformative ones opt_3 and opt_4 for both agents. The proofs of these two lemmas are appendix F.1.

Lemma 4.3. *Given the game defined in theorem 4.1, for any round* t*, Alice has* $R_{1,t} + R_{2,t} = R_{3,t} + R_{4,t} = 0$*, and Bob has* $S_{1,t} + S_{2,t} = S_{3,t} + S_{4,t} = 0$.

We now sketch the proof that consists of four steps. Informally, using lemmas 4.2 and 4.3, the first step says that the uninformative actions opt_3 and opt_4 can not completely dominate other actions. Specifically, both agents can not choose an option between opt_3 and opt_4 with a probability larger than 0.75. With the first step, lemma 4.2 and lemma 4.3, we only need to focus on agents' reward of the truthtelling opt_1 shown in fig. 1. We partition the space into three types of events. *Good events* happen when R_1, S_1 are both very large or very small, *bad events* happen when one of R_1, S_1 is very large and one of them is very small, and *intermediate states* are the states between them. The second step removes the possibility that Alice and Bob continue using different reports opt_1 and opt_2. Therefore, we can always escape "bad events" and enter "intermediate states". The third step further shows that if the game is in an intermediate state, there exists a constant probability that the game will get into a good events that leads to truthful convergence in a constant number of rounds. Hence,

after the game enters "good events", which leaves us the final step: showing their action converges to either both truth telling opt_1 or flipping opt_2 truthful convergence.

Now we discuss each steps in more details but defer all the proofs to appendix F.

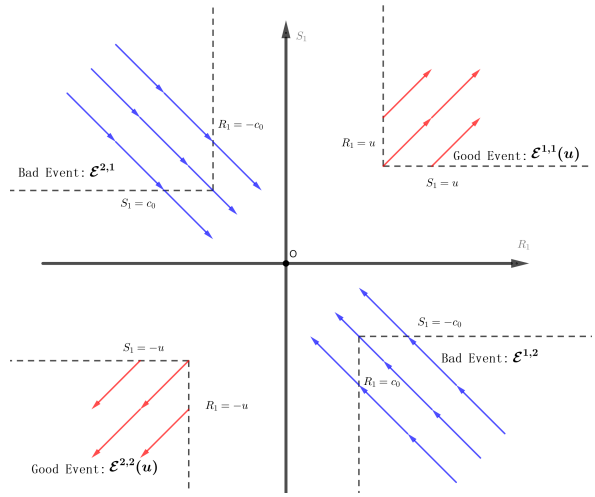

Figure 1: A schematic diagram of behaviors of $R_{1,t}$ and $S_{1,t}$.

4.1 Step 1: Choosing $\text{opt}_1, \text{opt}_2$ with Nonzero Probability

Combining Lemmas 4.2 and 4.3, we have the following Lemma 4.4, which completes step 1.

Lemma 4.4. *Given the game defined in theorem 4.1, for any round $t \geq 0$ and $i \in \{1, 2\}$, if $\sum_{j=1}^{t} r_{i,j} \geq 0$, the probability for Alice to choose opt_i is larger than $\frac{1}{4}$; if $\sum_{j=1}^{t} s_{i,t} \geq 0$, the probability for Bob to choose opt_i is larger than $\frac{1}{4}$.*

4.2 Step 2: Escaping Bad Events

Before introduce the formal statement of step 2, we define two "bad events". Given c_0 for all $t \geq 1$ $\mathcal{E}_t^{1,2} := \{R_{1,t} > c_0, \text{ and } S_{2,t} > c_0\}$ and $\mathcal{E}_t^{2,1} := \{R_{2,t} > c_0 \text{ and } S_{1,t} > c_0\}$. We will specify c_0 later. Intuitively, when c_0 is sufficiently large, $\mathcal{E}_t^{1,2}$ implies that Alice and Bob will choose $\text{opt}_1, \text{opt}_2$ respectively with a probability close to 1 in following rounds. For simplicity, we treat each event $\mathcal{E}$ as an indicator function, i.e., $\mathcal{E}$ happens if and only if $\mathcal{E} = 1$. In order to prove that these two bad events cannot go on forever, we want to show that when $\mathcal{E}_t^{1,2}$ happens, $R_{1,t} - R_{2,t}$ will tend to decrease at a rapid rate. Therefore, Alice will eventually deviate from opt_1 to choose opt_2 with a relatively high probability.

