Surrogate Scoring Rules and a Dominant Truth Serum for Information Elicitation

YANG LIU, Harvard University
YILING CHEN, Harvard University

We study information elicitation without verification (IEWV) and ask the following question: *Can we achieve truthfulness in dominant strategy in IEWV?* This paper considers two elicitation settings. The first setting is when the mechanism designer has access to a random variable that is a noisy or proxy version of the ground truth, with *known* biases. The second setting is the standard peer prediction setting where agents' reports are the only source of information that the mechanism designer has. We introduce *surrogate scoring rules* (SSR) for the first setting, which use the noisy ground truth to evaluate quality of elicited information, and show that SSR achieve truthful elicitation in dominant strategy. Built upon SSR, we develop a multi-task mechanism, *dominant truth serum* (DTS), to achieve truthful elicitation in dominant strategy when the mechanism designer only has access to agents' reports (the second setting). The method relies on an estimation procedure to accurately estimate the average bias in the reports of other agents. With the accurate estimation, a random peer agent's report serves as a noisy ground truth and SSR can then be applied to achieve truthfulness in dominant strategy. A salient feature of SSR and DTS is that they both quantify the quality or value of information despite lack of ground truth, just as proper scoring rules do for the *with* verification setting. Our work complements both the strictly proper scoring rule literature by solving the case where the mechanism designer only has access to a noisy or proxy version of the ground truth, and the peer prediction literature by achieving truthful elicitation in dominant strategy.

1 INTRODUCTION

Eliciting private information about events and tasks from human agents is an almost universal problem in our society when there is uncertainty and asymmetric information. For example, financial experts are asked to make predictions on whether the S & P 500 index will go up next week, seismologists attempt to predict future earthquake activities, students doing peer grading are asked to assess the correctness of others' solutions. From the perspective of incentive alignment, an important goal of a mechanism designer is to incentivize agents to truthfully reveal their private information so that high-quality information is obtained.[1]

Incentive design for information elicitation has been extensively studied. However, proposed solutions have achieved different notions of truthfulness depending on which one of the following two categories an elicitation problem falls into: information elicitation *with* verification and information elicitation *without* verification.

Information elicitation *with* verification refers to settings where the mechanism designer will have access to the ground truth at some point (e.g. whether S & P 500 index actually went up and whether a predicted earthquake actually happened) and can use the ground truth to evaluate the quality of elicited information. Strictly proper scoring rules [Brier, 1950, Gneiting and Raftery, 2007, Matheson and Winkler, 1976a, Murphy and Winkler, 1970] have been proposed and studied to incentivize truthful reporting of private predictions (probability assessment of event outcome). Facing a strictly proper scoring rule, an agent strictly maximizes his expected score by truthfully revealing his prediction. Hence, truthfulness in dominant strategy is achieved here. Moreover, the score of an agent measures the quality of his prediction.

[1] Many also consider incentivizing both effort exertion to improve the quality of private information and truthful information revelation.

Information elicitation *without* verification refers to settings where the mechanism designer does not have access to the ground truth (e.g. whether a student's homework solution is correct is unavailable as otherwise peer grading is not needed). An elegant family of incentive mechanisms [Dasgupta and Ghosh, 2013, Jurca and Faltings, 2006, Miller et al., 2005, Prelec, 2004, Radanovic and Faltings, 2013, Witkowski and Parkes, 2012], collectively called peer prediction, has been developed for this category to elicit private signals (e.g. a peer grader's judgement on the correctness of a homework solution). Peer prediction leverages the correlation of agents' private signals and scores an agent's report based on how it compares with the reports from other agents. Agents play a Bayesian game in peer prediction and there exists a Bayesian Nash Equilibrium (BNE) where every agent reports truthfully. Truthfulness at BNE is clearly a weaker incentive guarantee than truthfulness in dominant strategy as the former is sensitive to errors and mistakes of other agents. Peer prediction also suffers from multiplicity of equilibria as there are other non-truthful BNE.[2] In experiments, the truthful equilibrium was often not reached [Gao et al., 2014]. In addition, the peer prediction score of an agent measures the correlation of his report with those of others, but it doesn't reflect the quality of the agent's report.

We focus on information elicitation *without* verification in this paper and ask:

Can we achieve truthfulness in dominant strategy in information elicitation without verification?

We provide a positive answer to this question for binary information elicitation. Moreover, in our proposed mechanisms, the score of an agent measures the quality of his report as in information elicitation with verification.

Two specific information elicitation *without* verification settings are considered in this paper. The first setting is when the mechanism designer has access to a random variable that is a noisy or proxy version of the ground truth, with *known* biases. The second setting is the standard peer prediction setting where agents' reports are the only source of information that the mechanism designer has.

For the first setting, we introduce *surrogate scoring rules* (SSR), which use the noisy ground truth to evaluate quality of elicited information, and show that SSR achieve truthful elicitation in dominant strategy without access to the ground truth. These surrogate scoring rules are inspired by the use of surrogate loss functions in machine learning [Angluin and Laird, 1988, Bylander, 1994, Natarajan et al., 2013, Scott, 2015, Scott et al., 2013] and they remove bias from the noisy random variable such that in expectation a report is as if evaluated against the ground truth.

Built upon SSR, in the second setting where the mechanism designer only has access to agents' reports, we develop a multi-task mechanism, *dominant truth serum* (DTS), to achieve truthful elicitation in dominant strategy. The method relies on an estimation procedure to accurately estimate the average bias in the reports of other agents. With the accurate estimation, a random peer agent's report serves as a noisy ground truth and SSR can then be applied to achieve truthfulness in dominant strategy. We further show that DTS can be extended to achieve truthful elicitation in dominant strategy in settings where agents' signals are correlated conditioned on the ground truth. This allows us to overcome the failure of classic peer prediction in the presence of weak, correlated signals, observed by [Gao et al., 2016].

A salient feature of SSR and DTS is that they both quantify the quality or value of information despite lack of ground truth, just as proper scoring rules do for the *with* verification setting. In this sense, SSR and DTS unify the *with* and *without* verification settings by always rewarding agents

[2]Several recent peer prediction mechanisms [Dasgupta and Ghosh, 2013, Kong and Schoenebeck, 2016a, Shnayder et al., 2016] have made truthful equilibrium focal in the sense that it leads to the highest expected payoff to agents among all equilibria. But there is at least one other equilibrium that gives the same expected payoff to agents.

based on their quality of information (and hence achieving truthfulness in dominant strategy). In addition, SSR and DTS apply to both signal and prediction elicitation.

The rest of the paper is organized as follows. In Section 2, we describe our model and the two elicitation settings. Section 3 introduces SSR for the first elicitation setting. Section 4 details DTS for the second elicitation setting. Then, in Section 5, we extend DTS to correlated signals. Section 6 concludes with discussions on some additional directions.

1.1 Related work

The most relevant literature to our paper is *strictly proper scoring rule* and *peer prediction.* Scoring rules were developed for eliciting truthful prediction (probability) [Brier, 1950, Gneiting and Raftery, 2007, Jose et al., 2006, Matheson and Winkler, 1976b, Savage, 1971, Winkler, 1969]. For example, the pioneer works [Brier, 1950] proposed Brier scoring to verify the qualities of forecasts. A full characterization result is given for strictly proper scoring rules in [Gneiting and Raftery, 2007]. Our work complements the literature of scoring rule by proposing surrogate scoring rules to cope with the case when there is only access to a noisy or proxy version of the ground truth.

Our surrogate scoring rules can be viewed as a class of proxy scoring rules defined in [Witkowski et al., 2017]. [Witkowski et al., 2017] proposed to design proper proxy scoring rules by using an unbiased proxy random variable as the ground truth in a strictly proper scoring rule. Our proxy random variable is biased. We focus on designing a surrogate scoring function that removes bias in any noisy ground truth, and thus can deal with a more general setting.

Motivated by the complexity of reporting the probability for events with large outcome space, property elicitation, such as for mean, variance etc of probability distributions, has been studied more recently [Frongillo and Kash, 2015, Lambert et al., 2008]. In this paper we do not cover the property elicitation setting.

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. This line of research started with the celebrated *Bayesian Truth Serum* work [Prelec, 2004], where a surprisingly popular answer methodology is shown to be able to incentivize agents to truthfully report even they believe they hold the minority answer (but more likely to be true in their own opinion). The seminal work [Miller et al., 2005] established that strictly proper scoring rule [Gneiting and Raftery, 2007] can be adopted in the peer prediction setting for eliciting truthful reports (but the mechanism designer need to know details of agents' model); a sequence of followed up works have been done to relax the assumptions that have been imposed therein [Radanovic and Faltings, 2013, Witkowski and Parkes, 2012]. More recently, [Dasgupta and Ghosh, 2013, Witkowski et al., 2013] formally introduced and studied an effort sensitive model for binary signal data elicitation. Particularly [Dasgupta and Ghosh, 2013] proposed a multi-task peer prediction mechanism that can help remove undesirable equilibria that lead to low quality reports. These results are further strengthened and extended to a non-binary signal setting in [Kong and Schoenebeck, 2016a, Shnayder et al., 2016]. An information theoretical framework for studying this type of elicitation problems has been proposed in [Kong and Schoenebeck, 2016a,b]. [Kong and Schoenebeck, 2016b] also proposed a conceptual mechanism to achieve truthfulness in dominant strategy but it requires infinite number of tasks. Our work is the first one to establish truthfulness in dominant strategy with a finite number of tasks for a peer prediction setting.

It is worth to mention that our work borrows ideas from the machine learning literature on learning with noisy data [Bylander, 1994, Menon et al., 2015, Natarajan et al., 2013, Scott, 2015, van Rooyen and Williamson, 2015]. From a high level's perspective, our goal in this paper aligns with the goal in learning from noisy labels - both aim to evaluate a prediction when the ground truth is missing, but instead a noisy signal of the ground truth is available. Our work addresses additional challenge that the error rate of the noisy signal remains unknown a priori.

2 OUR MODEL

Prior work on information elicitation typically considers one of two types of information: private signals and private beliefs (i.e. probabilistic predictions). In this paper, we develop mechanisms that apply to both cases. Hence, in Section 2.1, we first introduce our model of information and notations to unify these two types of information. Then, Section 2.2 and 2.3 will introduce our model and design goals for two elicitation settings without accessing the ground truth. In the first setting (Section 2.2), the mechanism designer has access to a random variable that is a noisy version of the ground truth with *known* bias. The second setting (Section 2.3) is the typical setting in the peer prediction literature where the mechanism designer only has access to reports from agents. For both settings, our design goal is to achieve truthful elicitation in dominant strategy.

2.1 Model of information

Suppose we are interested in eliciting information about a binary event $y \in \{0,1\}$ from a set of human agents, indexed by $[N] := \{1,2,\dots,N\}$. The realization of the event y has prior distribution $\mathcal{P}_0 := \Pr[y=0]$, $\mathcal{P}_1 = \Pr[y=1]$. We assume a non-trivial prior distribution that $0 < \mathcal{P}_0, \mathcal{P}_1 < 1$. Each of the N agents holds a noisy observation of y, denoting as s_i. Agents' observations are conditionally independent: $\Pr[s_i, s_j|y] = \Pr[s_i|y] \cdot \Pr[s_j|y]$. We short-hand the following error rates:

$$e_{1,i} := \Pr[s_i = 0|y=1],\ e_{0,i} := \Pr[s_i = 1|y=0],$$

i.e., $e_{1,i}, e_{0,i}$ are the "error probabilities" or the subjective "human bias" of agent i's observation for y. We do *not* assume homogeneous agents, that is we allow agents to have different $e_{1,i}, e_{0,i}$. The error rates can also model subjectivity in agents' private belief and observation. Based on signal s_i, each agent can form a posterior belief about y, denoting as $p_i := \Pr[y=1|s_i]$. When there are multiple tasks, we assume the prior, $\mathcal{P}_0$ and $\mathcal{P}_1$, and each agent's error rates, $e_{1,i}$ and $e_{0,i}$, are homogeneous across tasks.

