# Sample Elicitation

**Jiaheng Wei***
UC Santa Cruz
jiahengwei@ucsc.edu

**Zuyue Fu***
Northwestern University
zuyue.fu@u.northwestern.edu

**Yang Liu**
UC Santa Cruz
yangliu@ucsc.edu

**Xingyu Li**
UC Santa Cruz
xli279@ucsc.edu

**Zhuoran Yang**
Princeton University
zy6@princeton.edu

**Zhaoran Wang**
Northwestern University
zhaoranwang@gmail.com

## Abstract

It is important to collect credible training samples $(x, y)$ for building data-intensive learning systems (e.g., a deep learning system). Asking people to report complex distribution $p(x)$, though theoretically viable, is challenging in practice. This is primarily due to the cognitive loads required for human agents to form the report of this highly complicated information. While classical elicitation mechanisms apply to eliciting a complex and generative (and continuous) distribution $p(x)$, we are interested in eliciting samples $x_i \sim p(x)$ from agents directly. We coin the above problem *sample elicitation*. This paper introduces a deep learning aided method to incentivize credible sample contributions from self-interested and rational agents. We show that with an accurate estimation of a certain $f$-divergence function we can achieve approximate incentive compatibility in eliciting truthful samples. We then present an efficient estimator with theoretical guarantees via studying the variational forms of the $f$-divergence function. We also show a connection between this sample elicitation problem and $f$-GAN, and how this connection can help reconstruct an estimator of the distribution based on collected samples. Experiments on synthetic data, MNIST, and CIFAR-10 datasets demonstrate that our mechanism elicits truthful samples. Our implementation is available at `https://github.com/weijiaheng/Credible-sample-elicitation.git`.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

## 1 Introduction

The availability of a large number of credible samples is crucial for building high-fidelity machine learning models. This is particularly true for deep learning systems that are data-hungry. Arguably, the most scalable way to collect a large amount of training samples is to crowdsource from a decentralized population of agents who hold relevant data. The most popular example is the build of ImageNet (Deng et al., 2009).

The main challenge in eliciting private information is to properly score reported information such that the self-interested agent who holds private information will be incentivized to report truthfully. Most of the existing works focused on eliciting simple categorical information, such as binary labels or multi-class categorical information. These solutions are not properly defined or scalable to more continuous or high-dimensional information elicitation tasks. For example, suppose that we are interested in collecting the calorie information of a set of food pictures, we can crowdsource to ask crowd workers to tell us how many calories are there in each particular image of food (how many calories in the hot dog shown in the picture). There exists no computable elicitation/scoring mechanism for reporting this more continuous spectrum of data (reported calorie).

In this work [1], we aim to collect credible samples from self-interested agents via studying the problem of *sample elicitation*. Instead of asking each agent to report the entire distribution $p$, we hope to elicit samples drawn from the distribution $\mathbb{P}$ truthfully. We consider the samples $x_p \sim \mathbb{P}$ and $x_q \sim \mathbb{Q}$. In analogy to strictly

[1]Correspondence to: {yangliu, jiahengwei}@ucsc.edu.

proper scoring rules[2], we aim to design a score function $S$ s.t. $\mathbb{E}_{x\sim\mathbb{P}}[S(x_p, x')] > \mathbb{E}_{x\sim\mathbb{P}}[S(x_q, x')]$ for any $q \neq p$, where $x'$ is a reference answer that can be defined using elicited reports.

Our challenge lies in accurately evaluating reported samples. We first observe that the $f$-divergence function between two properly defined distributions of the samples can serve the purpose of incentivizing truthful reports of samples. We proceed with using deep learning techniques to solve the score function design problem via a data-driven approach. We then propose a variational approach that enables us to estimate the divergence function efficiently using reported samples, via a variational form of the $f$-divergence function, through a deep neural network. These estimation results help us establish approximate incentive compatibility in eliciting truthful samples. It is worth noting that our framework also generalizes to the setting where there is no access to ground truth samples and we can only rely on reported samples. There we show that our estimation results admit an approximate Bayesian Nash Equilibrium for agents to report truthfully. Furthermore, in our estimation framework, we use a generative adversarial approach to reconstruct the distribution from the elicited samples. In addition to the analytical results, we demonstrate the effectiveness of our mechanism in eliciting truthful samples empirically using MNIST and CIFAR-10 datasets.

We want to emphasize that the deep learning based estimators considered above can handle complex data. With our deep learning solution, we are further able to provide estimates for the divergence functions used for our scoring mechanisms with provable finite sample complexity. In this paper, we focus on developing theoretical guarantees - other parametric families either can not handle complex data, e.g., it is hard to handle images using kernel methods, or do not have provable guarantees on the sample complexity.

Our results complement the elicitation task in crowdsourcing by providing a method to elicit feature data $X \sim \mathbb{P}(X|Y)$, as compared to previous works mainly focusing on eliciting labels $Y \sim \mathbb{P}(Y|X)$. The difference is previous works focus on eliciting a label for a particular image, but our method enables elicitation of an image in response to a particular label (suppose that we are interested in collecting training images that contain "Cats"), which is inherently a more complex piece of information to evaluate and score with. This capability can help us build high-quality datasets for more comprehensive applications from scratch.

**Related work.** The most relevant literature to our paper is *strictly proper scoring rules* and *property elicitation*. Scoring rules were developed for eliciting truthful prediction (probability) (Brier, 1950; Winkler, 1969; Savage, 1971; Matheson and Winkler, 1976; Jose et al., 2006; Gneiting and Raftery, 2007). Characterization results for strictly proper scoring rules are given in McCarthy (1956); Savage (1971); Gneiting and Raftery (2007). Property elicitation notices the challenge of eliciting complex distributions (Lambert et al., 2008; Steinwart et al., 2014; Frongillo and Kash, 2015b). For instance, Abernethy and Frongillo (2012) characterize the score functions for eliciting linear properties, and Frongillo and Kash (2015a) study the complexity of eliciting properties. Another line of relevant research is peer prediction, where solutions can help elicit private information when the ground truth verification might be missing (De Alfaro et al., 2016; Gao et al., 2016; Kong et al., 2016; Kong and Schoenebeck, 2018, 2019). Our work complements the information elicitation literature via studying the question of sample elicitation using a variational approach to estimate $f$-divergences. A parallel work has also studied the variational approach for eliciting truthful information Schoenebeck and Yu (2020). Our work focuses more on formalizing the sample elicitation problem. In addition, we provide sample complexity guarantees to our theorems by offering deep neural network-aided estimators and contribute to the community a practical solution.

