CREDIBLE SAMPLE ELICITATION BY DEEP LEARNING, FOR DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

It is important to collect credible training samples (x, y) for building data-intensive learning systems (e.g., a deep learning system). In the literature, there is a line of studies on eliciting distributional information from self-interested agents who hold a relevant information. Asking people to report complex distribution $p(x)$, though theoretically viable, is challenging in practice. This is primarily due to the heavy cognitive loads required for human agents to reason and report this high dimensional information. 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 and generative (and continuous) distribution $p(x)$ for this image data, we are interested in eliciting samples $x_i \sim p(x)$ from agents. This paper introduces a deep learning aided method to incentivize credible sample contributions from selfish and rational agents. The challenge to do so is to design an incentive-compatible score function to score each reported sample to induce truthful reports, instead of an arbitrary or even adversarial one. We show that with accurate estimation of a certain f-divergence function we are able to achieve approximate incentive compatibility in eliciting truthful samples. We then present an efficient estimator with theoretical guarantee via studying the variational forms of f-divergence function. Our work complements the literature of information elicitation via introducing the problem of *sample elicitation*. 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.

1 INTRODUCTION

The availability of a large quantity 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 for collecting a large amount of training samples is crowdsourcing from a decentralized population of agents who hold relevant sample information. 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 a private information will be incentivized to report truthfully. 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 & Winkler, 1976; Jose et al., 2006; Gneiting & Raftery, 2007): suppose we are interested in eliciting information about a random vector $X = [X_1, ..., X_{d-1}, Y] \in \Omega \subseteq \mathbb{R}^d$. Denote its distribution as p with measure $\mathbb{P}$. As the mechanism designer, if we have a sample $x \sim \mathbb{P}$ drawn from the true distribution, we can apply strictly proper scoring rules (SPSR) to elicit p: the agent who holds p will be scored using $S(p, x)$. S is called strictly proper if the following condition holds: $\mathbb{E}_{x\sim\mathbb{P}}[S(p, x)] > \mathbb{E}_{x\sim\mathbb{P}}[S(q, x)],\ \forall q \neq p$. The above elicitation approach has two main caveats that limited its application:

- 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. This partially inspired a line of follow-up works on eliciting property of the distributions, which we will discuss later.
- The mechanism designer may not possess any ground truth samples.

In this work we aim to collect credible samples from self-interested agents via studying the question of *sample elicitation*. Instead of asking each agent to report the complete distribution p, we hope to elicit samples drawn from the distribution $x \sim \mathbb{P}$ truthfully. Denote samples $x_p \sim \mathbb{P}, x_q \sim \mathbb{Q}$. In analogy to strictly proper scoring rules[1], we aim to design a score function S s.t. $\mathbb{E}_{x\sim\mathbb{P}}[S(x_p, x')] > \mathbb{E}_{x\sim\mathbb{P}}[S(x_q, x')],\ \forall q \neq p$, where x' is a reference answer that can be defined using elicited reports. This setting will relax the requirements of high reporting complexity, and has wide applications in collecting training samples for machine learning tasks. Indeed our goal resembles similarity to property elicitation (Lambert et al., 2008; Steinwart et al., 2014; Frongillo & Kash, 2015b), but we emphasize that our aims are different - property elicitation aims to elicit statistical properties of a distribution, while ours focus on eliciting samples drawn from the distributions. In certain scenarios, when agents do not have the complete knowledge or power to compute these properties, our setting enables elicitation of individual sample points.

Our challenge lies in accurately evaluating reported samples. We first observe that the f-divergence functions between two properly defined distributions of the samples can serve the purpose of incentivizing truthful report of samples. We then propose a variational approach that enables us to estimate the divergence functions efficiently using reported samples, via estimating the variational form of the f-divergence functions, through a deep neutral network. These estimation results help us establish an approximate incentive compatibility in eliciting truthful samples. It is worth to note our framework also generalizes to the setting where there is no access to ground truth sample, where we can only rely on reported samples. There we show 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.