We consider eliciting either s_i or p_i from agents,[3] but we are not able to access the ground truth y to verify the reported information from agents. We call the elicitation of these two types of information as *signal elicitation* and *prediction elicitation* respectively. The literature on proper scoring rules has been primarily considering prediction elicitation, while the literature on peer prediction has focused on signal elicitation. To unify both types of information, we denote the information space as $\mathcal{I}$, and each agent's information as I_i. For signal elicitation, $I_i = s_i$ and $\mathcal{I} = \{0,1\}$. For prediction elicitation, $I_i = p_i$ and $\mathcal{I} = [0,1]$. Denote agent i's report to a mechanism as $a_i \in I$. Then, for for signal elicitation and prediction elicitation, we have $a_i : s_i \to \{0,1\}$ and $a_i : p_i \to [0,1]$ respectively.

2.2 Model of elicitation with noisy ground truth

The first elicitation setting is when the mechanism designer has access to the realization of a binary random variable $z \in \{0,1\}$, which is a noisy or proxy version of the ground truth with known bias. The bias of z is again captured by the error rates:

$$e_{1,z} := \Pr[z=0|y=1],\ e_{0,z} := \Pr[z=1|y=0].$$

[3]For clarity of presentation, we assume that we elicit the same type of information from all agents. But our work can be extended to a mixed elicitation setting.

We cannot expect to do much if z is independent of y and hence assume that z and y are stochastically relevant, an assumption commonly adopted in the information elicitation literature [Miller et al., 2005].

Definition 2.1. z and y are *stochastically relevant* if there exists an $s \in \{0, 1\}$ such that

$$\Pr[y = s|z = 0] \neq \Pr[y = s|z = 1].$$

The following lemma shows that the stochastic relevance requirement directly translates to a constraint on the error rates, that is, $e_{1,z} + e_{0,z} \neq 1$.

Lemma 2.2. *y and z are stochastically relevant if and only if $e_{1,z} + e_{0,z} \neq 1$. And y and z are stochastically irrelevant (independent) if and only if $e_{1,z} + e_{0,z} = 1$.*

We consider both signal elicitation and prediction elicitation. Our goal is to design a scoring function $\varphi(a_i, z) : I \times \{0, 1\} \to \mathbb{R}_+$ such that when agent i's report a_i is evaluated against the realization of the noisy ground truth z, agent i strictly maximizes his expected score by reporting his true information $a_i = I_i$. Formally, we want to achieve

$$\textbf{Strict properness with noisy ground truth: } \mathbb{E}[\varphi(I_i, z)] > \mathbb{E}[\varphi(a_i, z)],\ \forall a_i \neq I_i, \tag{1}$$

where the expectation is taken with respect to the agent's subjective belief $\Pr[z|s_i]$.

It's worth noting that prediction elicitation in this model, as in the literature, doesn't need agents to form their prediction p_i according to the signal model described in Section 2.1, and instead only requires that agents each have a subjective belief p_i. For simplicity of presentation, we use the signal model in Section 2.1 throughout the paper.

2.3 Model of elicitation with peer reports

The second elicitation setting is close to the standard setting considered in the peer prediction literature. Here the only source of information that the mechanism designer has access to is the agents' reports. In this setting, we assume "differentiable" prior such that $\mathcal{P}_0 \neq \mathcal{P}_1$. This is a technical condition that we need. In practice, with multiple tasks, we can perturb the priors by adding some known tasks to (or deleting task from) one of the two classes.

2.3.1 Informativeness of signals. We adopt the standard assumption requiring that each agent's signal s_i and the ground truth y are stochastically relevant, i.e. $e_{1,i} + e_{0,i} \neq 1$. Moreover, we assume that signals are Bayesian informative of y, that is, $\Pr[y = s|s_i = s] > \Pr[y = s]$, $\forall i, s$. In other words, the posterior probability for $y = s$ is greater than the prior probability of $y = s$ if an agent receives a signal s. This assumption has been adopted by [Dasgupta and Ghosh, 2013, Miller et al., 2005], and has been shown by [Liu and Chen, 2017] to be equivalent to a constraint on the error rates of agent signals.

Lemma 2.3. *(Lemma 2.1 in [Liu and Chen, 2017]) s_i is Bayesian informative of y if and only if $e_{1,i} + e_{0,i} < 1$.*

The case $e_{1,i} + e_{0,i} > 1$ can be viewed as negative Bayesian informative as always flipping one's observation will return a Bayesian informative signal. In this paper we focus only on settings where every agent's signal is Bayesian informative of y. But this assumption is to simplify presentation as our results hold for negative Bayesian informative agents, as well as for a mixed population.

2.3.2 Knowledge of agents and the mechanism designer. For signal elicitation, agents do not need to be aware of their own values of $e_{1,i}$ and $e_{0,i}$, but they need to know that their own signals are Bayesian informative and the error rates satisfy $e_{1,i} + e_{0,i} < 1$. For prediction elicitation, agents need to know their own values of $e_{1,i}$ and $e_{0,i}$ to form their predictions. The above knowledge does not need to be common knowledge among agents. This is because we seek to achieve dominant strategy truthfulness and hence agents are as if not playing a Bayesian game.

The mechanism designer knows the priors of ground truth $\mathcal{P}_0, \mathcal{P}_1$, but not the realized y, nor the private information s_i's or p_i's. Further the designer does not know any of the biases of agents' information, $e_{1,i}$ and $e_{0,i}$, a-priori. This can capture scenarios when we know the prior distribution of answers for a certain type of questions, but not specifically the answer for a particular question, nor how human agents make wrong prediction of such events. In our proposed mechanism, we will show how to infer the true error rates of agents' reports.

2.3.3 Design goal. We again aim to achieve truthful elicitation in dominant strategy for both signal elicitation and prediction elicitation. Denote by $\mathcal{D}_{-i}$ the set of information collected from agents $j \neq i$, possibly for multiple tasks, i.e. $\mathcal{D}_{-i} := \{a_j\}_{j \neq i}$ For this part, our goal is to find a reference signal $z_{\mathcal{D}_{-i}} \in \{0,1\}$, as a function of $\mathcal{D}_{-i}$, and a scoring function

$$\varphi_{\mathcal{D}_{-i}}(a_i, z) : \mathcal{I} \times \{0,1\} \rightarrow \mathbb{R}_+,$$

also as a function of $\mathcal{D}_{-i}$, such that it is a dominant strategy for agent i to truthfully report his private information:

Truthfulness in dominant strategy with peer reports: $I_i = \text{argmax}_{a_i} \mathbb{E}[\varphi_{\mathcal{D}_{-i}}(a_i, z_{\mathcal{D}_{-i}})]$, (2)

where the expectation is taken with respect to the agent's subjective belief $\Pr[z_{\mathcal{D}_{-i}} | s_i]$. When it's clear in context, we'll use z as shorthand for $z_{\mathcal{D}_{-i}}$ and φ for $\varphi_{\mathcal{D}_{-i}}$.

3 SURROGATE SCORING RULES

In this section we propose *surrogates scoring rules* (SSR) to address the question of achieving strict properness when only having access to a noisy ground truth. SSR is built upon strictly proper scoring rules. Thus, we first briefly discuss strictly proper scoring rules. Then, we formally define SSR, prove its strict properness and give an implementation of it.

3.1 Strictly proper scoring rules

As mentioned earlier, when ground truth is available, strictly proper scoring rules were designed to elicit predictions, i.e. the p_i's. A scoring function $S : [0,1] \times \{0,1\} \rightarrow \mathbb{R}_+$ is strictly proper if and only if

$$\mathbb{E}[S(p_i, y)] > \mathbb{E}[S(\tilde{p}_i, y)], \ \forall \tilde{p}_i \neq p_i,$$

where the expectation is taken with respect to the agent's belief $\Pr[y|s_i]$. There is a rich family of strictly proper scoring rules, including Brier ($S(p_i, y) = 1 - (p_i - y)^2$), logarithmic and spherical scoring rules [Gneiting and Raftery, 2007].

Though not enjoying much attention in the literature, the above idea of defining strictly proper scoring rules also applies to signal elicitation, that is to design a function $S : \{0,1\}^2 \rightarrow \mathbb{R}_+$ such that

$$\mathbb{E}[S(s_i, y)] > \mathbb{E}[S(\tilde{s}_i, y)], \ \forall \tilde{s}_i \neq y_i,$$

where the expectation is taken with respect to the agent's belief $\Pr[y|s_i]$. For instance, with knowledge of the prior of y ($\Pr[y = s], s \in \{0,1\}$), the following prior dependent output agreement scoring function is strictly proper (1/Prior):

$$S(\tilde{s}_i, y) = \frac{1}{\Pr[y = \tilde{s}_i]} \cdot 1(\tilde{s}_i = y).$$

Lemma 3.1. *1/Prior scoring function is strictly proper.*

To unify prediction and signal elicitation, we will denote a strictly proper scoring rule for eliciting agent i's information I_i as $S(a_i, y) : I \times \{0,1\} \rightarrow \mathbb{R}_+$.

3.2 Surrogate scoring rules

We define *surrogate scoring rules* (SSR), when we only have access to z, a noisy and proxy signal of the ground truth y, and $e_{1,z} + e_{0,z} \neq 1$.

Definition 3.2 (Surrogate Scoring Rules). $\varphi : I \times \{0,1\} \rightarrow \mathbb{R}_+$ is a surrogate scoring rule if there exists a strictly proper scoring rule $S : I \times \{0,1\} \rightarrow \mathbb{R}_+$ and a strictly increasing function $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\forall a_i$,

$$\mathbb{E}_{y,z}[\varphi(a_i, z)] = f(\mathbb{E}_y[S(a_i, y)]). \tag{3}$$

If Eqn .(3) is true, the strict properness of $\varphi(\cdot)$ holds immediately because of the strict properness of S and monotonicity of $f(\cdot)$. For instance, a simple such f to aim for can be an affine function that there exist two constants $c_1 > 0, c_2$ such that

$$\mathbb{E}_{y,z}[\varphi(a_i, z)] = c_1 \cdot \mathbb{E}_y[S(a_i, y)] + c_2. \tag{4}$$

Theorem 3.3. *SSR is strictly proper for eliciting I_i.*

The above approach seeks a surrogate scoring function $\varphi(\cdot)$ that helps us remove the bias in z to return us a strictly proper score in expectation.

SSR can be viewed as a particular class of proxy scoring rules [Witkowski et al., 2017]. But the approach of [Witkowski et al., 2017] to achieve properness is to plug in an *unbiased* proxy ground truth to a strictly proper scoring rule. SSR on the other hand directly works with biased proxy and the scoring function is designed to de-bias the noise.

3.3 An implementation of SSR

Consider any strictly proper scoring function S for eliciting I_i using y, the ground truth. We define the following SSR, which we name as SSR_alpha:

$$\text{(SSR\_alpha)}: \; \varphi(a_i, z = s) = \frac{(1 - e_{1-s,z}) \cdot S(a_i, s) - e_{s,z} \cdot S(a_i, 1-s)}{1 - e_{1,z} - e_{0,z}}. \tag{5}$$

From above we note that the knowledge of the error rates $e_{1,z}, e_{0,z}$ is crucial for defining a SSR. The above scoring function is inspired by the work "learning with noisy label" [Natarajan et al., 2013]. Intuitively speaking, the linear transform will ensure that in expectation, a_i is scored against y:

Lemma 3.4. *When $e_{1,z} + e_{0,z} < 1$, we have for (SSR_alpha): $\mathbb{E}_{z|y}[\varphi(a_i, z)] = S(a_i, y)$, $\forall y \in \{0,1\}$.*

This can be proved fairly straightforwardly via spelling out the expectation. Interested readers are also referred to Lemma 1, [Natarajan et al., 2013]. In above, the scoring function $\varphi(\cdot)$ takes the form of linear combination of the scoring function evaluated at the noisy outcome z and the one evaluated at the opposite signal $1 - z$. The second term is to account for the fact that the reference signal is merely a noisy proxy for the ground truth signal. The linear weights, carefully defined over the bias level of the reference, balance the bias in the noisy signal with respect to the ground truth. Also note the above unbiasedness is achieved w.r.t. each possible realization of y, being independent of its prior distribution. This helps us remove the requirement of common knowledge over the priors among agents, and thus our solution works for the subjective prior setting. Now we establish the strict properness of SSR:

Theorem 3.5. *(SSR_alpha) is strictly proper.*

Proof. Consider agent i. Consider the case that $e_{1,z} + e_{0,z} < 1$. The proof is straightforward following the "unbiasedness" property established for $\varphi(\cdot)$ in Lemma 3.4:

$$\mathbb{E}[\varphi(a_i, z)] = \mathbb{E}\Big[\mathbb{E}[\varphi(a_i, z)|y]\Big] = \mathbb{E}[S(a_i, y)].$$

The theorem follows immediately from the strictly properness of S.