Our work is also related to works on divergence estimation. The simplest way to estimate divergence starts with the estimation of the density function (Wang et al., 2005; Lee and Park, 2006; Wang et al., 2009; Zhang and Grabchak, 2014; Han et al., 2016). Another method based on the variational form (Donsker and Varadhan, 1975) of the divergence function comes into play (Broniatowski and Keziou, 2004, 2009; Nguyen et al., 2010; Kanamori et al., 2011; Ruderman et al., 2012; Sugiyama et al., 2012), where the estimation of divergence is modeled as the estimation of density ratio between two distributions. The variational form of the divergence function also motivates the well-known Generative Adversarial Network (GAN) (Goodfellow et al., 2014), which learns the distribution by minimizing the Kullback-Leibler divergence. Follow-up works include Nowozin et al. (2016); Arjovsky et al. (2017); Gulrajani et al. (2017); Bellemare et al. (2017), with theoretical analysis in Liu et al. (2017); Arora et al. (2017); Liang (2018); Gao et al. (2019). See also Gao et al. (2017); Bu et al. (2018) for this line of work.

**Notations.** For the distribution $\mathbb{P}$, we denote by $\mathbb{P}_n$ the empirical distribution given a set of samples $\{x_i\}_{i=1}^n$ following $\mathbb{P}$, i.e., $\mathbb{P}_n = 1/n \cdot \sum_{i=1}^n \delta_{x_i}$, where $\delta_{x_i}$ is the Dirac measure at $x_i$. We denote by $\|v\|_s = (\sum_{i=1}^d |v^{(i)}|^s)^{1/s}$ the $\ell_s$ norm of the vec-

[2] Our specific formulation and goal will be different in details.

tor $v \in \mathbb{R}^d$ where $1 \leq s < \infty$ and $v^{(i)}$ is the $i$-th entry of $v$. We also denote by $\|v\|_\infty = \max_{1\leq i\leq d} |v^{(i)}|$ the $\ell_\infty$ norm of $v$. For any real-valued continuous function $f\colon \mathcal{X} \to \mathbb{R}$, we denote by $\|f\|_{L_s(\mathbb{P})} := [\int_{\mathcal{X}} |f(x)|^s \mathrm{d}\mathbb{P}]^{1/s}$ the $L_s(\mathbb{P})$ norm of $f$ and $\|f\|_s := [\int_{\mathcal{X}} |f(x)|^s \mathrm{d}\mu]^{1/s}$ the $L_s(\mu)$ norm of $f(\cdot)$, where $\mu$ is the Lebesgue measure. Also, we denote by $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$ the $L_\infty$ norm of $f(\cdot)$. For any real-valued functions $g(\cdot)$ and $h(\cdot)$ defined on some unbounded subset of the real positive numbers, such that $h(\alpha)$ is strictly positive for all large enough values of $\alpha$, we write $g(\alpha) \lesssim h(\alpha)$ and $g(\alpha) = \mathcal{O}(h(\alpha))$ if $|g(\alpha)| \leq c \cdot h(\alpha)$ for some positive absolute constant $c$ and any $\alpha > \alpha_0$, where $\alpha_0$ is a real number. We denote by $[n]$ the set $\{1, 2, \ldots, n\}$.

## 2 Preliminary

### 2.1 Sample Elicitation

We consider two scenarios. We start with an easier case where we, as the mechanism designer, have access to a certain number of group truth samples. Then we move to the harder case where the inputs to our mechanism can only be elicited samples from agents.

**Multi-sample elicitation with ground truth samples.** Suppose that the agent holds $n$ samples, with each of them independently drawn from $\mathbb{P}$, i.e., $x_i \sim \mathbb{P}$ [3] for $i \in [n]$. The agent can report each sample arbitrarily, which is denoted as $r_i(x_i) : \Omega \to \Omega$. There are $n$ data $\{x_i^*\}_{i\in[n]}$ independently drawn from the ground truth distribution $\mathbb{Q}$[4]. $x_i$s and $x_i^*$s are often correlating with each other. For example, $x_i^*$ corresponds to the true calorie level of the food contained in a picture. $x_i$ is the corresponding guess from the agent, possibly as a (randomized) function of $x_i^*$. Therefore the two distributions $\mathbb{P}$ and $\mathbb{Q}$ are not independent in general.

We are interested in designing a score function $S(\cdot)$ that takes inputs of each $r_i(\cdot)$ and $\{r_j(x_j), x_j^*\}_{j\in[n]}$: $S(r_i(x_i), \{r_j(x_j), x_j^*\}_{j\in[n]})$ such that if the agent believes that $x^*$ is drawn from the same distribution $x^* \sim \mathbb{P}$, then for any $\{r_j(\cdot)\}_{j\in[n]}$, it holds with probability at least $1 - \delta$:

$$\sum_{i=1}^n \mathbb{E}_{x,x^*\sim\mathbb{P}}\Big[S\big(x_i, \{x_j, x_j^*\}_{j\in[n]}\big)\Big] \geq \sum_{i=1}^n \mathbb{E}_{x,x^*\sim\mathbb{P}}\Big[S\big(r_i(x_i), \{r_j(x_j), x_j^*\}_{j\in[n]}\big)\Big] - n\cdot\epsilon.$$

We name the above as $(\delta, \epsilon)$**-properness** (per sample) for sample elicitation. When $\delta = \epsilon = 0$, it is reduced to the one that is similar to the properness definition in scoring rule literature (Gneiting and Raftery, 2007). We also shorthand $r_i = r_i(x_i)$ when there is no confusion. Agent believes that her samples are generated from the same distribution as that of the ground truth samples, i.e., $\mathbb{P}$ and $\mathbb{Q}$ are the same distributions.