By assumption 3.4 and lemma 4.2, given $\delta > 0$ there exists a constant c_1 such that when $R_{1,t} > c_1$, Alice chooses opt_1 with probability larger than $1 - \delta$; when $S_{2,t} > c_1$, Bob chooses opt_2 with probability larger than $1 - \delta$. Let $\gamma_1 = P_{X,Y}(1,1) + P_{X,Y}(0,0) - P_{X,Y}(1,0) - P_{X,Y}(0,1)$ $\gamma_2 = (P_{X,Y}(1,1) - P_{X,Y}(0,0))^2 - (P_{X,Y}(1,0) - P_{X,Y}(0,1))^2$.

Lemma 4.5. *Given the game defined in theorem 4.1, there exists a $\delta > 0$ and corresponding c_1 such that for any round t, $\mathbb{E}[r_{1,t+1} - r_{2,t+1} | r_{t-1} > c_1 + 1] \geq \frac{\gamma_1 - \gamma_2}{2} > 0$.*

Given such δ and c_1 in lemma 4.5, we set $c_0 = c_1 + \lceil \frac{1000}{\gamma_1 - \gamma_2} \rceil + 1$. Because in each round $R_{1,t}, R_{2,t}$ and $S_{1,t}, S_{2,t}$ vary by at most 1, if $\mathcal{E}_t^{1,2}$ happens Alice chooses opt_1 and Bob chooses opt_2 with probability larger than $1 - \delta$ independently for the next $\lceil \frac{1000}{\gamma_1 - \gamma_2} \rceil + 1$ rounds. Similar argument holds for $\mathcal{E}_t^{2,1}$. We use the above observation to show lemma 4.6.

Lemma 4.6. *Given the game defined in theorem 4.1, for all t and history $\mathcal{H}_t \in \mathcal{E}_t^{1,2}$, we have* $\mathbb{E}\left[\sum_{j=1}^{\lceil \frac{1000}{\gamma_1 - \gamma_2} \rceil + 1} (r_{1,t+j} - r_{2,t+j}) \middle| \mathcal{H}_t\right] \leq -100.$

This lemma formalizes the blue arrows in Fig. 1. With this lemma, we get the main result of step two.

Lemma 4.7. *Given the game defined in theorem 4.1,* $\Pr\left\{\limsup_{t\to\infty} \overline{\mathcal{E}_t^{1,2} \vee \mathcal{E}_t^{2,1}} = 1\right\} = 1.$

If we treat $R_{1,t} - R_{2,t}$ as money of Alice, Lemma 4.7 is similar to the gambler's ruin problem [7]. More specifically, $R_{1,t} - R_{2,t}$ has a negative expected growth each $\lceil \frac{1000}{\gamma_1 - \gamma_2} \rceil + 1$ rounds by Lemma 4.6, so $R_{1,t} - R_{2,t}$ will always become small enough to escape $\mathcal{E}^{1,2}$.

4.3 Step 3: From Intermediate States to Good Events

We define a series of "good events" at first. For all $u \in \mathbb{N}^+$ and $t \geq 1$, $\mathcal{E}_t^{1,1}(u) := \{R_{1,t} \geq u, \text{ and } S_{1,t} \geq u\}$ and $\mathcal{E}_t^{2,2}(u) := \{R_{2,t} \geq u \text{ and } S_{2,t} \geq u\}$.

To the end, we want to show that $\vee_{t\in\mathbb{N}^+}\mathcal{E}_t^{1,1}(u)$ and $\vee_{t\in\mathbb{N}^+}\mathcal{E}_t^{2,2}(u)$ happen with probability 1 for any $u \in \mathbb{N}^+$. Formally, we claim Lemma 4.8.