Now consider the case with $e_{1,z} + e_{0,z} > 1$. We cannot directly apply Lemma 3.4. Now let's define the "flip signal" $\hat{z}$ of z: $\hat{z} = 1 - z$. Easy to see that

$$e_{1,\hat{z}} := \Pr[\hat{z} = 0|z = 1] = \Pr[z = 1|z = 1] = 1 - e_{1,z},\ e_{0,\hat{z}} := \Pr[\hat{z} = 1|z = 0] = \Pr[z = 0|z = 0] = 1 - e_{0,z},$$

and $e_{1,\hat{z}} + e_{0,\hat{z}} < 1$. The scoring function for agent i then becomes:

$$\begin{aligned}
\varphi(a_i, z = s) &= \frac{(1 - e_{1-s,z}) \cdot S(a_i, s) - e_{s,z} \cdot S(a_i, 1-s)}{1 - e_{1,z} - e_{0,z}}. \\
&= \frac{[1 - (1 - e_{s,z})]S(a_i, 1-s) - (1 - e_{1-s,z})S(a_i, s)}{1 - (1 - e_{1,z}) - (1 - e_{0,z})} \\
&= \frac{(1 - e_{1-(1-s),\hat{z}})S(a_i, 1-s) - e_{1-s,\hat{z}}S(a_i, s)}{1 - e_{1,\hat{z}} - e_{0,\hat{z}}} \\
&= \varphi(a_i, \hat{z} = 1 - s).
\end{aligned} \tag{6}$$

Then we have

$$\mathbb{E}[\varphi(a_i, z)] = \mathbb{E}[\varphi(a_i, \hat{z})] = \mathbb{E}\Big[\mathbb{E}[\varphi(a_i, \hat{z})|y]\Big] = \mathbb{E}[S(a_i, y)],$$

where the last equality is due to the unbiasedness of $\varphi(\cdot)$ with respect to $\hat{z}$ (instead of z), and the fact that $e_{1,\hat{z}} + e_{0,\hat{z}} < 1$, so Lemma 3.4 can be applied. Again it is easy to see it would be agent i's best interest to tell the truth, due to the strict properness or truthfulness of $S(\cdot)$, finishing the proof. □

Remark 1. *We provide a simple yet effective scoring function to unify the two elicitation scenarios. Intuitively, the reason we can achieve above is that the core step of surrogate function $\varphi(\cdot)$ only operates with the reference answer z, and it is independent of the format of the information being elicited.*

4 DOMINANT TRUTH SERUM

The results in previous section are built upon the fact that there exists a reference signal for the ground truth and we know its error rates. In this section, we apply the idea of SSR to the peer prediction setting. A reasonable way to do so is to take agents' reports as the source for this noisy copy of the ground truth. Yet the mechanism designer cannot assume the knowledge of the noise in agents' reports. In this section we discuss how to select the reference answer z and how to learn $e_{1,z}, e_{0,z}$ accurately so that $\varphi(\cdot)$ will be well defined. Then applying Theorem 3.5, we will be able to establish a dominant strategy argument in the peer prediction setting.

4.1 Reference report

An intuitive way of selecting a reference report z is to uniformly randomly select a report from agents $j \neq i$, i.e., from $\mathcal{D}_{-i}$. Denote by $\tilde{e}_{1,i}, \tilde{e}_{0,i}$ the error rate of agent i's reported information a_i. For the signal elicitation setting, when agents truthfully report, i.e. $a_i = s_i$, we have $\tilde{e}_{1,i} = e_{1,i},\ \tilde{e}_{0,i} = e_{0,i}$, and when they revert their signals we have $\tilde{e}_{1,i} = 1 - e_{1,i},\ \tilde{e}_{0,i} = 1 - e_{0,i}$. For the case with prediction elicitation, we can similarly define a signal from agent i, via drawing a random sample: $a^s_i \sim \text{Bernoulli}(\tilde{p}_i)$. In this case, we will still call such a a^s_i as agent i's "signal" but we should keep in mind this signal is randomly drawn from his reported prediction, instead of being deterministic. We similarly define $\tilde{e}_{1,i}, \tilde{e}_{0,i}$, and assume conditional independence among a_i^ss.

$e_{1,z}$ and $e_{0,z}$ can then be characterized fairly easily as follows:

$$e_{1,z} := \frac{\sum_{j \neq i} \tilde{e}_{1,j}}{N-1}, \quad e_{0,z} := \frac{\sum_{j \neq i} \tilde{e}_{0,j}}{N-1}. \tag{7}$$

This can be obtained by the following argument: Denote the following even

$$\mathcal{W}_j := \{\text{agent } j \text{ selected as the reference agent}\}.$$

Then we have

$$\Pr[z=0|y=1] = \sum_{j \neq i} \Pr\Big[z=0,\ \mathcal{W}_j | y=1\Big]$$

$$= \sum_{j \neq i} \Pr\Big[z=0|y=1,\ \mathcal{W}_j\Big] \cdot \Pr\Big[\mathcal{W}_j\Big] = \frac{1}{N-1} \cdot \sum_{j \neq i} \tilde{e}_{1,j}\ .$$

We have similar interpretation for $e_{0,z}$. The next subsection will detail the steps towards estimating $e_{0,z}, e_{1,z}$. Note if we know the average error rates when agents truthfully report, the SSR already gives us an equilibrium implementation of truthful elicitation - as in the equilibrium argument, every other agent will truthfully report.

4.2 Bias learning with matching

Our learning algorithm for inferring $e_{0,z}, e_{1,z}$ relies on establishing three equations towards characterizing $e_{1,z}, e_{0,z}$. We will first show that the three equations, with knowing their true parameters, together will uniquely define $e_{1,z}, e_{0,z}$. Then we argue that with estimated and imperfect parameters from agents' report, the solution from the perturbed set of equations will approximate the true values of $e_{1,z}, e_{0,z}$, with guaranteed accuracy.

(1) **Posterior distribution:** The first equation is based on on the posterior distribution of 0/1 label information collected from $\mathcal{D}_{-i}$, denoting as $P_{0,-i} := \Pr[z=0], P_{1,-i} := \Pr[z=1]$. For $P_{0,-i}$ we have

$$P_{0,-i} = \mathcal{P}_1 e_{1,z} + \mathcal{P}_0 (1 - e_{0,z}). \tag{8}$$

Note the above equation is also equivalent with $P_{1,-i} = \mathcal{P}_1 (1 - e_{1,z}) + \mathcal{P}_0 e_{0,z}$.

(2) **Matching between two signals:** The second equation is derived from a second order statistics, namely the matching probability. Consider the following experiments: draw two independent signals (for the same task) from the reference agents [4], denote them as z_1, z_2. Denote the matching-on-1 probability of the two signals as

$$\Pr[z_1 = z_2 = 1] = q_{2,-i}.$$

Further the matching probability can be written as a function of $e_{0,z}, e_{1,z}$:[5]

$$q_{2,-i} = \mathcal{P}_1 (1 - e_{1,z})^2 + \mathcal{P}_0 e_{0,z}^2. \tag{9}$$

[4]Note in practice, we won't obtain independent copies of answers for the same question from the same agents for inferring the matching probability. However when the number of agent is large enough, we will show that drawing two or three agents randomly from the population (without replacement) can approximate these independent matching probabilities with small and diminishing errors (as a function of number of agents N).

[5]Note here we implicitly assumed that agents' reported signals are also conditional independent of the ground truth signal. In practice, without knowing the ground truth, it is hard for agents to coordinate based on the ground truth. In fact if the reporting strategies are functions of each agent's observation, we will have the conditional independence established immediately, according to the conditional independence of agents' observations.

(3) **Matching among three signals:** The third equation is obtained by going one order higher that, we check the matching-on-1 probability over three independent signals drawn from reference agents. Similarly as defined for (2) matching between two signals, draw three independent signals (for the same task) from the reference agents, denote them as z_1, z_2, z_3. Denote the matching-on-1 probability of the three signals as $\Pr[z_1 = z_2 = z_3 = 1] = q_{3,-i}$. Then we also have that

$$q_{3,-i} = \mathcal{P}_1(1 - e_{1,z})^3 + \mathcal{P}_0 e_{0,z}^2. \tag{10}$$

The reasons we selected the above three equations are two-fold. The first reason is that, as we will show below, the above three equations uniquely define $e_{0,z}, e_{1,z}$. The second reason is that in practice we only need to know $P_{1,-i}, q_{2,-i}, q_{3,-i}$ in order to establish the three equations, since we have assumed the knowledge of priors $\mathcal{P}_0, \mathcal{P}_1$. All three parameters $P_{1,-i}, q_{2,-i}, q_{3,-i}$ can be estimated from agents' reports, without the need of knowing any ground truth labels.

As promised, we first establish the following theorem:

Theorem 4.1. $(e_{0,z}, e_{1,z})$ *is the unique pair of solution to Eqn.(8, 9, 10), when* $\mathcal{P}_0 \neq \mathcal{P}_1$.

Proof. First of all, there are at most two solutions from Eqn. (8) and (9), since the equations are at most order two. Further we prove the following properties of its solutions:

Lemma 4.2. *For the solutions to Eqn.(8) and (9), we have*

(1) When the reports are uninformative such that $e_{0,z} + e_{1,z} = 1$, *the above equation return exactly one solution satisfying that* $e_{0,z} + e_{1,z} = 1$.

(2) When the reports are informative such that $e_{0,z} + e_{1,z} \neq 1$, *Eqn.(8) and (10) jointly return two pair of solutions*

$$[e_{0,z}(1), e_{1,z}(1)],\ [e_{0,z}(2), e_{1,z}(2)]$$

and exactly one of them satisfies $e_{0,z}(1)+e_{1,z}(1) < 1$ *and the other one that* $e_{0,z}(2)+e_{1,z}(2) > 1$. *Further their summations of the two pairs have the same distance to the 1 (uninformative):*

$$1 - (e_{0,z}(1) + e_{1,z}(1)) = (e_{0,z}(2) + e_{1,z}(2)) - 1.$$

We shall shorthand the following: $x_1 := e_{0,z},\ x_2 := 1 - e_{1,z}$.