**Sample elicitation with peer samples.** Suppose there are $n$ agents each holding a sample $x_i \sim \mathbb{P}_i$, where the distributions $\{\mathbb{P}_i\}_{i\in[n]}$ are not necessarily the same - this models the fact that agents can have subjective biases or local observation biases. This is a more standard peer prediction setting. We denote by their joint distribution as $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2 \times .... \times \mathbb{P}_n$.

Similar to the previous setting, each agent can report her sample arbitrarily, which is denoted as $r_i(x_i) : \Omega \to \Omega$ for any $i \in [n]$. We are interested in designing and characterizing a score function $S(\cdot)$ that takes inputs of each $r_i(\cdot)$ and $\{r_j(x_j)\}_{j\neq i}$: $S(r_i(x_i), \{r_j(x_j)\}_{j\neq i})$ such that for any $\{r_j(\cdot)\}_{j\in[n]}$, it holds with probability at least $1 - \delta$ that

$$\mathbb{E}_{x\sim\mathbb{P}}\Big[S\big(x_i, \{r_j(x_j) = x_j\}_{j\neq i}\big)\Big] \geq \mathbb{E}_{x\sim\mathbb{P}}\Big[S\big(r(x_i), \{r_j(x_j) = x_j\}_{j\neq i}\big)\Big] - \epsilon.$$

We name the above as $(\delta, \epsilon)$**-Bayesian Nash Equilibrium** (BNE) in truthful elicitation. We only require that agents are all aware of the above information structure as common knowledge, but they do not need to form beliefs about details of other agents' sample distributions. Each agent's sample is private to herself.

**Connection to the proper scoring rule** At a first look, this problem of eliciting quality data is readily solvable with the seminal solution for eliciting distributional information, called the strictly proper scoring rule (Brier, 1950; Winkler, 1969; Savage, 1971; Matheson and Winkler, 1976; Jose et al., 2006; Gneiting and Raftery, 2007): suppose we are interested in eliciting information about a random vector $X = (X_1, ..., X_{d-1}, Y) \in \Omega \subseteq \mathbb{R}^d$, whose probability density function is denoted by $p$ with distribution $\mathbb{P}$. As the mechanism designer, if we have a sample $x$ drawn from the true distribution $\mathbb{P}$, we can apply strictly proper scoring rules to elicit $p$: the

[3] Though we use $x$ to denote the samples we are interested in, $x$ potentially includes both the feature and labels $(x, y)$ as in the context of supervised learning.

[4] The number of ground truth samples can be different from $n$, but we keep them the same for simplicity of presentation. It will mainly affect the terms $\delta$ and $\epsilon$ in our estimations.

agent who holds $p$ will be scored using $S(p, x)$. $S$ is called strictly proper if it holds for any $p$ and $q$ that $\mathbb{E}_{x\sim\mathbb{P}}[S(p, x)] > \mathbb{E}_{x\sim\mathbb{P}}[S(q, x)]$. The above elicitation approach has two main caveats that limited its application: (1) When the outcome space $|\Omega|$ is large and is even possibly infinite, it is practically impossible for any human agents to report such a distribution with reasonable efforts. Consider the example where we are interested in building an image classifier via first collecting a certain category of high-dimensional image data. While classical elicitation results apply to eliciting a complex, generative and continuous distribution $p(x)$ for this image data, we are interested in eliciting samples $x_i \sim p(x)$ from agents. (2) The mechanism designer may not possess any ground truth samples.

### 2.2 $f$-divergence

It is well known that maximizing the expected proper scores is equivalent to minimizing a corresponding Bregman divergence (Gneiting and Raftery, 2007). More generically, we take the perspective that divergence functions have great potentials to serve as score functions for eliciting samples. We define the $f$-divergence between two distributions $\mathbb{P}$ and $\mathbb{Q}$ with probability density function $p$ and $q$ as

$$D_f(q\|p) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu. \tag{2.1}$$

Here $f(\cdot)$ is a function satisfying certain regularity conditions, which will be specified later. Solving our elicitation problem involves evaluating the $D_f(q\|p)$ successively based on the distributions $\mathbb{P}$ and $\mathbb{Q}$, without knowing the probability density functions $p$ and $q$. Therefore, we have to resolve to a form of $D_f(q\|p)$ which does not involve the analytic forms of $p$ and $q$, but instead sample forms. Following from Fenchel's convex duality, it holds that

$$D_f(q\|p) = \max_{t(\cdot)} \mathbb{E}_{x\sim\mathbb{Q}}[t(x)] - \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))], \tag{2.2}$$

where $f^\dagger(\cdot)$ is the Fenchel duality of the function $f(\cdot)$, which is defined as $f^\dagger(u) = \sup_{v\in\mathbb{R}}\{uv - f(v)\}$, and the max is taken over all functions $t(\cdot)\colon \Omega \subset \mathbb{R}^d \to \mathbb{R}$.

## 3 Sample Elicitation: A Variational Approach

Recall from (2.2) that $D_f(q\|p)$ admits the following variational form:

$$D_f(q\|p) = \max_{t(\cdot)} \mathbb{E}_{x\sim\mathbb{Q}}[t(x)] - \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))]. \tag{3.1}$$

We highlight that via functional derivation, (3.1) is solved by $t^*(x; p, q) = f'(\theta^*(x; p, q))$, where $\theta^*(x; p, q) = q(x)/p(x)$ is the density ratio between $p$ and $q$. Our elicitation builds upon such a variational form (3.1) and the following estimators,

$$\widehat{t}(\cdot; p, q) = \operatorname*{argmin}_{t(\cdot)} \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{x\sim\mathbb{Q}_n}[t(x)],$$
$$\widehat{D}_f(q\|p) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\widehat{t}(x)] - \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(\widehat{t}(x))].$$

### 3.1 Error Bound and Assumptions

Suppose we have the following error bound for estimating $D_f(q\|p)$: for any probability density functions $p$ and $q$, it holds with probability at least $1-\delta(n)$ that

$$\left|\widehat{D}_f(q\|p) - D_f(q\|p)\right| \le \epsilon(n), \tag{3.2}$$

where $\delta(n)$ and $\epsilon(n)$ will be specified later in Section §4. To obtain such an error bound, we need the following assumptions.