Our contributions are three-folds. (1) We tackle the problem of eliciting complex distribution via proposing a sample elicitation framework. Our deep learning aided solution concept makes it practical to solicit complex sample information from human agents. (2) Our framework covers the case when the mechanism designer has no access to ground truth information, which adds contribution to the peer prediction literature. (3) On the technical side, we develop estimators via deep learning techniques with strong theoretical guarantees. This not only helps us establish approximate incentive-compatibility, but also enables the designer to recover the targeted distribution from elicited samples. Our contribution can therefore be summarized as

"eliciting credible training samples by deep learning, for deep learning".

1.1 Related works

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 & Winkler, 1976; Jose et al., 2006; Gneiting & Raftery, 2007). Characterization results for strictly proper scoring rules are given in (McCarthy, 1956; Savage, 1971; Gneiting & Raftery, 2007). Property elicitation noticed the challenge of eliciting complex distributions (Lambert et al., 2008; Steinwart et al., 2014; Frongillo & Kash, 2015b). For instance, (Abernethy & Frongillo, 2012) characterizes the scoring functions for eliciting linear properties. (Frongillo & Kash, 2015a) studies the complexity of eliciting properties. Another line of relevant research is peer prediction, which 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 & Schoenebeck, 2018; 2019). Our work complements the information elicitation literature via proposing and studying the question of sample elicitation via a variational approach to estimate f-divergence functions.

As mentioned, our work borrows ideas from the statistical learning literature on divergence estimation. The simplest way to estimate divergence starts with the estimation of density functions, see (Wang et al., 2009; Lee & Park, 2006; Wang et al., 2005) and the references therein. In recent years, another method based on the Donsker-Varadhan representation (Donsker & Varadhan, 1975) of the divergence function comes into play. Related works include (Ruderman et al., 2012; Nguyen et al., 2010; Kanamori et al., 2011; Sugiyama et al., 2012; Broniatowski & Keziou, 2004; 2009), where the estimation of divergence is modeled as the estimation of density ratio between two distributions. The

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

Donsker-Varadhan representation of the divergence function also motivates the well-know Generative Adversarial Network (GAN) (Goodfellow et al., 2014), which learns the distribution by minimizing the Kullback-Leibler divergence (Kullback & Leibler, 1951). Follow-up works involve f-GAN (Nowozin et al., 2016), Wasserstein-GAN (Arjovsky et al., 2017; Gulrajani et al., 2017) and Cramér-GAN (Bellemare et al., 2017), where different divergence functions are used to learn the distribution. Theoretical analysis of GAN are given in (Liang, 2018; Liu et al., 2017; Arora et al., 2017).

1.2 NOTATIONS

For the probability measure $\mathbb{P}$, we denote by $\mathbb{P}_n$ the empirical measure given a set of samples $\{x_i\}_{i=1}^n$ following $\mathbb{P}$; in other words, $\mathbb{P}_n = 1/n \cdot \sum_{i=1}^n \delta_{x_i}$, where δ_{x_i} is the Dirac measure at x_i. We denote by $\|v\|_s$ the ℓ_s norm of vector $v \in \mathbb{R}^d$ where $1 \leq s < \infty$, and $\|v\|_\infty = \max_{1 \leq i \leq d} |v^{(i)}|$, where $v^{(i)}$ is the i-th entry 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 $f(x)$ and $\|f\|_s := [\int_{\mathcal{X}} |f(x)|^s \, \mathrm{d}m]^{1/s}$ the $L_s(m)$ norm of $f(x)$, where m is the Lebesgue measure. Also, we denote by $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. For any real-valued functions g and h defined on some unbounded subset of the real positive numbers, such that $h(\alpha)$ is strictly positive for all large enough values of α, 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 α_0 is a real number. We denote by $[n]$ the set $\{1, 2, \ldots, n\}$.