Proof. Eqn.(8) and (9) become equivalent with the following

$$\text{(I)} : \mathcal{P}_0 x_1 + \mathcal{P}_1 x_2 = P_{1,-i},\ \text{(II)} : \mathcal{P}_0 x_1^2 + \mathcal{P}_1 x_2^2 = q_{2,-i}.$$

From the first equation (I) we know that $x_1 = \frac{P_{1,-i} - \mathcal{P}_1 \cdot x_2}{\mathcal{P}_0}$. Then

$$x_1 - x_2 = \frac{P_{1,-i} - \mathcal{P}_1 \cdot x_2}{\mathcal{P}_0} - x_2 = \frac{P_{1,-i} - x_2}{\mathcal{P}_0}$$

Further plug x_1 into the second equation (II) we have

$$\mathcal{P}_1 x_2^2 - 2\mathcal{P}_1 P_{1,-i} x_2 + P_{1,-i}^2 - \mathcal{P}_0 q_{2,-i} = 1 \Rightarrow x_2 = P_{1,-i} \pm \Delta,$$

$$\text{where } \Delta := \frac{\sqrt{(2\mathcal{P}_1 P_{1,-i})^2 - 4\mathcal{P}_1(P_{1,-i}^2 - \mathcal{P}_0 q_{2,-i})}}{2\mathcal{P}_1}.$$

From above we have derived the following $e_{0,z} + e_{1,z} - 1 = x_1 - x_2 = \pm\frac{\Delta}{\mathcal{P}_0}$. When $e_{0,z} + e_{1,z} = 1$, we know we must have $\Delta = 0$ so we land at an unique solution. When $e_{0,z} + e_{1,z} \neq 1$, we know that $\Delta > 0$ and we have two solutions – but both solutions satisfy that

$$|e_{0,z} + e_{1,z} - 1| = \frac{\Delta}{\mathcal{P}_0},$$

finishing the proof. □

We now show the following fact:

$$
\begin{aligned}
&\text{Eqn.(10)} - \text{Eqn.(8)} \cdot \text{Eqn.(9)} \\
\Leftrightarrow &\mathcal{P}_0 x_1^3 + \mathcal{P}_1 x_2^3 - (\mathcal{P}_0 x_1 + \mathcal{P}_1 x_2)(\mathcal{P}_0 x_1^2 + \mathcal{P}_1 x_2^2) = q_{3,i} - q_{2,i} P_1 \\
\Leftrightarrow &\mathcal{P}_0 \mathcal{P}_1 (x_1^3 + x_2^3 - x_1^2 x_2 - x_1 x_2^2) = q_{3,-i} - q_{2,-i} P_{1,-i} \\
\Leftrightarrow &\mathcal{P}_0 \mathcal{P}_1 (x_1 - x_2)^2 (x_1 + x_2) = q_{3,-i} - q_{2,-i} P_{1,-i} \\
\Leftrightarrow &x_1 + x_2 = \frac{(q_{3,-i} - q_{2,-i} P_{1,-i}) \mathcal{P}_0}{\Delta^2 \mathcal{P}_1}
\end{aligned}
$$

The above equation, together with equation (I), will uniquely identify the solution for x_1, x_2 when $\mathcal{P}_0 \neq \mathcal{P}_1$:

$$
\begin{aligned}
x_1 &= \frac{1}{\mathcal{P}_1 - \mathcal{P}_0} \left(\frac{q_{3,-i} - q_{2,-i} P_{1,-i}}{q_{2,-i} - P_{1,-i}^2} \mathcal{P}_1 - P_{1,-i} \right), \\
x_2 &= \frac{1}{\mathcal{P}_1 - \mathcal{P}_0} \left(P_{1,-i} - \frac{q_{3,-i} - q_{2,-i} P_{1,-i}}{q_{2,-i} - P_{1,-i}^2} \mathcal{P}_0 \right).
\end{aligned}
$$

This completes the proof. □

This result implies that when there is no ground truth data, knowing how frequently human agents reach consensus with each other will help us characterize their (average) subjective biases.

We conjecture that the above estimation method can be extended to the case with unknown priors of the ground truth. The idea is very straightforward following our approach: We can relax the knowledge of the priors to have two additional unknown variables. Then we will go two orders up to have additional two equations characterizing the matchings among four and five reference signals, to cop with the two more unknown variables. This merits future studies.

4.3 Statistical consistency results

Now we detail the learning procedure. Our method relies on multiple tasks (thus our mechanism is naturally a multi-task one). We will collect information on multiple independently drawn events from priors $y(1), y(2), \cdots \sim \mathcal{P}_0, \mathcal{P}_1$ by randomly assigning them to agents. Each task will be randomly assigned to three agents. Suppose there are K tasks that have been assigned to agents $j \neq i$ - this is not hard to guarantee $\forall i$ when we assign more than K tasks to the entire set of agents randomly, and when N is large.

The algorithm first estimates the following three quantities for each agent i: $\tilde{q}_{2,-i}, \tilde{q}_{3,-i}$ and $\tilde{P}^i_{1,-i}$ based on above information. Denote the three agents that are assigned task $k = 1, \ldots, K$ as $r_1(k), r_2(k), r_3(k) \neq i$, and their reports for task k as $\tilde{y}_{r_{\text{idx}}(k)}(k)$, idx $= 1, 2, 3$. Then we estimate the three parameters as follows:

$$
\begin{aligned}
\tilde{P}_{1,-i} &= \frac{\sum_{k=1}^K 1(\tilde{y}_{r_1(k)}(k) = 1)}{K}, \\
\tilde{q}_{2,-i} &= \frac{\sum_{k=1}^K 1(\tilde{y}_{r_1(k)}(k) = \tilde{y}_{r_2(k)}(k) = 1)}{K}, \\
\tilde{q}_{3,-i} &= \frac{\sum_{k=1}^K 1(\tilde{y}_{r_1(k)}(k) = \tilde{y}_{r_2(k)}(k) = \tilde{y}_{r_3(k)}(k) = 1)}{K}.
\end{aligned}
$$

Mechanism 1 Learning of $e_{1,z}, e_{0,z}$

1. Estimate $\tilde{q}_{2,-i}, \tilde{q}_{3,-i}$ and $\tilde{P}_{1,-i}$ using $\{\tilde{y}_{r_{\text{idx}}(k)}(k), k = 1, 2, ..., K, \text{idx} = 1, 2, 3\}$.
2. Compute the following:

$$\tilde{e}_{0,z} := \frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left(\frac{\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2} \cdot \mathcal{P}_1 - \tilde{P}_{1,-i}\right), \tag{11}$$

$$\tilde{e}_{1,z} := 1 - \frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left(\tilde{P}_{1,-i} - \frac{\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2} \cdot \mathcal{P}_0\right), \tag{12}$$

Remark 2. *We notice that the sensitivity in computing $e_{0,z}, e_{1,z}$ is inversely proportional to $\mathcal{P}_1 - \mathcal{P}_0$. This implies that the difficulty in selecting the right pair of root increases when the priors are less differentiable.*

We would like to show that the estimation error in estimating reports' error rate can be bounded as a function of K and N. The first source of errors is the imperfect estimations of $q_{2,-i}, q_{3,-i}, P_{1,-i}$. The second is due to estimation errors for matching probability with heterogeneous agents. Formally we have the following theorem:

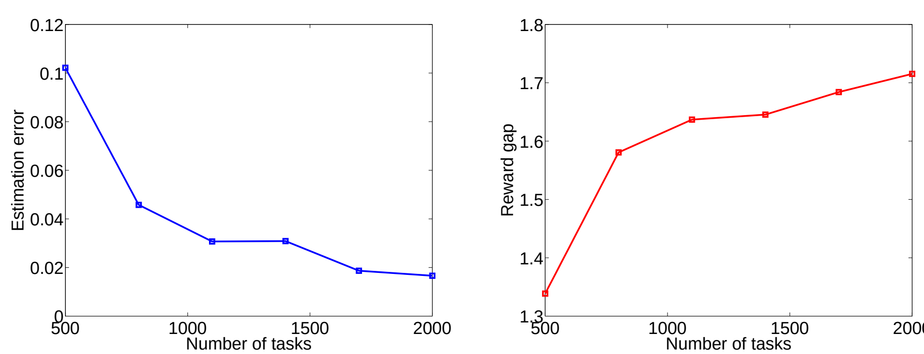

Fig. 1. Convergence of the estimation. Left: convergence with different number of tasks. $e_{0,z} = 0.85, e_{1,z} = 0.83$. Right: gap in reward functions (1/Prior scores) between truth-telling and mis-reporting.

Theorem 4.3. *The mechanism designer can learn noisy copies of $e_{1,z}, e_{0,z}, \varphi(\cdot)$ for each agent i using data collected from agents $j \neq i$, satisfying that*

(1) $|\tilde{e}_{1,z} - e_{1,z}| \leq \epsilon$, $|\tilde{e}_{0,z} - e_{0,z}| \leq \epsilon$ with probability at least $1 - \delta$.
(2) For the scoring function $\varphi(\cdot)$ defined using $\tilde{e}_{1,z}, \tilde{e}_{0,z}$, we have with probability at least $1 - \delta_1$ that $|\tilde{\varphi}(t, y) - \varphi(t, y)| \leq \epsilon_1$, $\forall t, y$.

All terms $\epsilon = O\left(\frac{1}{N} + \sqrt{\frac{\log K}{K}}\right), \delta = O(\frac{1}{K}), \epsilon_1 = O\left(\frac{1}{N} + \sqrt{\frac{\log K}{K}}\right), \delta_1 = O(\frac{1}{K})$ can be made arbitrarily small with increasing number of samples K and N.

Proof can be found in Appendix. The main steps of the proof are based on Theorem 4.1 and sensitivity analysis of the solutions Eqn. (11,12). A set of simple simulation results are shown in Figure 1. We simulated a scenario with $e_{0,z} = 0.85, e_{1,z} = 0.83$, i.e., majority people are wrong from the reference set of agents. Nevertheless the Left side figure shows a nice decay in the errors of estimating $e_{0,z}, e_{1,z}$. The Right side figure shows the gap in expected score between truthful reporting and misreporting (denoted as "Reward gap") with the estimated $e_{0,z}, e_{1,z}$ (for the same simulation setting). Indeed we show truthful reporting is better (positive gap, increasing with more accurate estimations of $e_{0,z}, e_{1,z}$) even when majority of people are wrong (as $e_{0,z} = 0.85, e_{1,z} = 0.83$).

4.4 Truth-telling is a dominant strategy

The above results readily imply that our strategy is asymptotically (in K, N) dominant. To see this, suppose both $\tilde{e}_{1,z}, \tilde{e}_{0,z}$ have converged to their true values. We define *dominant truth serum* as follows:

Mechanism 2 Dominant Truth Serum (DTS)

1. For each task agent i received and reported information, randomly select one of the rest two reference reports (recall each task is randomly assigned to three agents) as the reference z.
2. When z is uninformative that $\tilde{e}_{1,z} + \tilde{e}_{0,z} = 1$:
   - Score agent i zero regardless of his report.
3. When z is informative that $\tilde{e}_{1,z} + \tilde{e}_{0,z} \neq 1$:
   - Score agents via (SSR_alpha) defined using $z, \tilde{e}_{1,z}, \tilde{e}_{0,z}$. (with possible scalings)

Note again in above, if the collected reports are in the format of prediction, we will draw a signal according to the reported prediction - this is the z that we will use.

Theorem 4.4. *Asymptotically it is a dominant strategy for agent i to report truthfully under DTS.*

Proof. First for the case that $\tilde{e}_{1,z} + \tilde{e}_{0,z} = 1$, it is indifferent for agent i to truthfully report, or to misreport, or to randomize between the two strategies. Thus truth-telling is a weakly dominant strategy. When $\tilde{e}_{1,z} + \tilde{e}_{0,z} \neq 1$, the dominant strategy argument follows from the strictly properness of (SSR_alpha). □

Remark 3. *Several remarks follow.*

- *We would like to emphasize again that both z and φ come from $\mathcal{D}_{-i}$: z will be decided by agents $j \neq i$'s reports $\mathcal{D}_{-i}$. φ not only has z as input, but its definition also depends on $e_{1,z}$ and $e_{0,z}$, which will be learned from $\mathcal{D}_{-i}$.*
- *When making decisions on reporting, we show under our mechanisms agents can choose to be oblivious of how much error presents in others' reports. This removes the practical concern of implementing a particular Nash Equilibrium.*
- *Another salient feature of our mechanism is that we have migrated the cognitive load for having prior knowledge from agents to the mechanism designer. Yet we do not assume the designer has direct knowledge neither; instead we will leverage the power of estimation from reported data to achieve our goal.*

Now we argue the dominant strategy in finite sample regime, under noisy estimations. We first define the informative region. When $e_{0,z} + e_{1,z}$ is arbitrarily close to 1, we have the difficulty in determining the number of samples needed for the learning process. With this in mind, we will modify our mechanism to the following:

- Pre-select a small constant $0 < \kappa < 1$ (but close to 0).
- When $|\tilde{e}_{1,z} + \tilde{e}_{0,z} - 1| \leq \kappa$ (instead of setting it to be exactly 0; this characterizes our informative region), pay agent i nothing;
- Else follow our original mechanism and pay according to (SSR_alpha).