**Assumption 3.1** (Bounded Density Ratio)**.** The density ratio $\theta^*(x; p, q) = q(x)/p(x)$ is bounded such that $0 < \theta_0 \le \theta^* \le \theta_1$ holds for positive absolute constants $\theta_0$ and $\theta_1$.

The above assumption is standard in related literature (Nguyen et al., 2010; Suzuki et al., 2008), which requires that the probability density functions $p$ and $q$ lie on the same support. For simplicity of presentation, we assume that this support is $\Omega \subset \mathbb{R}^d$. We define the $\beta$-Hölder function class on $\Omega$ as follows.

**Definition 3.2** ($\beta$-Hölder Function Class)**.** The $\beta$-Hölder function class with radius $M$ is defined as

$$\mathcal{C}_d^\beta(\Omega, M) = \bigg\{ t(\cdot)\colon \Omega \subset \mathbb{R}^d \to \mathbb{R}\colon \sum_{\|\alpha\|_1 < \beta} \|\partial^\alpha t\|_\infty + \sum_{\|\alpha\|_1 = \lfloor\beta\rfloor} \sup_{\substack{x,y\in\Omega,\\ x\neq y}} \frac{|\partial^\alpha t(x) - \partial^\alpha t(y)|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} \le M \bigg\},$$

where $\partial^\alpha = \partial^{\alpha_1}\cdots\partial^{\alpha_d}$ with $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$.

We impose the following assumptions.

**Assumption 3.3** ($\beta$-Hölder Condition)**.** The function $t^*(\cdot; p, q) \in \mathcal{C}_d^\beta(\Omega, M)$ for some positive absolute constants $M$ and $\beta$, where $\mathcal{C}_d^\beta(\Omega, M)$ is the $\beta$-Hölder function class in Definition 3.2.

**Assumption 3.4** (Regularity of Divergence Function)**.** The function $f(\cdot)$ is smooth on $[\theta_0, \theta_1]$ and $f(1) = 0$. Also, $f$ is $\mu_0$-strongly convex, and has $L_0$-Lipschitz continuous gradient on $[\theta_0, \theta_1]$, where $\mu_0$ and $L_0$ are positive absolute constants, respectively.

We highlight that we only require that the conditions in Assumption 3.4 hold on the interval $[\theta_0, \theta_1]$, where the absolute constants $\theta_0$ and $\theta_1$ are specified in Assumption 3.1. Thus, Assumption 3.4 is mild and it holds for many commonly used functions in the definition of $f$-divergence. For example, in Kullback-Leibler (KL) divergence, we take $f(u) = -\log u$, which satisfies Assumption 3.4.

We will show that under Assumptions 3.1, 3.3, and 3.4, the bound (3.2) holds. See Theorem 4.3 in Section §4 for details.

### 3.2 Multi-sample elicitation with ground truth samples

In this section, we focus on multi-sample elicitation with ground truth samples. Under this setting, as a reminder, the agent will report multiple samples. After the agent reported her samples, the mechanism designer obtains a set of ground truth samples $\{x_i^*\}_{i\in[n]} \sim \mathbb{Q}$ to serve the purpose of evaluation. This falls into the standard strictly proper scoring rule setting. Our mechanism is presented in Algorithm 1.

**Algorithm 1** $f$-scoring mechanism for multiple-sample elicitation with ground truth

1. Compute $\widehat{t}(\cdot; p, q) = \operatorname{argmin}_{t(\cdot)} \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{x^*\sim\mathbb{Q}_n}[t(x^*)]$.
2. For $i \in [n]$, pay reported sample $r_i$ using $S\big(r_i, \{r_j, x_j^*\}_{j=1}^n\big) := a - b\big(\mathbb{E}_{x\sim\mathbb{Q}_n}[\widehat{t}(x; p, q)] - f^\dagger\big(\widehat{t}(r_i; p, q)\big)\big)$ for some constants $a, b > 0$.

Algorithm 1 consists of two steps: Step 1 is to compute the function $\widehat{t}(\cdot; p, q)$, which enables us, in Step 2, to pay agent using a linear-transformed estimated divergence between the reported samples and the true samples. We have the following result.

**Theorem 3.5.** The $f$-scoring mechanism in Algorithm 1 achieves $(2\delta(n), 2b\epsilon(n))$-properness.

The proof is mainly based on the error bound in estimating $f$-divergence and its non-negativity. Not surprisingly, if the agent believes her samples are generated from the same distribution as the ground truth sample, and that our estimator can well characterize the difference between the two sets of samples, she will be incentivized to report truthfully to minimize the difference. We defer the proof to Section §10.1.

### 3.3 Single-task elicitation without ground truth samples

The above mechanism in Algorithm 1, while intuitive, has two caveats: 1. The agent needs to report multiple samples (multi-task/sample elicitation); 2. Multiple samples from the ground truth distribution are needed. To deal with such caveats, we consider the single point elicitation in an elicitation without a verification setting. Suppose there are $2n$ agents each holding a sample $x_i \sim \mathbb{P}_i$ [5]. We randomly partition the agents into two groups and denote the joint distributions for each group's samples as $\mathbb{P}$ and $\mathbb{Q}$ with probability density functions $p$ and $q$ for each of the two groups. Correspondingly, there are a set of $n$ agents for each group, respectively, who are required to report their *single* data point according to two distributions $\mathbb{P}$ and $\mathbb{Q}$, i.e., each of them holds $\{x_i^p\}_{i\in[n]} \sim \mathbb{P}$ and $\{x_i^q\}_{i\in[n]} \sim \mathbb{Q}$. As an interesting note, this is also similar to the setup of a Generative Adversarial Network (GAN), where one distribution corresponds to a generative distribution $x \mid y = 1$,[6] and another $x \mid y = 0$. This is a connection that we will further explore in Section §5 to recover distributions from elicited samples.