2 PRELIMINARY

We formulate the question of sample elicitation.

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. This is a setting that resembles similarity to the proper scoring rule setting. 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 the agent holds n samples, with each of them drawn from $\mathbb{P}$, i.e., $x_i \sim \mathbb{P}$ [2]. The agent can report each sample arbitrarily $r_i(x_i) : \Omega \to \Omega$. There are n data $x_1^*, ..., x_n^*$ drawn from the ground truth distribution $\mathbb{Q}$[3]. We are interested in designing score functions $S(\cdot)$ that takes inputs of each $r_i(\cdot)$ and $\{r_j(x_j), x_j^*\}_{j=1}^n$: $S\big(r_i(x_i), \{r_j(x_j), x_j^*\}_{j=1}^n\big)$ such that if the agent believes that x^* is drawn from the same distribution $x^* \sim \mathbb{P}$, with probability at least $1 - \delta$, $\forall \, \{r_j(\cdot)\}_{j=1}^n$,

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

We name above as (δ, ϵ)-**properness** (per sample) for sample elicitation. When $\delta = \epsilon = 0$, it is reduced to a one that is similar to the properness definition in scoring rule literature. When there is no confusion we will also shorthand $r_i := r_i(x_i)$. Agent believes that his samples are generated from the same distribution as of the ground truth samples ($p = q$).

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

Again each agent can report arbitrarily $r_i(x_i) : \Omega \to \Omega$. We are interested in designing and characterizing score function $S(\cdot)$ that takes inputs of each $r_i(\cdot)$ and $\{r_j(x_j)\}_{j \neq i}$: $S\big(r_i(x_i), \{r_j(x_j)\}_{j \neq i}\big)$

[2] 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.

[3] 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 δ, ϵ terms resulting from our estimations.

such that $\forall\ \{r_j(\cdot)\}_{j=1}^n$, with probability at least $1-\delta$,

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

We name above an (δ,ϵ)**-Bayesian Nash Equilibrium** (BNE) in truthful elicitation. We only require that agents are all aware of 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 themselves.

2.2 f-DIVERGENCE

It is well known that maximizing the expected proper scores equals to minimizing a corresponding Bregman divergence (Gneiting & Raftery, 2007). More generically, we take the perspective that divergence functions have great potentials to serve as scoring functions for eliciting samples. Denote the f-divergence between two distributions p and q as $D_f(q\|p)$:

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

Here f is a function satisfying certain regularity conditions, which will be specified in the following section. Solving our elicitation problem involves evaluating the value of $D_f(q\|p)$ 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 exact form of p and q, but instead a sample form. Following from Fenchel's convex duality, it holds that

$$D_f(q\|p) = \max_t \mathbb{E}_{y\sim\mathbb{Q}}[t(y)] - \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))], \tag{2.1}$$

where $\mathbb{P}$ and $\mathbb{Q}$ correspond to the distributions with probability density functions p and q, and $f^\dagger$ is the Fenchel duality of f, 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 GENERATIVE ADVERSARIAL APPROACH

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

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

We highlight that via functional derivative, the above variational form 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. Our elicitation builds upon above variational form (3.1) and the following estimators:

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

3.1 CONCENTRATION AND ASSUMPTIONS

Suppose we have the following concentration bound for estimating $D_f(q\|p)$: with probability at least $1-\delta(n)$

$$|\widehat{D}_f(q\|p) - D_f(q\|p)| \leq \epsilon(n),\ \forall p,q. \tag{3.2}$$

This concentration bound will be established based on the following assumptions.

Assumption 3.1 (Bounded Density Ratio)**.** We assume that the density ratio $\theta^*(x;p,q) = q(x)/p(x)$ is bounded from above and below such that $0<\theta_0\leq\theta^*\leq\theta_1$ holds for some constants θ_0 and θ_1.