κ can then help us quantify the number of K, N needed. With this we prove the following theorem:

Theorem 4.5. *When set κ small enough, and K, N large enough, we have that with finite number of tasks:*

(1) For signal elicitation, it is a dominant strategy for each agent i to report s_i truthfully under DTS;
(2) For prediction elicitation, it is a dominant strategy to report p_i truthfully when $S(\cdot)$ is strictly concave in report a_i and Lipschitz.
(3) For prediction elicitation, it is an $\epsilon(K, N)$-dominant strategy to report p_i truthfully for any Lipschitz $S(\cdot)$, where $\epsilon(K, N) = O\left(\frac{1}{N} + \sqrt{\frac{\log K}{K} + \frac{1}{K}}\right)$ is a diminishing term in both K and N.

Above we adopt the classical approximate-dominant strategy definition that truth-telling is ϵ-dominant strategy if $\mathbb{E}[\varphi(I_i, z)] > \mathbb{E}[\varphi(a_i, z)] - \epsilon, \ \forall a_i \neq I_i$.

The proof for the signal elicitation case is rather straightforward. We need to show that when the estimation errors of the error rates $e_{0,z}, e_{1,z}$ are small enough, the scores under each basic reporting strategy (truthful, always reverting the observation, always reporting 1 and always reporting 0) will be close to the true score. Then the truthfulness follows from the strict properness of S. It is slightly more challenging for the prediction elicitation setting (reporting p_is), due to the continuous reporting space. The main idea is that when the estimation error drops below the regularization term that corresponds to S's strict concavity, any possible gain from deviating from truthful reporting due to the error terms is less than the loss due to the drop in the regularization term. Note it is not hard to find strictly concave scoring function S - both Brier and logarithmic scores are strictly concave.

4.5 Weak dominance v.s. strong dominance

We have shown when we cannot learn an informative enough reference report from the rest of the agents, there is very little we can do regarding that. With this note, our deterministic payment strategy is only weakly dominant. We argue that if agents' reports can be modeled as being from certain distributions, the weak dominance case happens rarely.

Denote by $X_i = \tilde{e}_{1,i} + \tilde{e}_{0,i}$, the weakly dominant case happens only when $\sum_{j \neq i} X_j = 1$. Potentially there are infinitely many strategies that lead to this state. However, suppose agents' reporting strategies are drawn from a distribution defined over a continuous space (mixed strategy space), then we can easily argue that $\{\omega : \sum_{j \neq i} X_j(\omega) = 1\}$ (via taking $\sum_{j \neq i} X_j$ as a random variable) is a zero measure event.

Also we would like to point out that, as long as each agent believes that this uninformative state happens with probability < 1, DTS induces strong dominant strategy in truthful reporting, as a non-trivial mix between weak and strong dominance returns strong dominance.

Also, as a matter of fact, $\sum_{j \neq i} X_j = 1$ corresponds to the case that the reference signal is *stochastically irrelevant* to the ground truth. Theoretically speaking, there is little one can do for eliciting private signals using such a reference signal, via a peer prediction method. In practice, one can use screening tasks to remove some workers to move away from this uninformative state.

5 CORRELATED SIGNALS

Earlier on, we have assumed that agents' private signals are conditional independent (on ground truth). In this section, we discuss the generalization of our method to the case when agents' observations are correlating with each other. The intuition for why we can hope for this extension is that we do not explore the correlation structure among signals explicitly. Technically this is due to our scoring system's dependence on the reference answer z through $e_{1,z}, e_{0,z}$, which is independent of whether z is drawn from a set of independent signals or a set of correlated ones. Being correlated or not, as long as we can learn the average error rate from the rest of agents, we will have an unbiased score for each agent w.r.t. the ground truth signal. To extend our results, the only challenge we need to address is to accurately estimate $\Pr[z=1|y=0], \Pr[z=0|y=1]$.

5.1 Learning the error rate

However, learning $\Pr[z=1|y=0], \Pr[z=0|y=1]$ with correlated signals is not as straightforward as before. The reason is in Eqn. (9, 10) of our proposed method in Section 4.2, the second and third-order terms for characterizing the matching probability will not hold. Correlation among signals will violate the product structure of the matching equations we had earlier. Nonetheless we would like to stick with our idea of using the matching probabilities among agents to obtain additional information. We show this is not entirely hopeless:

Lemma 5.1. *The second order matching probabilities are the same for the cases with correlated and independent signals:*

$$
\begin{aligned}
&\Pr[z_1=1, z_2=1|y=0] + \Pr[z_1=0, z_2=0|y=0] \\
&= \Pr[z_1=1|y=0]\Pr[z_2=1|y=0] + \Pr[z_1=0|y=0]\Pr[z_2=0|y=0] \\
&\Pr[z_1=1, z_2=1|y=1] + \Pr[z_1=0, z_2=0|y=1] \\
&= \Pr[z_1=1|y=1]\Pr[z_2=1|y=1] + \Pr[z_1=0|y=1]\Pr[z_2=0|y=1]
\end{aligned}
$$

With above lemma, we know the correlation structure does not affect estimating the second order matching probability. Similar to what we have detailed earlier that, the first order and second order equation together will help us reduce the possible solutions to at most two pairs. Particularly we can do the following:

1. Estimate $\tilde{q}_i$ using $\{\tilde{y}_{r_{\text{idx}}(k)}(k), k=1,2,...,K, \text{idx}=1,2,3\}$:
$$\tilde{q}_{-i} = \frac{\sum_{k=1}^{K} 1(\tilde{y}_{r_1(k)}(k) = \tilde{y}_{r_2(k)}(k))}{K}.$$
i.e., $\tilde{q}_{-i}$ is the empirical matching probability (on both 0 & 1).
2. Solve the following set of equations.
$$
\begin{aligned}
&\text{(I): } \mathcal{P}_0[e_{0,z}^2 + (1-e_{0,z})^2] + \mathcal{P}_1[e_{1,z}^2 + (1-e_{1,z})^2] = \tilde{q}_{-i}, \\
&\text{(II): } \mathcal{P}_0 e_{0,z} + \mathcal{P}_1(1-e_{1,z}) = \tilde{P}_{1,-i}.
\end{aligned}
$$

First notice that Eqn. (I) doesn't contain additional information than

$$\mathcal{P}_0 e_{0,z}^2 + \mathcal{P}_1(1-e_{1,z})^2 = q_{2,-i} \text{ (please refer to Eqn. (9) in previous estimation)}$$

, in light of $\mathcal{P}_0 e_{0,z} + \mathcal{P}_1(1-e_{1,z})$ as:

$$
\begin{aligned}
\mathcal{P}_0[p_{0,z}^2 + (1-p_{0,z})^2] + \mathcal{P}_1[p_{1,z}^2 + (1-p_{1,z})^2] = {} & 2[\mathcal{P}_0 p_{0,z}^2 + \mathcal{P}_1(1-p_{1,z})^2] \\
& - 2[\mathcal{P}_0 p_{0,z} + \mathcal{P}_1(1-p_{1,z})] + 1
\end{aligned}
$$

So they necessarily lead to the same set of solutions, and thus Lemma 4.2 will hold too.

We need a third equation for picking the exact solution. The idea is to similarly check one order higher matching, but we will use $\Pr[z_1 = z_2, z_3 = 1]$ instead of $\Pr[z_1 = z_2 = z_3 = 1]$. To let our presentation stay focused, we make the following assumption:

Assumption 1. *With three drawn reference signals z_1, z_2, z_3 from $j \neq i$, we have* $\Pr[z_1, z_2, z_3|y] = \Pr[z_1, z_2|y] \cdot \Pr[z_3|y]$.

That is the conditional dependency is up to order two. This is not entirely unreasonable to assume: when signals' correlations are due to a common set of observed attributes (e.g., essay length to essay quality), we can assume that one another signal z_2 already encodes sufficient correlational information with z_1, and thus the conditional independence between (z_1, z_2) and z_3. We then note the following fact: adopting the short-hand notation $x_1 := e_{0,z},\ x_2 := 1 - e_{1,z}$ we show

$$\begin{aligned}\Pr[z_1 = z_2, z_3 = 1] &= \mathcal{P}_0 \Pr[z_1 = z_2, z_3 = 1 | y = 0] + \mathcal{P}_1 \Pr[z_1 = z_2, z_3 = 1 | y = 1] \\ &= \mathcal{P}_0 \Pr[z_1 = z_2 | y = 0] \Pr[z_3 = 1 | y = 0] + \mathcal{P}_1 \Pr[z_1 = z_2 | y = 1] \Pr[z_3 = 1 | y = 1] \\ &= \mathcal{P}_0 (x_1^2 + (1 - x_1)^2) x_1 + \mathcal{P}_1 (x_2^2 + (1 - x_2)^2) x_2.\end{aligned}$$

The first equation is based on the Bayesian structure of ground truth signal. The second is due to Assumption 1. The third equation is due to Lemma 5.1. The following Lemma can then be established:

Lemma 5.2.

$$x_1 + x_2 = \frac{\Pr[z_1 = z_2, z_3 = 1] - \Pr[z_1 = z_2] \Pr[z_1 = 1]}{2\mathcal{P}_0 \mathcal{P}_1 (x_1 - x_2)^2} + \frac{1}{\mathcal{P}_0 \mathcal{P}_1 (x_1 - x_2)^2}.$$

Similarly with knowing $|x_1 - x_2|$ (jointly determined by the first order and second order equations, as detailed in Lemma 4.2), we are able to formulate another equation regarding x_1, x_2.

All three parameters $\Pr[z_1 = z_2, z_3 = 1], \Pr[z_1 = z_2], \Pr[z_1 = 1]$ are estimable from agents' reports. Together with $\mathcal{P}_0 x_1 + \mathcal{P}_1 x_2 = P_{1,-i}$, we are able to solve for x_1, x_2. The rest of arguments follow the proof of Theorem 4.1.

5.2 Application: removing curse of weak signals

Extending our results to the correlated signal case not only removes one technical assumption, but also helps remove a curse for peer prediction introduced by the existence of weak but strongly correlated signals. Consider the case where besides the quality signal (i.e., the true signal we really want to elicit) s_i, each agent i also observes a cheap signal, denoting as s_i^l. Such cheap signals often exist in practice. For example in peer grading, the length of an assignment is one of such signals. Cheap signals are often less accurate than the quality signals in that

$$\Pr[s_i^l = 1 | y = 0] > \Pr[s_i = 1 | y = 0],\ \Pr[s_i^l = 0 | y = 1] > \Pr[s_i = 0 | y = 1],$$

and further such weak signals do not always satisfy conditional independence in that $\Pr[s_i^l, s_j^l | y] \neq \Pr[s_i^l | y] \Pr[s_j^l | y]$. Nonetheless it is also often true that weak signals correlate with each other more than the quality signals: $\Pr[s_i^l, s_j^l | y] > \Pr[s_i, s_j | y]$.

The existence of weak signals has been observed to be harmful in that reporting such weak yet more correlated signals may lead to higher utilities for agents [Gao et al., 2016], as peer prediction essentially checks the correlation among signals. Suppose s_i and s_i^l are conditionally independent given y, formally we prove the following results.