We denote by the joint distribution of $p$ and $q$ as $p \oplus q$ (distribution as $\mathbb{P}\oplus\mathbb{Q}$), and the product of the marginal distribution as $p\times q$ (distribution as $\mathbb{P}\times\mathbb{Q}$). We consider the divergence between the two distributions: $D_f(p \oplus q\|p \times q) = \max_{t(\cdot)} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}\oplus\mathbb{Q}}[t(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}\times\mathbb{Q}}[f^\dagger(t(\mathbf{x}))]$. Motivated by the connection between mutual information and KL divergence, we define generalized $f$-mutual information in the follows, which characterizes the generic connection between a generalized $f$-mutual information and $f$-divergence.

**Definition 3.6** (Kong and Schoenebeck (2019))**.** The generalized $f$-mutual information between $p$ and $q$ is defined as $I_f(p; q) = D_f(p \oplus q\|p \times q)$.

Further it is shown in Kong and Schoenebeck (2018, 2019) that the data processing inequality for mutual information holds for $I_f(p; q)$ when $f$ is strictly convex. We define the following estimators,

$$\begin{aligned}\widehat{t}(\cdot; p \oplus q, p \times q) =& \underset{t(\cdot)}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n}[f^\dagger(t(\mathbf{x}))] \\ &\quad - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n}[t(\mathbf{x})], \\ \widehat{D}_f(p \oplus q\|p \times q) =& \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n}[\widehat{t}(\mathbf{x}; p \oplus q, p \times q)] \\ &\quad - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n}[f^\dagger(\widehat{t}(\mathbf{x}; p \oplus q, p \times q))],\end{aligned} \tag{3.3}$$

where $\mathbb{P}_n$ and $\mathbb{Q}_n$ are empirical distributions of the reported samples. We denote $\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n \mid r_i$ as the conditional distribution when the first variable is fixed with realization $r_i$. Our mechanism is presented in Algorithm 2.

Similar to Algorithm 1, the main step in Algorithm 2 is to estimate the $f$-divergence between $\mathbb{P}_n \times \mathbb{Q}_n$ and $\mathbb{P}_n \oplus \mathbb{Q}_n$ using reported samples. Then we pay agents using a linear-transformed form of it. We have the following result.

[5] This choice of $2n$ is for the simplicity of presentation.

[6] "$\mid$" denotes the conditional distribution.

**Algorithm 2** $f$-scoring mechanism for sample elicitation

1: Compute $\widehat{t}(\cdot; p \oplus q, p \times q)$ as

$$\widehat{t}(\cdot) = \operatorname*{argmin}_{t(\cdot)} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n}[f^\dagger(t(\mathbf{x}))] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n}[t(\mathbf{x})].$$

2: Pay each reported sample $r_i$ using: for some constants $a, b > 0$,

$$S(r_i, \{r_j\}_{j\neq i}) := a + b\Big(\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n|r_i}\left[\widehat{t}(\mathbf{x}; p\oplus q, p\times q)\right] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n|r_i}\left[f^\dagger(\widehat{t}(\mathbf{x}; p\oplus q, p\times q))\right]\Big)$$

**Theorem 3.7.** The $f$-scoring mechanism in Algorithm 2 achieves $(2\delta(n), 2b\epsilon(n))$-BNE.

The theorem is proved by error bound in estimating $f$-divergence, a max argument, and the data processing inequality for $f$-mutual information. We defer the proof in Section §10.2.

The job left for us is to establish the error bound in estimating the $f$-divergence to obtain $\epsilon(n)$ and $\delta(n)$. Roughly speaking, if we solve the optimization problem (3.3) via deep neural networks with proper structure, it holds that $\delta(n) = 1 - \exp\{-n^{d/(2\beta+d)}\log^5 n\}$ and $\epsilon(n) = c\cdot n^{-\beta/(2\beta+d)}\log^{7/2} n$, where $c$ is a positive absolute constant. We state and prove this result formally in Section §4.

**Remark 3.8.** (1) When the number of samples grows, it holds that $\delta(n)$ and $\epsilon(n)$ decrease to 0 at least polynomially fast, and our guaranteed approximate incentive-compatibility approaches a strict one. (2) Our method or framework handles arbitrary complex information, where the data can be sampled from high dimensional continuous space. (3) The score function requires no prior knowledge. Instead, we design estimation methods purely based on reported sample data. (4) Our framework also covers the case where the mechanism designer has no access to the ground truth, which adds contribution to the peer prediction literature. So far peer prediction results focused on eliciting simple categorical information. Besides handling complex information structures, our approach can also be viewed as a data-driven mechanism for peer prediction problems.

# 4 Estimation of $f$-divergence

In this section, we introduce an estimator of $f$-divergence and establish the statistical rate of convergence, which characterizes $\epsilon(n)$ and $\delta(n)$. For the simplicity of presentation, in the sequel, we estimate the $f$-divergence $D_f(q\|p)$ between distributions $\mathbb{P}$ and $\mathbb{Q}$ with probability density functions $p$ and $q$, respectively. The rate of convergence of the estimated $f$-divergence can be easily extended to that of the estimated mutual information.