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

Definition 3.2 (β-Hölder Function)**.** The ball of β-Hölder functions with radius M is defined as

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

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

We assume that the function $t^*(\cdot; p, q)$ is β-Hölder, which characterizes the smoothness of $t^*(\cdot; p, q)$.

Assumption 3.3 (β-Hölder Condition)**.** We assume that $t^*(\cdot; p, q) \in \mathcal{C}_d^\beta(\Omega, M)$ for some positive constants $M > 0$, where $\mathcal{C}_d^\beta(\Omega, M)$ is the β-Hölder function class in Definition 3.2.

In addition, we assume that the following regularity conditions hold for the function f.

Assumption 3.4 (Regularity of Divergence Function)**.** The function f is smooth on $[\theta_0, \theta_1]$ and $f(1) = 0$. Furthermore,

(i) f is μ_0-strongly convex on $[\theta_0, \theta_1]$, where $\mu_0 > 0$ is a constant;
(ii) f has L_0-Lipschitz continuous gradient on $[\theta_0, \theta_1]$, where $L_0 > 0$ is a constant.

We highlight that we only require that the conditions hold on the interval $[\theta_0, \theta_1]$ in Assumption 3.4, where the constants θ_0 and θ_1 are specified in Assumption 3.1. Thus, the above assumptions are mild and they hold for many commonly used functions. For example, in Kullback-Leibler (KL) divergence, we take $f(u) = -\log u$, which satisfies Assumption 3.4; while in Jenson-Shannon divergence, we take $f(u) = u \log u - (u + 1) \log(u + 1)$, which also satisfies Assumption 3.4.

We will show that under Assumptions 3.1, 3.3, and 3.4, the bound (3.2) holds. See Theorem 4.2 in Section 4 for details.

3.2 Multi-sample elicitation with ground truth samples

In this setting, as a reminder, the agent will report multiple samples. After the agent reported his samples, the mechanism designer has a set of ground truth samples $x_1^*, ..., x_n^* \sim \mathbb{Q}$ to serve the purpose of evaluation. This falls into the standard strictly proper scoring rule setting.

Our mechanism is presented in Algorithm 1. It consists two steps: one is to compute $\widehat{t}(x)$, which will enable us, in step 2, to pay agent using a linear-transformed estimated divergence between the reported samples and the true samples.

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

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

And we have the following results.

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 concentration of f-divergence function and its non-negativity. Not surprisingly, if the agent believes his samples are generated from the same distribution as the ground truth sample, and that our estimator can well characterize the difference between the two set of samples, he will be incentivized to report truthfully to minimize the difference. We defer the proof to Section B.1.

3.3 Single-task elicitation without ground truth samples

The above mechanism, while intuitive, has two caveats:

- The agent needs to report multiple samples (multi-task/sample elicitation);
- Multiple samples from the ground truth distribution are needed.

Now consider the single point elicitation in an elicitation without verification setting. Suppose there are $2n$ agents each holding a sample $x_i \sim \mathbb{P}_i$ [4]. We randomly partition the agents into two groups, and denote the joint distributions for each group's samples as p and q with measures $\mathbb{P}$ and $\mathbb{Q}$ 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 is holding $x_1^p, ..., x_n^p \sim \mathbb{P}$ and $x_1^q, ..., x_n^q \sim \mathbb{Q}$. As an interesting note, this is also similar to the setup of a Generative Adversarial Network (GAN) - one distribution corresponds to a generative distribution $x|y = 1$, and another $x|y = 0$. This is a connection that we will further explore in Section 5 to recover distributions from elicited samples.

Denote the joint distribution of p and q as $p \oplus q$ (measure as $\mathbb{P} \oplus \mathbb{Q}$), and the product of the marginal distribution as $p \times q$ (measure as $\mathbb{P} \times \mathbb{Q}$). Consider the divergence between the two distributions:

$$D_f(p \oplus q || p \times q) = \max_t \mathbb{E}_{\mathbf{x} \sim \mathbb{P} \oplus \mathbb{Q}}[t(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P} \times \mathbb{Q}}[f^\dagger(t(\mathbf{x}))] \tag{3.3}$$

The results below connect mutual information with divergence functions. The most famous one is the relationship between KL divergence and mutual information, but the generic connection between a generalized f-mutual information definition and f-divergence function holds too.