Lemma 4.8. *Given the game defined in theorem 4.1, for all u there exists λ_u so that for any T with history $\mathcal{H}_T \in \overline{\mathcal{E}_T^{1,2} \vee \mathcal{E}_T^{2,1}}$, we have*

$$\Pr\left\{\left(\vee_{i=T}^{T+4(u+c_0)+100}\mathcal{E}_t^{1,1}(u)\right) \vee \left(\vee_{i=T}^{T+4(u+c_0)+100}\mathcal{E}_t^{2,2}(u)\right) = 1 \middle| \mathcal{H}_T\right\} \geq \lambda_u,$$

This lemma generally says that when the agents are in an intermediate state that $\overline{\mathcal{E}_T^{1,2} \vee \mathcal{E}_T^{2,1}} = 1$, they have a constant probability λ_u to get into a good event in the next $4(u + c_0) + 100$ rounds. In order to prove this lemma, we define a nice event $\mathcal{P}_T$ such that $\mathcal{P}_T$ happens with probability larger than λ_u and $\mathcal{P}_T$ implies the good events happen in no more than $4(u + c_0) + 100$ rounds. Formally, event $\mathcal{P}_T$ is defined as the following: First, for $j = T, \ldots, T_1$ until some round $T_1 \geq T$ such that $S_{1,T_1} \geq 0$, $x_{j+1} = 1 - \hat{y}_j, y_{j+1} = 1 - \hat{x}_j$, Alice uses strategy opt_1, Bob uses strategy opt_2. Then, Alice and Bob uses strategy opt_1 for $4u + 50$ rounds and signals are generated as $x_{j+1} = y_{j+1} = -x_j$ for $j \geq T_1$. We can use Lemma 4.4 to prove that $\lambda_u = (\min_{i,j\in\{0,1\}}\{P_{X,Y}(i, j)\} \times 0.25^2)^{100+4(u+c_0)}$ is a feasible lower bound.

4.4 Step 4: Good Events Lead to Truthful Convergence

Similar to step 2, we have lemma 4.9 to show that $\mathcal{E}_t^{1,1}(u)$ can lead to increasing of $R_{1,t} - R_{2,t}$ at first. The idea is completely similar to Lemma 4.6 and it formalize the red arrows in Fig. 1.

Lemma 4.9. *Given the game defined in theorem 4.1, for all $u > 2c_0 + 1$ and t if $\mathcal{H}_t \in \mathcal{E}_t^{1,1}\left(\lfloor \frac{u}{2} \rfloor\right)$, we have* $\mathbb{E}\left[\sum_{j=1}^{\lceil \frac{1000}{\gamma_1-\gamma_2} \rceil+1} (r_{1,t+j} - r_{2,t+j}) \middle| \mathcal{H}_t\right] \geq 100,$

Using Lemma 4.9, we are able to prove that for any ε, there exists a u such that when $\mathcal{E}_t^a(u)$ or $\mathcal{E}_t^b(u)$ happens, Alice and Bob will tend to choose $(\text{opt}_1, \text{opt}_1)$ or $(\text{opt}_2, \text{opt}_2)$ with increasingly higher probability. Formally, we propose Lemma 4.10.

Lemma 4.10. *Given the game defined in theorem 4.1, for all $\epsilon > 0$ there exists $u \in \mathbb{N}^+$ such that given a history $\mathcal{H}_T \in \mathcal{E}_T^{1,1}(u) \vee \mathcal{E}_T^{2,2}(u)$, we have*

$$\Pr\left\{\forall i \in \mathbb{N}, \mathcal{E}^{1,1}_{T+\left(\lceil \frac{1000}{\gamma_1-\gamma_2} \rceil+1\right)i}\left(\lfloor \frac{u}{2} \rfloor + i\right) \vee \mathcal{E}^{2,2}_{T+(\lceil \frac{1000}{\gamma_1-\gamma_2} \rceil+1)i}\left(\lfloor \frac{u}{2} \rfloor + i\right) = 1 \middle| \mathcal{H}_T\right\} \geq 1 - \varepsilon,$$

We can verify that $M_i = R_{1,T+i\left(\lceil \frac{1000}{\gamma_1-\gamma_2} \rceil+1\right)} - 100i$ is a sub-martingale, and use Azuma-Hoeffding inequality to prove lemma 4.10.