Following from the analysis in Section §3, by Fenchel duality, estimating $f$-divergence between $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to solving the following optimization problem,

$$t^*(\cdot; p, q) = \operatorname*{argmin}_{t(\cdot)} \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))] - \mathbb{E}_{x\sim\mathbb{Q}}[t(x)],$$
$$D_f(q\|p) = \mathbb{E}_{x\sim\mathbb{Q}}[t^*(x;p,q)] - \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t^*(x;p,q))]. \tag{4.1}$$

A natural way to estimate the divergence $D_f(q\|p)$ is to solve the empirical counterpart of (4.1):

$$t^\natural(\cdot; p, q) = \operatorname*{argmin}_{t\in\Phi} \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{x\sim\mathbb{Q}_n}[t(x)],$$
$$D_f^\natural(q\|p) = \mathbb{E}_{x\sim\mathbb{Q}_n}[t^\natural(x;p,q)] - \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t^\natural(x;p,q))], \tag{4.2}$$

where $\Phi$ is a function space with functions whose infinity norm is bounded by a constant $M$. We establish the statistical rate of convergence with general function space $\Phi$ as follows, and defer the case where $\Phi$ is a family of deep neural networks in Section §8 of the appendix. We introduce the following definition of the covering number.

**Definition 4.1** (Covering Number)**.** Let $(V, \|\cdot\|_{L_2})$ be a normed space, and $\Phi \subset V$. We say that $\{v_1,\ldots,v_N\}$ is a $\delta$-covering over $\Phi$ of size $N$ if $\Phi \subset \cup_{i=1}^N B(v_i,\delta)$, where $B(v_i,\delta)$ is the $\delta$-ball centered at $v_i$. The covering number is defined as $N_2(\delta,\Phi) = \min\{N : \exists\ \delta\text{-covering over } \Phi \text{ of size } N\}$.

We impose the following assumption on the covering number of the space $\Phi$, which characterizes the representation power of $\Phi$.

**Assumption 4.2.** $N_2(\delta,\Phi) = \mathcal{O}(\exp\{\delta^{-\gamma_\Phi}\})$, where $0 < \gamma_\Phi < 2$.

In the following theorem, we establish the statistical convergence rate of the estimator proposed in (4.2). For the simplicity of discussion, we assume that $t^* \in \Phi$.

**Theorem 4.3.** Suppose that Assumptions 3.1, 3.4, and 4.2 hold, and $t^*(\cdot; p, q) \in \Phi$. With probability at least $1 - \exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)})$, we have $|D_f^\natural(q\|p) - D_f(q\|p)| \lesssim n^{-1/(\gamma_\Phi+2)}$.

We defer the proof of Theorem 4.3 to Section §10.3. By Theorem 4.3, the estimator in (4.2) achieves the optimal non-parametric rate of convergence (Stone, 1982).

## 5 Connection to $f$-GAN and Reconstruction of Distribution

After sample elicitation, a natural question to ask is how to learn a representative probability density function from the samples. Denote the probability density function from elicited samples as $p$. Then, learning the probability density function $p$ is to solve for

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} D_f(q \| p), \tag{5.1}$$

where $\mathcal{Q}$ is the probability density function space. By the non-negativity of $f$-divergence, $q^* = p$ solves (5.1), which implies that by solving (5.1), we reconstruct the representative probability from the samples.

To see the connection between (5.1) and the formulation of $f$-GAN (Nowozin et al., 2016), by (2.2) and (5.1), we have $q^* = \operatorname{argmin}_{q \in \mathcal{Q}} \max_t \mathbb{E}_{x \sim \mathbb{Q}}[t(x)] - \mathbb{E}_{x \sim \mathbb{P}}[f^\dagger(t(x))]$, which is the formulation of $f$-GAN. We now propose the following estimator,

$$q^\natural = \operatorname*{argmin}_{q \in \mathcal{Q}} D_f^\natural(q \| p), \tag{5.2}$$

where $D_f^\natural(q \| p)$ is defined in (4.2). We defer the case where deep neural networks are used to construct the estimators in Section §8 of the appendix. We impose the following assumption.

**Assumption 5.1.** $N_2(\delta, \mathcal{Q}) = \mathcal{O}(\exp\{\delta^{-\gamma_\Phi}\})$.

The following theorem characterizes the error bound of estimating $q^*$ by $q^\natural$.

**Theorem 5.2.** Under the same assumptions in Theorem 4.3, further if Assumption 5.1 holds, for sufficiently large sample size $n$, with probability at least $1 - 1/n$, we have

$$D_f(q^\natural \| p) \lesssim n^{-1/(\gamma_\Phi + 2)} \cdot \log n + \min_{\widetilde{q} \in \mathcal{Q}} D_f(\widetilde{q} \| p).$$

We defer the proof of Theorem 5.2 to Section §10.5 in Appendix. In the upper bound of Theorem 5.2, the first term characterizes the generalization error of the estimator in (5.2), while the second term is the approximation error.

## 6 Experiment results

We use the synthetic dataset, MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) test dataset to validate the incentive property of our mechanism.

### 6.1 Experiments on synthetic data

In this section, the scores are estimated based on the variational approach we documented earlier and the method used in (Nguyen et al., 2010) for estimating the estimator. The experiments are based on synthetic data drawn from 2-dimensional Gaussian distributions [7]. We randomly generate 2 pairs of the means ($\mu$) and covariance matrices ($\Sigma$) (see Table 1 for details.) Experiment results show that truthful reports lead to higher scores that are close to analytical MI.