Definition 3.6 (Kong & Schoenebeck (2019))**.** A generalized f-mutual information between p and q is defined as the f-divergence between the joint distribution of $p \oplus q$ and the product of marginal distribution $p \times q$:

$$I_f(p; q) = D_f(p \oplus q || p \times q)$$

Further it is shown in Kong & Schoenebeck (2018; 2019) that the data processing inequality for mutual information holds for $I_f(p; q)$ when f is strictly convex. Again define the following estimators: (we use $\mathbf{x}$ to denote a sample drawn from the joint distribution)

$$\widehat{t}(\mathbf{x}; p \oplus q, p \times q) = \operatorname*{argmin}_t \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \times \mathbb{Q}_n}[f^\dagger(t(x))] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n}[t(\mathbf{x})] \tag{3.4}$$

$$\widehat{D}_f(p \oplus q || p \times q) = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n}[\widehat{t}(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \times \mathbb{Q}_n}[f^\dagger(\widehat{t}(\mathbf{x}))]$$

Recall that $\mathbb{P}_n$ and $\mathbb{Q}_n$ are the empirical distributions of reported samples. We denote $\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n | r_i$ 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 is to estimate divergence between $\mathbb{P}_n \times \mathbb{Q}_n$ and $\mathbb{P}_n \oplus \mathbb{Q}_n$ using the reported samples. Then we pay agents using a linear-transformed form of it.

Algorithm 2 f-scoring mechanism for sample elicitation

1. Compute $\widehat{t}(\mathbf{x}; p \oplus q, p \times q) = \operatorname{argmin}_t \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \times \mathbb{Q}_n}[f^\dagger(t(\mathbf{x}))] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n}[t(\mathbf{x})]$.
2. Pay each reported sample r_i using:

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

for some constants $a, b > 0$.

And we have the following results.

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

The theorem is proved by our concentration results in estimating f-divergence, a max argument, and the data processing inequality for f-mutual information. We defer the proof in Section B.2.

The job left for us is to estimate the divergence functions as accurate as possible to reduce ϵ and δ. Roughly speaking, if we solve the optimization problem (3.4) using deep neural networks with proper structure, it holds that $\delta(n) = \exp\{-n^{(d-2\beta)/(2\beta+d)} \log^{14} n\}$ and $\epsilon(n) = c \cdot n^{-2\beta/(2\beta+d)} \log^7 n$, where c is a positive absolute constant. We state and prove this result formally in Section 4.

Several remarks follow:

[4] This choice of $2n$ is simply for exposition.

Remark 3.8. (1) When the number of samples grows, $\delta(n) \to 0, \epsilon(n) \to 0$ at least polynomially fast, and our guaranteed approximate incentive-compatibility approaches a strict one. (2) Our method or framework handles arbitrary complex information - x can be sampled from high dimensional continuous space. (3) The score function requires no prior knowledge - we design estimation methods purely based on reported sample data. (4) Our framework also covers the case when the mechanism designer has no access to 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 structure, our approach can also be viewed as a data-driven mechanism for peer prediction problems too.

4 ESTIMATION OF f-DIVERGENCE

In this section, we introduce an estimator of f-divergence and establish the statistical rate of convergence. In this way, we provide estimates of $\epsilon(n)$ and $\delta(n)$. For the ease of exposition, in the sequel, we consider estimating f-divergence $D_f(q\|p)$ between two given distribution $\mathbb{P}$ and $\mathbb{Q}$ with probability density functions $p(x)$ and $q(y)$, respectively. The results in this section can be easily extended to mutual information.