Theorem 5.3. *When S is taken as the 1/Prior scoring rule, reporting s_i (instead of the weak signal s_i^l) is a dominant strategy with DTS.*

6 DISCUSSIONS AND CONCLUDING REMARKS

Recovering the ground truth. Though being a fundamental step, truthful elicitation is also often not the ultimate goal, as the mechanism designer may want to infer the ground truth answer in the end. As a by-product, our framework for inferring the error rates from agents would allow us to do so. The reasoning is as follows: first ask each agent $j \neq i$ to report a label. If the ground truth label is 0, the posterior distribution of agents' reports being 1 will converge to $e_{0,z}$ with high probability; while on the other hand, if the ground truth is 1, this posterior distribution converges to $1 - e_{1,z}$. When $e_{0,z} \neq 1 - e_{1,z}$ (reference answer being informative), we will know the ground truth label with high probability via posterior distribution testing.

Unbiased estimator of the true score. Our paper proposes a novel and robust way for quantifying the quality of information when there is no ground truth verification, using peer evaluation, as well as an estimation procedure for learning the bias in peer evaluation. This may find applications in a sequential elicitation setting, where such scores can provide informative feedbacks to contributed information before the outcome reveals.

Effort exertion. So far our discussion focuses on eliciting truthful reports, our results extend to the scenario for eliciting high quality data. For instance, when workers can choose to exert costly effort (consider a binary effort level case, and denote the cost as c for exerting efforts) to improve the quality of their answers, and once the reference signal is informative (either positively or negatively), we can scale up the scoring functions to cover the cost c, as similarly done in [Miller et al., 2005], to establish the dominance of reporting a high-effort signal.

Heterogeneous error rates. In order to estimate error rates correctly, we needed to make the assumption that agents have homogeneous error rates over multiple tasks. We now discuss the applicability of our method in light of the heterogeneity issue. First we would like to emphasize that in practice, for each set of experiments we can choose to group tasks according to their types (e.g., image labeling, solving puzzles, objects recognition), and run our mechanism over each group separately. For this setting, we can assume the homogeneity more comfortably. When it is not quite possible to group tasks together, suppose that each human agent's error rates are also task contingent in that for each possible task, we show our solution can be viewed as a linear approximation for this task contingent case. We have more details on this matter in the Appendix.

A Machine Learning method. To echo recent works on using machine learning techniques [Liu and Chen, 2017] to learn a machine learning model to generate a reference answer, instead soliciting from other agents, we show this idea is also ready to be plugged into our SSR solution framework. Consider the current task that needs to be elicited and denote its feature vector as $\mathbf{x} \in \mathbb{R}^d$. Suppose we have learned (following the results in [Liu and Chen, 2017], such a classifier is learnable purely from agent's reported noisy data) a good classifier $\tilde{f}^*(\mathbf{x})$ for predicting its true outcome. Replace the reference answer z with $\tilde{f}^*(\mathbf{x})$ and plug in its error rates. The rest of job is to reason about $e_{1,z}, e_{0,z}$ (for $\tilde{f}^*(\mathbf{x})$). There are possibly many different ways to do so. We sketch its possibility when there is again no ground truth label being available in Appendix.

Conclusion. In this paper we propose *surrogate scoring rules* (SSR), which complement the literature of strictly proper scoring rules by considering the setting when there is only access to a noisy copy of the ground truth. We further extend SSR to the peer prediction setting where the noisy ground truth can only be inferred via collecting reports from other reference agents. This returns us a *dominant truth serum* (DTS), which achieves truthful elicitation in dominant strategy.

REFERENCES

Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

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

Tom Bylander. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the seventh annual conference on Computational learning theory*. ACM, 340–347.

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

Rafael Frongillo and Ian A Kash. 2015. Vector-valued property elicitation. In *Conference on Learning Theory*. 710–727.

Alice Gao, James R Wright, and Kevin Leyton-Brown. 2016. Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. *arXiv preprint arXiv:1606.07042* (2016).

Xi Alice Gao, Andrew Mao, Yiling Chen, and Ryan Prescott Adams. 2014. Trick or treat: putting peer prediction to the test. In *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 507–524.

Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Amer. Statist. Assoc.* 102, 477 (2007), 359–378.

Victor Richmond Jose, Robert F. Nau, and Robert L. Winkler. 2006. Scoring Rules, Generalized Entropy and utility maximization. (2006). Working Paper, Fuqua School of Business, Duke University.

R. Jurca and B. Faltings. 2006. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce (EC '06)*. ACM, 190–199.

Yuqing Kong and Grant Schoenebeck. 2016a. Equilibrium selection in information elicitation without verification via information monotonicity. *arXiv preprint arXiv:1603.07751* (2016).

Yuqing Kong and Grant Schoenebeck. 2016b. A Framework For Designing Information Elicitation Mechanisms That Reward Truth-telling. *arXiv preprint arXiv:1605.01021* (2016).

N.S. Lambert, D.M. Pennock, and Y. Shoham. 2008. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce (EC '08)*. ACM, 129–138.

Yang Liu and Yiling Chen. 2017. Machine Learning aided Peer Prediction. *ACM EC* (June 2017).

Yang Liu and Mingyan Liu. 2015. An Online Learning Approach to Improving the Quality of Crowd-Sourcing. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '15)*. ACM, New York, NY, USA, 217–230. DOI: http://dx.doi.org/10.1145/2745844.2745874

James E Matheson and Robert L Winkler. 1976a. Scoring rules for continuous probability distributions. *Management science* 22, 10 (1976), 1087–1096.

James E. Matheson and Robert L. Winkler. 1976b. Scoring Rules for Continuous Probability Distributions. *Management Science* 22, 10 (1976), 1087–1096.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. 2015. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*. 125–134.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science* 51, 9 (2005), 1359 –1373.

Allan H Murphy and Robert L Winkler. 1970. Scoring rules in probability assessment and evaluation. *Acta psychologica* 34 (1970), 273–286.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*. 1196–1204.

Dražen Prelec. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306, 5695 (2004), 462–466.

G. Radanovic and B. Faltings. 2013. A Robust Bayesian Truth Serum for Non-Binary Signals. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI '13)*.

Leonard J Savage. 1971. Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* 66, 336 (1971), 783–801.

Clayton Scott. 2015. A Rate of Convergence for Mixture Proportion Estimation, with Application to Learning from Noisy Labels.. In *AISTATS*.

Clayton Scott, Gilles Blanchard, Gregory Handy, Sara Pozzi, and Marek Flaska. 2013. Classification with Asymmetric Label Noise: Consistency and Maximal Denoising.. In *COLT*. 489–511.

V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. 2016. Informed Truthfulness in Multi-Task Peer Prediction. *ACM EC* (March 2016). arXiv:cs.GT/1603.03151

Brendan van Rooyen and Robert C Williamson. 2015. Learning in the Presence of Corruption. *arXiv preprint arXiv:1504.00091* (2015).

Robert L. Winkler. 1969. Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.* 64, 327 (1969), 1073–1078.

Jens Witkowski, Pavel Atanasov, Lyle H Ungar, and Andreas Krause. 2017. Proper proxy scoring rules.

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

J. Witkowski and D. Parkes. 2012. A Robust Bayesian Truth Serum for Small Populations. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI '12)*.

APPENDIX

PROOF OF LEMMA 2.2

Proof. Suppose not we will have

$$\Pr[y=0|s_i=0] = \Pr[y=0|s_i=1], \tag{13}$$
$$\Pr[y=1|s_i=0] = \Pr[y=1|s_i=1]. \tag{14}$$

From Eqn. (13) we know that

$$\frac{\Pr[y=0, s_i=1]}{\Pr[s_i=1]} = \frac{\Pr[y=0, s_i=0]}{\Pr[s_i=0]}$$
$$\Leftrightarrow \frac{\Pr[y=0]p_{0,i}}{\Pr[s_i=1]} = \frac{\Pr[y=0](1-e_{0,i})}{\Pr[s_i=0]},$$

when $\Pr[y=0] \neq 0$ we know that

$$\frac{\Pr[s_i=1]}{\Pr[s_i=0]} = \frac{e_{0,i}}{1-e_{0,i}}. \tag{15}$$

Similarly from Eqn. (14) we know

$$\frac{\Pr[s_i=1]}{\Pr[s_i=0]} = \frac{1-e_{1,i}}{e_{1,i}}. \tag{16}$$

Therefore we know $\frac{e_{0,i}}{1-e_{0,i}} = \frac{1-e_{1,i}}{e_{1,i}}$ from which we know $e_{0,i} + e_{1,i} = 1$. Contradiction. □

PROOF OF LEMMA 3.1

Proof. This can be fairly easily established by noting the following: by truthful reporting agent i's expected score is

$$\mathbb{E}[S(s_i, y)|s_i] = \frac{\Pr[y=s_i|s_i]}{\Pr[y=s_i]} > \frac{\Pr[y=s_i]}{\Pr[y=s_i]} = 1.$$

On the other hand, by reverting the observation, the agent has

$$\mathbb{E}[S(1-s_i, y)|s_i] = \frac{\Pr[y=1-s_i|s_i]}{\Pr[y=1-s_i]} < \frac{\Pr[y=1-s_i]}{\Pr[y=1-s_i]} = 1.$$

Both of above inequalities are due to Bayesian informativeness (and correspondingly negative Bayesian informativeness). □

PROOF OF THEOREM 4.3

Proof. *Estimation error due to heterogeneous agents:* The first challenge lies in the fact that the higher order equations doesn't capture the true matching probability with heterogeneous workers. If we assign one task (to be labeled) to two different workers, due to the asymmetric error rate, the LHS of Eqn. (I) is not precise– drawing a worker without replacement leads to a different labeling accuracy, and will complicate the solution for the system of equations. We show that our estimation, though being ignoring above bias, will not affect our results by too much: denoting by e_1 the accuracy of first drawn sample (for checking the matching) when $y=1$ (we can similarly argue for $y = 0$). Then we have $\mathbb{E}[e_1] = e_{1,z}$. And the average accuracy of the second drawn sample e_2 (conditional on e_1) is given by $\frac{e_{1,z}(N-1)-e_1}{N-2}$. Further the matching probability satisfies the following:

$$\mathbb{E}[e_1 \cdot e_2] = \mathbb{E}\Bigg[\mathbb{E}[e_1 \cdot e_2|e_1]\Bigg]$$
$$=\mathbb{E}\left[e_1 \cdot \frac{e_{1,z}(N-1)-e_1}{N-2}\right]$$
$$=\mathbb{E}[e_1] \cdot \frac{e_{1,z}(N-1)}{N-2} - \frac{\mathbb{E}[e_1^2]}{N-2}$$
$$=\frac{N-1}{N-2}e_{1,z}^2 - \frac{\mathbb{E}[e_1^2]}{N-2}$$

Note both $e_{1,z}$ and $\mathbb{E}[e_1^2]$ are no more than 1. Then

$$|\frac{N-1}{N-2}e_{1,z} - \frac{\mathbb{E}[e_1^2]}{N-2} - e_{1,z}^2| \le \frac{|e_{1,z}^2|}{N-2} + \frac{\mathbb{E}[e_1^2]}{N-2} \le \frac{2}{N-2}.$$

That the estimated quantity is bounded away from the true parameter by at most $\Theta(1/N)$. Similarly for the third order equation we have

$$|\mathbb{E}[e_1 \cdot e_2 \cdot e_3] - e_{1,z}^3| \le \frac{3}{N-3}.$$

Estimation errors due to finite estimation samples: The second sources of errors come from the estimation errors of $\tilde{q}_{2,-i}, \tilde{q}_{3,-i}$ and $\tilde{P}_{1,-i}$. We have the following lemma:

Lemma 6.1. *When there are K samples for estimating $\tilde{q}_{2,-i}, \tilde{q}_{3,-i}$ and $\tilde{P}_{1,-i}$ respectively (total budgeting 3K) and $K \ge \frac{\log 6/\delta}{2\epsilon^2}$ for any $\epsilon, \delta > 0$, we have with probability at least $1-\delta$*

$$|\tilde{q}_{2,-i} - q_{2,-i}| \le \epsilon + \frac{2}{N-2},\ |\tilde{q}_{3,-i} - q_{3,-i}| \le \epsilon + \frac{3}{N-3},\ |\tilde{P}_{1,-i} - P_{1,-i}| \le \epsilon.$$

The above lemma can be easily established using Chernoff bound, and our arguments in estimating with heterogeneous agents above. We now bound the error in estimating $p_{0,z}, p_{1,z}$, under the (ϵ,δ)-event proved in Lemma 6.1. First of all

$$|\tilde{e}_{0,z} - e_{0,z}| = |\frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left(\frac{\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2}\mathcal{P}_1 - \tilde{P}_{1,-i}\right) - \frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left(\frac{q_{3,-i} - q_{2,-i}P_1}{q_{2,-i} - P_1^2}\mathcal{P}_1 - P_{1,-i}\right)$$
$$\le \frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left(|\frac{\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2} - \frac{q_{3,-i} - q_{2,-i}P_1}{q_{2,-i} - P_1^2}|\mathcal{P}_1 + |\tilde{P}_{1,-i} - P_{1,-i}|\right).$$

According to Lemma 7 of [Liu and Liu, 2015], which we reproduce as follows:

Lemma 6.2. *For $k \ge 1$ and two sequences $\{l_i\}_{i=1}^m$ and $\{q_i\}_{i=1}^m$ and $0 \le l_i, q_i \le 1, \forall i = 1, ..., k.$, we have*

$$|\prod_{i=1}^m l_i - \prod_{j=1}^m q_j| \le \sum_{i=1}^m |l_i - q_i| . \quad (17)$$

We know the following facts

$$|\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2) - (q_{2,-i} - P_{1,-i}^2)| \leq |(\tilde{P}_{1,-i})^2 - P_{1,-i}^2| + |\tilde{q}_{2,-i} - q_{2,-i}| \leq 3\epsilon + \frac{2}{N-2}$$

$$|(\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}) - (q_{3,-i} - q_{2,-i}P_{1,-i})| \leq |\tilde{q}_{3,-i} - q_{3,-i}| + |\tilde{q}_{2,-i}\tilde{P}_{1,-i} - q_{2,-i}P_{1,-i}|$$
$$\leq 3\epsilon + \frac{2}{N-2} + \frac{3}{N-3}.$$

First we prove that

$$q_{2,-i} - P_{1,-i}^2 = \mathcal{P}_0 x_1^2 + \mathcal{P}_1 x_2^2 - (\mathcal{P}_0 x_1 + \mathcal{P}_1 x_2)^2 = \mathcal{P}_0\mathcal{P}_1(x_1 - x_2)^2 \geq \mathcal{P}_0\mathcal{P}_1\kappa^2.$$

Let $3\epsilon + \frac{2}{N-2} \leq \mathcal{P}_0\mathcal{P}_1\kappa^2/2$, we know that (using mean-value theorem/inequality)

$$|\frac{\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2} - \frac{q_{3,-i} - q_{2,-i}P_{1,-i}}{q_{2,-i} - P_1^2}| \leq |\frac{\tilde{q}_{3,-i} - \tilde{q}_{2,-i}\tilde{P}_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2} - \frac{q_{3,-i} - q_{2,-i}P_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2}|$$
$$+ |\frac{q_{3,-i} - q_{2,-i}P_{1,-i}}{\tilde{q}_{2,-i} - (\tilde{P}_{1,-i})^2} - \frac{q_{3,-i} - q_{2,-i}P_{1,-i}}{q_{2,-i} - P_{1,-i}^2}|$$
$$\leq 2\frac{\epsilon + \frac{2}{N-2} + \frac{3}{N-3}}{\kappa^2\mathcal{P}_0^2\mathcal{P}_1^2} + 2\frac{\epsilon + \frac{2}{N-2}}{(\kappa^2\mathcal{P}_0^2\mathcal{P}_1^2)^2}.$$

Together we proved that

$$|\tilde{e}_{0,z} - e_{0,z}| \leq \frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left[2\mathcal{P}_1\frac{\epsilon + \frac{2}{N-2} + \frac{3}{N-3}}{\kappa^2\mathcal{P}_0^2\mathcal{P}_1^2} + 2\mathcal{P}_1\frac{\epsilon + \frac{2}{N-2}}{(\kappa^2\mathcal{P}_0^2\mathcal{P}_1^2)^2} + \epsilon\right].$$

Similarly we are able to work out sensitivity analysis for $e_{1,z}$ that

$$|\tilde{e}_{1,z} - e_{1,z}| \leq \frac{1}{\mathcal{P}_1 - \mathcal{P}_0}\left[2\mathcal{P}_0\frac{\epsilon + \frac{2}{N-2} + \frac{3}{N-3}}{\kappa^2\mathcal{P}_0^2\mathcal{P}_1^2} + 2\mathcal{P}_0\frac{\epsilon + \frac{2}{N-2}}{(\kappa^2\mathcal{P}_0^2\mathcal{P}_1^2)^2} + \epsilon\right].$$

Summarizing and set $\delta = O(\frac{1}{K})$ we have $\epsilon = O\left(\frac{1}{N} + \sqrt{\frac{\log K}{K}}\right)$. With the error rates bounds for $e_{0,z}, e_{1,z}$, from Lemma 5.4 [Liu and Chen, 2017], which we reproduce in our contexts as follows:

Lemma 6.3. *When K, N are large enough s.t. $\epsilon \leq (1 - e_{0,z} - e_{1,z})/4 := \Delta_e$, with probability at least $1 - 2\delta$, $|\tilde{\varphi}(t, y) - \varphi(t, y)| \leq \frac{1+\Delta_e}{2\Delta_e^2} \cdot \epsilon$.*

We then obtain a bound on estimating $\varphi(\cdot)$. □

PROOF OF THEOREM 4.5

Proof. Suppose $|e_{0,z} + e_{1,z} - 1| \geq 2\kappa$. Then when the estimation errors are small enough such that $|\tilde{e}_{0,z} - e_{0,z}| + |\tilde{e}_{1,z} - e_{1,z}| \leq \kappa$, we know that $|\tilde{e}_{0,z} + \tilde{e}_{1,z} - 1| \geq \kappa$. Denote by

$$\Delta_\varphi = \mathbb{E}[S(s_i, y)] - \max_{a_i \neq s_i} \mathbb{E}[S(a_i, y)],$$

the minimum gap between the scores for truthfully reporting and mis-reporting. We can easily show that $\Delta_\varphi > 0$ for signal elicitation case, as we can take a_i to be the three basis of mis-reporting: always reverting the observation, always reporting 1 and always reporting 0. Then with a noisy estimation of $\varphi(\cdot)$, we have (using Theorem 4.3)

$$|\mathbb{E}[\tilde{\varphi}(a_i, z)] - \mathbb{E}[\varphi(a_i, z)]| \leq \epsilon_1 + \delta_1 \cdot \max \tilde{\varphi}, \ \forall a_i.$$

Above, we implicitly assumed the boundedness of $\tilde{\varphi}(\cdot)$: notice with bounded scoring function S, indeed we know that $\max \tilde{\varphi} \leq \frac{2 \max S}{\kappa}$. Since $\mathbb{E}[\varphi(s_i,z)] = \mathbb{E}[S(s_i,y)]$, choose ϵ_1, δ_1 such that $\epsilon_1 + \delta_1 \cdot \max \tilde{\varphi} < \Delta_\varphi/2$ we will have (for $a_i \neq s_i$)

$$\begin{aligned}\mathbb{E}[\tilde{\varphi}(s_i,z)] &> \mathbb{E}[S(s_i,y)] - \Delta_\varphi/2 \\ &> \mathbb{E}[S(a_i,y)] + \Delta_\varphi/2 > \mathbb{E}[\tilde{\varphi}(a_i,z)].\end{aligned}$$

i.e., the strict properness will preserve, under the noisy estimations. From above results we also observe that a larger Δ_φ will allow more noisy estimations. We thus can trade more payment with sample complexity, via designing the strictly proper scoring functions to increase Δ_φ.

Now consider the prediction elicitation case. Suppose that $S(p,y)$ is strictly concave w.r.t. p $\forall y$ with parameter λ.

$$\tilde{S}(a_i,y) = S(a_i,y) + \epsilon(a_i,y)\ ,$$

where $\epsilon(a_i,y)$ indicates the error term. First notice that

$$\mathbb{E}[S(a_i,y)] - \mathbb{E}[S(a'_i,y)] \geq \lambda |a_i - a'_i|.$$

Further we notice that

$$\begin{aligned}\epsilon(a_i,y) &= \tilde{S}(a_i,y) - S(a_i,y) \\ =& \Big(\frac{1-\tilde{e}_{1-y,z}}{1-\tilde{e}_{1,z}-\tilde{e}_{0,z}} - \frac{1-e_{1-y,z}}{1-e_{1,z}-e_{0,z}}\Big) S(a_i,y) \\ &- \Big(\frac{e_{y,z}}{1-e_{1,z}-e_{0,z}} - \frac{e_{y,z}}{1-e_{1,z}-e_{0,z}}\Big) S(a_i,-y).\end{aligned}$$

Due to the sample complexity results we know that with probability at least $1-\delta$ that

$$\begin{aligned}&\Big|\frac{1-\tilde{e}_{1-y,z}}{1-\tilde{e}_{1,z}-\tilde{e}_{0,z}} - \frac{1-e_{1-y,z}}{1-e_{1,z}-e_{0,z}}\Big| \leq \epsilon_1, \\ &\Big|\frac{e_{y,z}}{1-e_{1,z}-e_{0,z}} - \frac{e_{y,z}}{1-e_{1,z}-e_{0,z}}\Big| \leq \epsilon_1.\end{aligned}$$

Suppose $S(p,y)$ is also Lipschitz w.r.t. p $\forall y$ with parameter L. By Lipschitz conditions we know that

$$|\mathbb{E}[\epsilon(a_i,y)] - \mathbb{E}[\epsilon(a_i,y)]| \leq \epsilon_1 L \cdot |a_i - a'_i| + \frac{\delta}{\kappa^2} L \cdot |a_i - a'_i|.$$

Therefore when ϵ_1, δ are small enough such that

$$\epsilon_1 L + \frac{\delta}{\kappa^2} L < \lambda,$$

no deviation is profitable.

The $\epsilon(K,N) = \epsilon_1 L + \frac{\delta}{\kappa^2} L = O\Big(\frac{1}{N} + \sqrt{\frac{\log K}{K} + \frac{1}{K}}\Big)$-dominant strategy argument follows naturally from above error term analysis. We will not repeat the details. □

HETEROGENEOUS ERROR RATES ACROSS TASKS

In order to estimate error rates correctly, we needed to make the assumption that human workers have homogeneous error rates across multiple tasks. We now discuss the applicability of our method in light of the heterogeneity issue. Before starting, we would like to emphasize that in practice, for each set of experiments we can choose to group tasks according to their types (e.g., image labeling, solving puzzles, objects recognition), and run our mechanism over each group separately. For this setting, we can assume the homogeneity more comfortably.

Nonetheless when it is not quite possible to group tasks together, suppose that each human agent's error rates are also task contingent in that for each possible task x with $y \sim \mathcal{P} := \{\mathcal{P}_0, \mathcal{P}_1\}$ we have $e_{0,i}(y), e_{1,i}(y)$. Redefine the error rate of the reference answer as follows:

$$e_{1,z} = \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=1}[\tilde{e}_{1,j}(x)]}{N-1},\ e_{0,z} = \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=0}[\tilde{e}_{0,j}(x)]}{N-1}.$$

Again $e_{1,z}, e_{0,z}$ captures the error rate of the reference answer. With knowing $e_{1,z}, e_{0,z}$, the dominant strategy argument in Theorem 3.4 holds.