Table 1: Score comparison among truthful reports, random shift and random reports.

| | Exp1 | Exp2 |
|---|---|---|
| $\mu$ | $\begin{pmatrix} -2.97 \\ 8.98 \end{pmatrix}$ | $\begin{pmatrix} 6.98 \\ 8.39 \end{pmatrix}$ |
| $\Sigma$ | $\begin{pmatrix} 1.28 & 4.39 \\ 4.39 & 16.19 \end{pmatrix}$ | $\begin{pmatrix} 10.56 & 16.18 \\ 16.18 & 26.43 \end{pmatrix}$ |
| Analytical | 1.30 | 1.40 |
| Truthful | **1.36** ± 0.06 | **1.32** ± 0.05 |
| Random shift | 1.08 ± 0.05 | 1.11 ± 0.04 |
| Random report | 0.13 ± 0.02 | 0.20 ± 0.02 |

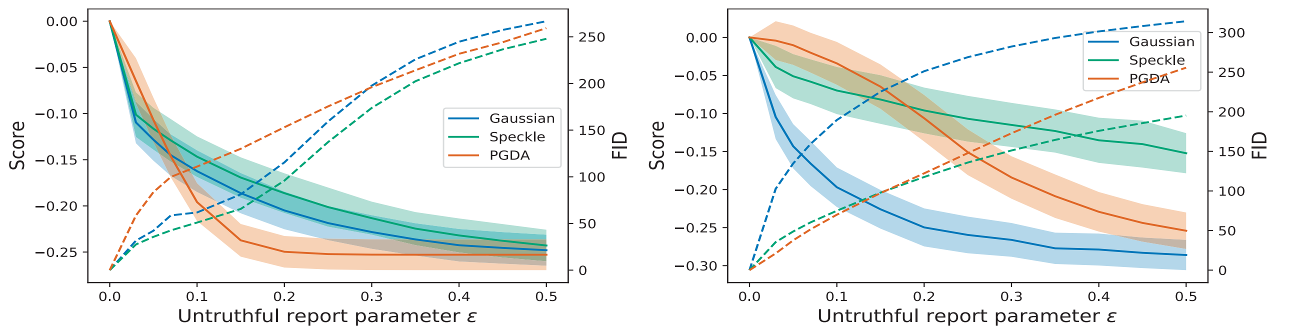
Two numerical experiments are shown in Table 1. In each of the above experiments, we draw 1000 pairs of samples $(x_i, y_i)$ from the corresponding Gaussian distribution. This set of pairs reflect the joint distribution $\mathbb{P} \oplus \mathbb{Q}$, while the sets $\{x_i\}$ and $\{y_i\}$ together correspond to the marginal distribution $\mathbb{P} \times \mathbb{Q}$. For each experiment, we repeat ten times and calculated the mean estimated score and the corresponding standard deviation.

For simplicity, we adopted $a = 0$, $b = 1$, and $f(\cdot) = -\log(\cdot)$ in estimating the scores, which make the expected score nothing but the Mutual Information (MI) between distributions $\mathbb{P}$ and $\mathbb{Q}$. The analytical values and the estimated scores of the MI for the two experiments are listed in the 4th and 5th row in Table 1, respectively. To demonstrate the effects of untruthful reports (misreports) on our score, we consider two types of untruthful reporting:

- **Random Shift:** The agent draws random noise from the uniform distribution $U(0, 3)$ and add to $\{x_i\}$.
- **Random Report;** Agent simply reports random signals drawn from the uniform distribution $U(0, 2\sigma)$, where $\sigma$ is the standard deviation of the marginal distribution $\mathbb{P}$. This models the case when agents contribute uninformative information.

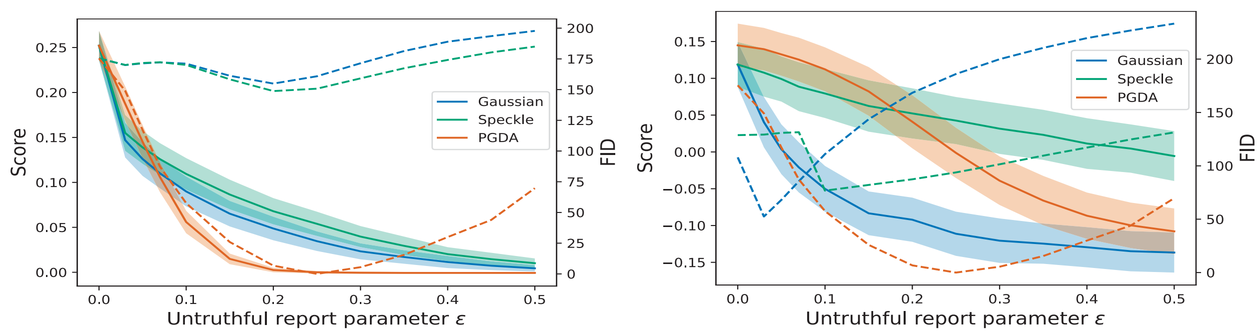
As expected, the scores of untruthful reports are generally lower than the scores of truthful ones.

[7] We choose simpler distribution so we can compute the scores analytically for verification purpose




(a) MNIST (b) CIFAR-10

Figure 1: Scores and FID value w.r.t. $\epsilon$ with ground truth verification. **Dashed lines** represent FID.




(a) MNIST (b) CIFAR-10

Figure 2: Scores and FID value w.r.t. $\epsilon$ with no ground truth verification but only peer samples. **Dashed lines** represent FID.

### 6.2 Experiments on Image data

We use the test dataset of MNIST and CIFAR-10 to further validate the robustness of our mechanism. Since the images are high-dimensional data, we choose to skip Step 1 in Algorithm 1 and 2 and instead adopt $\widehat{t}$ and $f^\dagger$ as suggested by (Nowozin et al., 2016) (please refer to Table 6 therein).

We take Total-Variation as an example in our experiments, and use (Nowozin et al., 2016) $\widehat{t}(x) = f^\dagger(\widehat{t}(x)) = \frac{1}{2}\tanh(x)$. We adopt $a = 0, b = 1$ for simplicity. For other divergences, please refer to Table 2 in the Appendix.

Untruthful reports are simulated by inducing the following three types of noise, with $\epsilon$ being a hyper-parameter controlling the degree of misreporting:

- **Gaussian Noise:** add Gaussian-distributed (mean=0, variance=$\epsilon$) additive noise in an image.
- **Speckle Noise:** add Speckle noise (mean=0, variance=$\epsilon$) in an image. Speckle noise (Van der Walt et al., 2014) is categorized into multiplicative noise of the clean image and Gaussian noised image.
- **Adversarial Attack:** use a pre-trained model to apply $L_\infty$ PGDAttack (PGDA) in an image. We adopt the default setting of (Ding et al., 2019) while replacing the hyper-parameter epsilon with our $\epsilon$.