Recall that from Section 2.2, estimating f-divergence between $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to solving the following optimization problem:

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

We now proceed to propose an estimator of $D_f(q\|p)$. Following from Assumption 3.3, we need to solve the above optimization problem (4.1) on the function class $\mathcal{C}_d^\beta(\Omega, M)$, which is usually intractable. However, due to the strong ability to approximate any proper functions, deep neural networks help us solve the problem. We formally define the family of neural networks as follow.

Definition 4.1. Given a vector $k = (k_0, \ldots, k_{L+1}) \in \mathbb{N}^{L+2}$, where $k_0 = d$ and $k_{L+1} = 1$, the family of neural network is defined as

$$\Phi(L, k) = \{\varphi(x; W, v) = W_{L+1}\sigma_{v_L} \cdots W_2\sigma_{v_1}W_1x \colon W_j \in \mathbb{R}^{k_j\times k_{j-1}}, v_j \in \mathbb{R}^{k_j}\}.$$

Here $\sigma_v(x)$ is short for $\sigma(x - v)$, and $\sigma(\cdot)$ is the ReLU activation.

In deep learning literature, to avoid overfitting, related works usually assume the sparsity of the neural network; in other words, most parameters in the network vanish. In practice, this can be done through special techniques (e.g., dropout (Srivastava et al., 2014)) or special network architecture (e.g., convolutional neural network (Krizhevsky et al., 2012)). Motivated by this, we propose the following family of sparse networks:

$$\begin{aligned}\Phi_M(L, k, s) =&\{\varphi(x; W, v) \in \Phi(L, d) \colon \|\varphi\|_\infty \le M,\ \|W_j\|_\infty \le 1 \text{ for } j \in [L+1],\\ &\|v_j\|_\infty \le 1 \text{ for } j \in [L],\ \sum_{j=1}^{L+1} \|W_j\|_0 + \sum_{j=1}^{L} \|v_j\|_0 \le s\},\end{aligned} \tag{4.2}$$

where the number of nonzero parameters is limited to be at most s. Another approach by using the norm of parameters to characterize the networks is proposed in the appendix. See Section A.2 for details. We focus on sparse networks here. Consider the following estimators:

$$\widehat{t}(x; p, q) = \operatorname*{argmin}_{t\in\Phi_M(L,k,s)} \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{y\sim\mathbb{Q}_n}[t(y)],$$
$$\widehat{D}_f(q\|p) = \mathbb{E}_{y\sim\mathbb{Q}_n}[\widehat{t}(x; p, q)] - \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(\widehat{t}(x; p, q))]. \tag{4.3}$$

The following theorem characterizes the statistical rate of convergence.

Theorem 4.2. We consider the family of neural networks $\Phi_M(L, k, s)$. Here the parameters L, k, and s satisfy that $L = \mathcal{O}(\log n)$, $s = \mathcal{O}(N \log n)$, and $k = (d, d, \mathcal{O}(dN), \mathcal{O}(dN), \ldots, \mathcal{O}(dN), 1)$, where $N = n^{d/(2\beta+d)}$. Then under Assumptions 3.1, 3.3, and 3.4, it holds that $\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} \lesssim n^{-\beta/(2\beta+d)} \log^{7/2} n$ with probability at least $1 - \exp\{-n^{d/(2\beta+d)} \log^5 n\}$. In addition, it holds that

$$|D_f(q\|p) - \widehat{D}_f(q\|p)| \lesssim n^{-\frac{2\beta}{2\beta+d}} \log^7 n$$

with probability at least $1 - \exp\{-n^{(d-2\beta)/(2\beta+d)} \log^{14} n\}$.

The result in this theorem indeed achieves the optimal nonparametric rate of convergence (Stone, 1982) up to a logarithmic term. We defer the proof of the theorem in Section B.3. From (3.2) and Theorem 4.2, we have

$$\delta(n) = \exp\{-n^{(d-2\beta)/(2\beta+d)} \cdot \log^{14} n\}, \qquad \epsilon(n) = c \cdot n^{-2\beta/(2\beta+d)} \cdot \log^{7} n,$$

where c is a positive absolute constant.