5 Simulations

In this section, we run simulations on the CA mechanism with various learning algorithms: the Hedge algorithms, follow the perturbed leader, follow the leader, and ϵ-greedy. We ran the process for 400 times for 800 rounds on each algorithm. The converge proportion in round t is defined

as the percentage of the simulations where both agents report truthfully opt_1 (or both use opt_2) in all the subsequent rounds. We show that all our algorithms converge to truth-telling. Note that all reward-based online learning algorithms (the Hedge algorithms, follow the perturbed leader, and follow the leader) exhibit truthful convergence that aligns with our theoretical result, theorem 4.1. We also find that the ϵ-greedy with time decreasing ϵ, which is not a reward-based online learning algorithm, also shows truthful convergence. This suggests that the CA mechanism may have truthful convergence beyond reward-based online learning algorithms.

In our simulations, we use the following private signal distribution that satisfies assumption 2.2: $P_{X,Y}(0,0) = P_{X,Y}(1,1) = 0.4, P_{X,Y}(1,0) = P_{X,Y}(0,1) = 0.2$. Moreover, Alice and Bob are using the same learning algorithms in our simulations that are listed below: First, Follow the Leader algorithm (FTL) chooses i with probability proportional to $\mathbb{I}[R_i = \max\{R_1, R_2, R_3, R_4\}]$. Follow the perturbed leader (FPL*, where * can be 1, 4 or 8) adds a uniform random noise between 0 and $*$ and choose action i with probability $\Pr\left\{R_i + p_i = \max_{j\in[4]}\{R_j + p_j\} \middle| p_j \overset{\text{iid}}{\sim} \mathcal{U}[0,*]\right\}$. We consider FPL1, FPL4, and FPL8. Hedge algorithm 1 choose i with probability proportional to $3^{R_i/2}$ that is an implementation by choosing $\epsilon = 0.5$ for the multiplicative weights algorithm of Arora et al. [2]. Hedge algorithm 2 chooses i with probability proportional to e^{R_i} that is an implementation by choosing $\beta = 1$ of exponentially weighted averaged forecaster introduced by Freund and Schapire [10]. Finally, ϵ-greedy algorithm uses time varying $\epsilon = \frac{1}{(t+1)^2}$ at round t. Note that ϵ-greedy is not in $\mathcal{A}$ but still achieves truthful convergence.

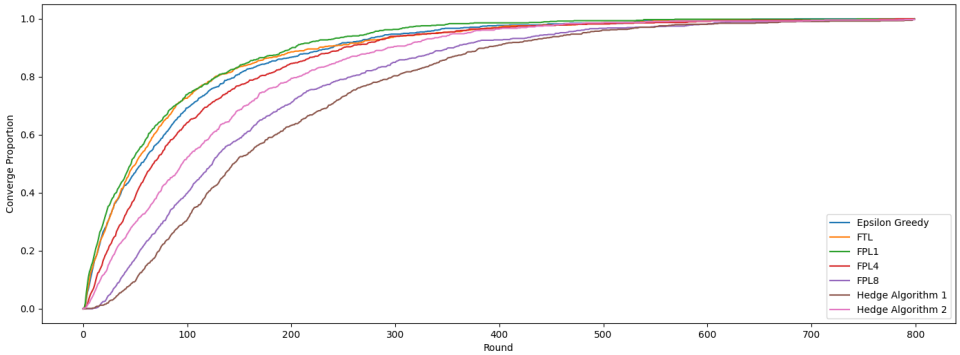

Figure 2: Convergence Rates of Learning Algorithms in CA.

Figure 2 shows that all algorithms experimented converge to truthful reporting. Moreover, although FTL is generally not no-regret, CA mechanism still works well with it. Additionally, we observe that when an algorithm explores less (e.g. FPL4 vs FPL8), it converges faster, but very little exploration does not further improve the convergence rate (e.g. FTL vs FPL1).

6 Conclusions

In this paper, we study sequential peer prediction with learning agents and prove that the notion of no-regret alone is not sufficient for truthful convergence. We then define a family of reward-based learning algorithms and show that the CA mechanism is able to achieve truthful convergence when agents use algorithms in this family. Finally, we give a discussion on the converge rates of different learning agents based on simulations.