So far so good. However it becomes less clear in how to estimate $e_{1,z}, e_{0,z}$. If we follow the idea in Section 4.2, the first order equation holds as before

$$\mathcal{P}_0 \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=0}[\tilde{e}_{0,j}(x)]}{N-1} + \mathcal{P}_1 (1 - \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=1}[\tilde{e}_{1,j}(x)]}{N-1}) = P_{1,-i},$$

i.e., $\mathcal{P}_0 e_{0,z} + \mathcal{P}_1(1 - e_{1,z}) = P_{1,-i}$. However the higher order matching statistics appear to be very different. For example, the second order matching probability between two agents on label 1 becomes

$$\mathcal{P}_0 \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=0}[\tilde{e}^2_{0,j}(x)]}{N-1} + \mathcal{P}_1 \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=1}[1-\tilde{e}^2_{1,j}(x)]}{N-1} = P_{1,-i},$$

Instead we would like a equation as a function of

$$\left(\frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=0}[\tilde{e}_{0,j}(x)]}{N-1}\right)^2, \left(\frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=1}[1-\tilde{e}_{1,j}(x)]}{N-1}\right)^2$$

in order to identify the true error rates on average. To get around of this issue, we can adopt the following approximation for $\tilde{e}^2_{0,j}(x)$ using Taylor expansion:

$$\tilde{e}^2_{0,j}(x) \approx e^2_{0,z} + 2e_{0,z}(\tilde{e}_{0,j}(x) - e_{0,z}) = 2e_{0,z} \cdot \tilde{e}_{0,j}(x) - e^2_{0,z}.$$

Then we claim that

$$\begin{aligned}\frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=0}[\tilde{e}^2_{0,j}(x)]}{N-1} &\approx \frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=0}[2e_{0,z} \cdot \tilde{e}_{0,j}(x) - e^2_{0,z}]}{N-1} \\ &= 2e_{0,z} \cdot e_{0,z} - e^2_{0,z} = e^2_{0,z}\end{aligned}$$

Similarly we have

$$\frac{\sum_{j\neq i} \mathbb{E}_{x,y|y=1}[1-\tilde{e}^2_{1,j}(x)]}{N-1} \approx (1-e_{1,z})^2.$$

From above, we see we are able to recover the matching equations defined in our earlier solution for task homogeneous cases; and it can be viewed as a linear approximation for this task contingent case.

PROOF OF LEMMA 5.1

Proof. W.l.o.g., consider $y = 0$. We first derive the following difference term

$$\begin{aligned}&\Pr[z_1 = 1, z_2 = 1|y = 0] - \Pr[z_1 = 1|y = 0]\Pr[z_2 = 1|y = 0] \\ &= (\Pr[z_1 = 1|z_2 = 1, y = 0] - \Pr[z_1 = 1|y = 0])\Pr[z_2 = 1|y = 0] \\ &= (\Pr[z_1 = 1|z_2 = 1, y = 0] - \Pr[z_1 = 1|z_2 = 1, y = 0]\Pr[z_2 = 1|y = 0] \\ &\quad - \Pr[z_1 = 1|z_2 = 0, y = 0]\Pr[z_2 = 0|y = 0])\Pr[z_2 = 1|y = 0] \\ &= \Pr[z_2 = 0|y = 0]\Pr[z_2 = 1|y = 0](\Pr[z_1 = 1|z_2 = 1, y = 0] - \Pr[z_1 = 1|z_2 = 0, y = 0]).\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\Pr[z_1 = 0, z_2 = 0|y = 0] - \Pr[z_1 = 0|y = 0]\Pr[z_2 = 0|y = 0] \\
&= (\Pr[z_1 = 0|z_2 = 0, y = 0] - \Pr[z_1 = 0|y = 0])\Pr[z_2 = 0|y = 0] \\
&= (\Pr[z_1 = 0|z_2 = 0, y = 0] - \Pr[z_1 = 0|z_2 = 0, y = 0]\Pr[z_2 = 0|y = 0] \\
&\quad - \Pr[z_1 = 0|z_2 = 1, y = 0]\Pr[z_2 = 1|y = 0])\Pr[z_2 = 0|y = 0] \\
&= \Pr[z_2 = 0|y = 0]\Pr[z_2 = 1|y = 0](\Pr[z_1 = 0|z_2 = 1, y = 0] - \Pr[z_1 = 0|z_2 = 0, y = 0]).
\end{aligned}$$

The above two terms add up to 0. The proof for $y = 1$ is symmetric. Proved. □

PROOF OF LEMMA 5.2

Proof.

$$\begin{aligned}
&\Pr[z_1 = z_2, z_3 = 1] - \Pr[z_1 = z_2]\Pr[z_1 = 1] = \mathcal{P}_0(x_1^2 + (1 - x_1)^2)x_1 + \mathcal{P}_1(x_2^2 + (1 - x_2)^2)x_2 \\
&- \left(\mathcal{P}_0(x_1^2 + (1 - x_1)^2) + \mathcal{P}_1(x_2^2 + (1 - x_2)^2)\right)(\mathcal{P}_0 x_1 + \mathcal{P}_1 x_2) \\
&= \mathcal{P}_0\mathcal{P}_1(x_1 - x_2)^2(2(x_1 + x_2) - 1).
\end{aligned} \tag{18}$$

Similarly with knowing $|x_1 - x_2|$ (jointly determined by the first order and second order equations, as detailed in Lemma 4.2), we are able to formulate another equation regarding x_1, x_2 that

$$x_1 + x_2 = \frac{\Pr[z_1 = z_2, z_3 = 1] - \Pr[z_1 = z_2]\Pr[z_1 = 1]}{2\mathcal{P}_0\mathcal{P}_1(x_1 - x_2)^2} + \frac{1}{\mathcal{P}_0\mathcal{P}_1(x_1 - x_2)^2}.$$

□

PROOF OF THEOREM 5.3

Proof. We only need to prove for the case when $s_i \neq s_i^l$, there is no incentive for agents to report s_i^l. With accurately learning $e_{0,z}, e_{1,z}$, we know

$$\mathbb{E}[\varphi(a_i, z)|s_i, s_i^l] = \mathbb{E}[S(a_i, y)|s_i, s_i^l] = \frac{\Pr[y = a_i|s_i, s_i^l]}{\Pr[y = a_i]}$$

Suppose $s_i = 1 - \hat{s}, s_i^l = \hat{s}$. Next we prove that

$$\Pr[y = 1 - \hat{s}|s_i = 1 - \hat{s}, s_i^l = \hat{s}] > \mathcal{P}_{1-\hat{s}}$$
$$\Pr[y = \hat{s}|s_i = 1 - \hat{s}, s_i^l = \hat{s}] < \mathcal{P}_{\hat{s}}$$

following which, our claim is clearly true. First

$$\begin{aligned}
\Pr[y = \hat{s}|s_i = 1 - \hat{s}, s_i^l = \hat{s}] &= \frac{\Pr[y = \hat{s}, s_i = 1 - \hat{s}, s_i^l = \hat{s}]}{\Pr[s_i = 1 - \hat{s}, s_i^l = \hat{s}]} \\
&= \frac{\Pr[s_i = 1 - \hat{s}, s_i^l = \hat{s}|y = \hat{s}]\mathcal{P}_{y=\hat{s}}}{\Pr[s_i = 1 - \hat{s}, s_i^l = \hat{s}]}.
\end{aligned}$$

Then

$$\begin{aligned}
&\Pr[y = \hat{s}|s_i = 1 - \hat{s}, s_i^l = \hat{s}] < \mathcal{P}_{\hat{s}} \\
\Leftrightarrow &\Pr[s_i = 1 - \hat{s}, s_i^l = \hat{s}|y = \hat{s}] < \Pr[s_i = 1 - \hat{s}, s_i^l = \hat{s}] \\
\Leftrightarrow &\Pr[s_i = 1 - \hat{s}|y = \hat{s}] \Pr[s_i^l = \hat{s}|y = \hat{s}] < \sum_{\hat{y}} \mathcal{P}_{\hat{y}} \cdot \Pr[s_i = 1 - \hat{s}|y = \hat{y}] \Pr[s_i^l = \hat{s}|y = \hat{y}] \\
\Leftrightarrow &\Pr[s_i = 1 - \hat{s}|y = \hat{s}](1 - \Pr[s_i^l = 1 - \hat{s}|y = \hat{s}]) < (1 - \Pr[s_i = 1 - \hat{s}|y = \hat{s}]) \Pr[s_i^l = 1 - \hat{s}|y = \hat{s}] \\
\Leftrightarrow &\Pr[s_i = 1 - \hat{s}|y = \hat{s}] < \Pr[s_i^l = 1 - \hat{s}|y = \hat{s}],
\end{aligned}$$

which holds by the definition of weak signal (a higher error rate). □

A MACHINE LEARNING METHOD

Suppose that each task is associated with a corresponding feature vector. Consider the current task that needs to be elicited and denote its feature vector as $\mathbf{x} \in \mathbb{R}^d$. Suppose we have learned (following the results in [Liu and Chen, 2017], such a classifier is learnable purely from agent's reported noisy data) a good classifier $\tilde{f}^*(\mathbf{x})$ for predicting its true outcome. Our ML-aided SSR simply works in the following ways: replacing the reference answer z with $\tilde{f}^*(\mathbf{x})$ and plug in its error rates. The rest of job is to reason about $e_{1,z}, e_{0,z}$ (for $\tilde{f}^*(\mathbf{x})$). There are possibly many different ways of doing so. We demonstrate its possibility when there is again no ground truth label being available with the following simple estimation procedure:

- Assign K tasks $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_K$ to three randomly drawn peers, denote them as $z_i^{\text{ref}}(\mathbf{x}_k), i = 1, 2, 3,\ k = 1, 2, ..., K$.
- Estimate $\Pr[\tilde{f}^*(\mathbf{x}_i) = z_1^{\text{ref}}(\mathbf{x}_i) = 1]$ and $\Pr[\tilde{f}^*(\mathbf{x}_i) = 1]$.
- Estimate $\Pr[z_1^{\text{ref}}(\mathbf{x}_i) = 1|y = 0]$ (follow Mechanism 1) using the three reference answers $z_i^{\text{ref}}(\mathbf{x}_k), i = 1, 2, 3,\ k = 1, 2, ..., K$.

Then note the following fact:

$$\begin{aligned}
\Pr[\tilde{f}^*(\mathbf{x}_i) = z_1^{\text{ref}}(\mathbf{x}_i) = 1] = &\mathcal{P}_0 \Pr[\tilde{f}^*(\mathbf{x}_i) = 1|y = 0] \Pr[z_1^{\text{ref}}(\mathbf{x}_i) = 1|y = 0] \\
&+ \mathcal{P}_1 \Pr[\tilde{f}^*(\mathbf{x}_i) = 1|y = 1] \Pr[z_1^{\text{ref}}(\mathbf{x}_i) = 1|y = 1]
\end{aligned}$$

Together with

$$\Pr[\tilde{f}^*(\mathbf{x}_i) = 1] = \mathcal{P}_0 \Pr[\tilde{f}^*(\mathbf{x}_i) = 1|y = 0] + \mathcal{P}_1 \Pr[\tilde{f}^*(\mathbf{x}_i) = 1|y = 1],$$

we will be able to solve for $e_{1,z}, e_{0,z}$ when we have a good estimates of $\Pr[\tilde{f}^*(\mathbf{x}_i) = z_1^{\text{ref}}(\mathbf{x}_i) = 1]$ and $\Pr[\tilde{f}^*(\mathbf{x}_i) = 1]$, and that $\Pr[z_1^{\text{ref}}(\mathbf{x}_i) = 1|y = 0] \neq \Pr[z_1^{\text{ref}}(\mathbf{x}_i) = 1|y = 1]$. With above argument, we see that we will be able to achieve dominant strategy scorings without even the need of reassigning all tasks.