Each test dataset consists of 10K images. As mentioned before, we use $\epsilon$ to denote the degree of untruthful reports by referring to truthful reports (clean images). For both MNIST and CIFAR-10, we split the dataset evenly into 200 groups (200 agents each holding a group of images). For $\epsilon \in \{0.03, 0.05, 0.07, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$, we assume each agent submits 50 images with $\epsilon-$noise to the central designer at one time. The mean and standard deviation of agents' scores are calculated concerning these 200 submissions.

Fréchet Inception Distance (FID score) (Heusel et al., 2017) is a widely accepted measure of similarity between two datasets of images. In our experiments, the FID score is considered as a measure of the truthfulness of agents' reports by referring to images for verification.

**Interpretation of the visualization** In Figure 1 and 2, the $x-$axis indicates the level of untruthful report—a large $\epsilon$ represents high-level untruthful (noisy) reports. The left $y-$axis is the score (given by our proposed method) of submitted $\epsilon-$level untruthful reports, while the right $y-$axis is the FID score. The curves visualize the relationship between the untruthful reports and the corresponding score/payment given by these two scoring methods. To validate the incentive property of our proposed score functions, the score of submitted reports is supposed to be monotonically decreasing w.r.t. the increasing $\epsilon-$level of untruthfulness. FID score has the incentive property if it is monotonically increasing w.r.t. the increasing $\epsilon-$level of untruthfulness.

**With ground truth verification** In this case, we consider the test images of MNIST and CIFAR-10 as ground truth images for verification. We report the average scores/payments of all 200 agents with their standard deviation. As shown in Figure 1, for untruthful reports using Gaussian and Speckle noise, a larger $\epsilon$ will lead to consistently lower score/payment, establishing the incentive-compatibility of our scoring mechanism. In this case, the untruthful report also leads to a higher FID score (less similarity) - we think this is an interesting observation implying that FID can also serve as a heuristic metric for evaluating image samples when we have ground truth verification. For PGDA untruthful reports, our mechanism is robust especially when $\epsilon$ is not too large.

**Without ground-truth verification** When we do not have access to the ground truth, we use only peer-reported images for verification. Again we report the average payment with the standard deviation. As shown in Figure 2, FID fails to continue to be a valid measure of truthfulness when we are using peer samples (reports) for verification. However, it is clear that our mechanism is robust to peer reports for verification: truthful reports result in a higher score.

## 7 Concluding Remarks

In this work, we introduce the problem of sample elicitation as an alternative to elicit complicated distribution. Our elicitation mechanism leverages the variational form of $f$-divergence functions to achieve accurate estimation of the divergences using samples. We provide the theoretical guarantee for both our estimators and the achieved incentive compatibility. Experiments on a synthetic dataset, MNIST, and CIFAR-10 test dataset further validate incentive properties of our mechanism. It remains an interesting problem to find out more “organic" mechanisms for sample elicitation that requires (i) less elicited samples; and (ii) induced strict truthfulness instead of approximated ones.

## Acknowledgement

This work is partially supported by the National Science Foundation (NSF) under grant IIS-2007951.

### References

Jacob D Abernethy and Rafael M Frongillo. A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, pages 27–1, 2012.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pages 224–232, 2017.

Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

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

Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences. Technical report, Citeseer, 2004.

Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, 2009.

Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of KL divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.

Luca De Alfaro, Michael Shavlovsky, and Vassilis Polychronopoulos. Incentives for truthful peer grading. *arXiv preprint arXiv:1604.03178*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. Advertorch v0.1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*, 2019.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. I. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Rafael Frongillo and Ian Kash. On elicitation complexity. In *Advances in Neural Information Processing Systems*, pages 3258–3266, 2015a.

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

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

Chao Gao, Yuan Yao, and Weizhi Zhu. Generative adversarial nets for robust scatter estimation: A proper scoring rule perspective. *arXiv preprint arXiv:1903.01944*, 2019.

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Density functional estimators with k-nearest neighbor bandwidths. In *International Symposium on Information Theory*, pages 1351–1355, 2017.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. *arXiv preprint arXiv:1605.09124*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

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

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. $f$-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2011.

Yuqing Kong and Grant Schoenebeck. Water from two rocks: Maximizing the mutual information. In *Conference on Economics and Computation*, pages 177–194, 2018.

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

Yuqing Kong, Katrina Ligett, and Grant Schoenebeck. Putting peer prediction under the micro (economic) scope and making truth-telling focal. In *International Conference on Web and Internet Economics*, pages 251–264. Springer, 2016.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

N.S. Lambert, D.M. Pennock, and Y. Shoham. Eliciting properties of probability distributions. In *Conference on Electronic Commerce*, pages 129–138, 2008.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Young Kyung Lee and Byeong U Park. Estimation of Kullback–Leibler divergence by local likelihood. *Annals of the Institute of Statistical Mathematics*, 58 (2):327–340, 2006.

Xingguo Li, Junwei Lu, Zhaoran Wang, Jarvis Haupt, and Tuo Zhao. On tighter generalization bound for deep neural networks: CNNs, ResNets, and beyond. *arXiv preprint arXiv:1806.05159*, 2018.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

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

John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42(9):654–655, 1956.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* MIT press, 2018.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11): 5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

Avraham Ruderman, Mark Reid, Darío García-García, and James Petterson. Tighter variational representations of $f$-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*, 2012.

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

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Grant Schoenebeck and Fang-Yi Yu. Learning and strongly truthful multi-task peer prediction: A variational approach. *arXiv preprint arXiv:2009.14730*, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Conference on Learning Theory*, pages 482–526, 2014.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pages 5–20, 2008.

Sara A van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2: e453, 2014.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via $k$-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

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

Zhiyi Zhang and Michael Grabchak. Nonparametric estimation of Kullback–Leibler divergence. *Neural computation*, 26(11):2570–2593, 2014.

Xingyu Zhou. On the Fenchel duality between strong convexity and Lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.