This is the first theoretical study on peer prediction with learning agents. There are many open problems and future directions to extend this work. For the private signals, we believe the same techniques can be used to prove truthful convergence of the CA mechanism used on reward-based learning agents when signals are generated by a Markov chain given some assumptions on the transition matrix. Moreover, this work is only restricted to binary signals and it is still an open problem whether there exists a mechanism for non-binary settings that can promise truthful convergence. For the learning agents, one could consider a more general family of learning algorithms such as when f is time-varying.

References

[1] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

[2] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

[3] M. Braverman, J. Mao, J. Schneider, and S. M. Weinberg. Selling to a no-regret buyer. *CoRR*, abs/1711.09176, 2017. URL `http://arxiv.org/abs/1711.09176`.

[4] M. K. Camara, J. D. Hartline, and A. Johnsen. Mechanisms for a no-regret agent: Beyond the common prior. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–270. IEEE, 2020.

[5] A. Dasgupta and A. Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330, 2013.

[6] Y. Deng, J. Schneider, and B. Sivan. Strategizing against no-regret learners. *Advances in neural information processing systems*, 32, 2019.

[7] R. A. Epstein. *The theory of gambling and statistical logic*. Academic Press, 2012.

[8] B. Faltings, R. Jurca, P. Pu, and B. D. Tran. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing*, 2014.

[9] W. Feller. *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons, 2008.

[10] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[11] J. Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–139, 1957.

[12] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[13] V. Kamble, N. Shah, D. Marn, A. Parekh, and K. Ramachandran. Truth serums for massively crowdsourced evaluation tasks. *arXiv preprint arXiv:1507.07045*, page 96, 2015.

[14] Y. Kong. Dominantly truthful multi-task peer prediction with a constant number of tasks. *CoRR*, abs/1911.00272, 2019. URL `http://arxiv.org/abs/1911.00272`.

[15] Y. Kong and G. Schoenebeck. Equilibrium selection in information elicitation without verification via information monotonicity. *arXiv preprint arXiv:1603.07751*, 2016.

[16] Y. Kong and G. Schoenebeck. A framework for designing information elicitation mechanisms that reward truth-telling. *CoRR*, abs/1605.01021, 2016. URL `http://arxiv.org/abs/1605.01021`.

[17] Y. Liu and Y. Chen. Surrogate scoring rules and a dominant truth serum for information elicitation. *arXiv preprint arXiv:1802.09158*, 2018.

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

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

[20] D. Prelec, H. S. Seung, and J. McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

[21] G. Radanovic and B. Faltings. Incentives for truthful information elicitation of continuous signals. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 770–776, 2014.

[22] G. Schoenebeck and F. Yu. Learning and strongly truthful multi-task peer prediction: A variational approach. *CoRR*, abs/2009.14730, 2020. URL `https://arxiv.org/abs/2009.14730`.

[23] G. Schoenebeck, F.-Y. Yu, and Y. Zhang. Information elicitation from rowdy crowds. In *Proceedings of the Web Conference 2021*, pages 3974–3986, 2021.

[24] V. Shnayder, A. Agarwal, R. M. Frongillo, and D. C. Parkes. Informed truthfulness in multi-task peer prediction. *CoRR*, abs/1603.03151, 2016. URL `http://arxiv.org/abs/1603.03151`.

[25] V. Shnayder, R. Frongillo, and D. C. Parkes. Measuring performance of peer prediction mechanisms using replicator dynamics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016.

[26] J. Witkowski and D. C. Parkes. A robust bayesian truth serum for small populations. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[27] S. Zheng, F. Yu, and Y. Chen. The limits of multi-task peer prediction. *CoRR*, abs/2106.03176, 2021. URL `https://arxiv.org/abs/2106.03176`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work focuses on pure theoretical problems, thus it does not have any negative social impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Our code is included in supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Training details are specified in section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We give error bars in appendix G instead of in section 5 for clearness of fig. 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We illustrate our hardware details in appendix G.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]