5 Connection to GAN and reconstruction of distribution

After sample elicitation, a natural question to ask is how to learn a representative distribution from the samples. Denote the distribution from elicited samples as p. Then, learning the density 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. Combining (2.1) and (5.1), we obtain the standard formulation of f-GAN (Nowozin et al., 2016):

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \max_{t} \mathbb{E}_{x\sim\mathbb{Q}}[t(x)] - \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))]. \tag{5.2}$$

Here the function $t(\cdot)$ acts as a discriminator of f-GAN. By the non-negativity of f-divergence, $q^* = p$ solve the above optimization problem (5.1). Based on the convergence result of the estimated f-divergence $\widehat{D}_f(q\|p)$ in Section 4, a natural estimator $\widehat{q}$ of q^* solves the following problem

$$\widehat{q} = \operatorname*{argmin}_{q \in \mathcal{Q}} \widehat{D}_f(q\|p), \tag{5.3}$$

where $\widehat{D}_f(q\|p)$ is given in (4.3).

Before we state the convergence result, we introduce the following notation of covering number.

Definition 5.1 (Covering Number)**.** Let $(V, \|\cdot\|_{L_2})$ be a normed space, and $\mathcal{Q} \subset V$. We say that $\{v_1, \ldots, v_N\}$ is an δ-covering over $\mathcal{Q}$ of size N if $\mathcal{Q} \subset \cup_{i=1}^N B(v_i, \delta)$, where $B(v_i, \delta)$ is the δ-ball centered at v_i. The covering number is defined as $N_2(\delta, \mathcal{Q}) = \min\{N \colon \exists \epsilon\text{-covering over } \mathcal{Q} \text{ of size } N\}$.

We make the following assumption on the covering number of the function space $\mathcal{Q}$.

Assumption 5.2. The covering number $N_2(\delta, \mathcal{Q})$ satisfies that $N_2(\delta, \mathcal{Q}) = \mathcal{O}(\exp\{1/\delta^{d/(2\beta)-1}\})$.

Recall that $q^* = p$ is the unique minimizer of the problem (5.1); therefore, the divergence $D_f(\widehat{q}\|p)$ characterizes convergence of $\widehat{q}$ towards p^*. Under Assumption 5.2, we use results in Section 4 to establish the rate of convergence in the following theorem.

Theorem 5.3. We set the parametesr L, k and s of $\Phi_M(L, k, s)$ as chosen in Theorem 4.2. Then under Assumptions 3.1, 3.3, 3.4, and 5.2, for sufficiently large sample size n, we have

$$D_f(\widehat{q}\|p) \lesssim n^{-\frac{2\beta}{2\beta+d}} \cdot \log^7 n + \inf_{\widetilde{q} \in \mathcal{Q}} D_f(\widetilde{q}\|p)$$

with probability at least $1 - 1/n$.

The rate of convergence in Theorem 5.3 achieves the optimal nonparametric rate of convergence (Stone, 1982) up to a logarithmic term, and shows that the f-divergence between the reconstructed distribution and the true distribution is small. This characterizes the convergence of the distribution estimator in (5.3). We defer the proof of the theorem in Section B.4.

6 Concluding remarks

In this work, we introduce the problem of sample elicitation as an alternative to eliciting complicated distribution. Our elicitation mechanism leverages the variational form of f-divergence functions to achieve accurate estimation of the divergences using samples. We provide theoretical guarantee for both our estimators and the achieved incentive compatibility.

It reminds 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.

REFERENCES

Jacob D Abernethy and Rafael M Frongillo. A characterization of scoring rules for linear properties. In *Conference on Learning Theory*, pp. 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 *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 224–232. JMLR. org, 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.

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 *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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*, pp. 3258–3266, 2015a.

Rafael Frongillo and Ian A Kash. Vector-valued property elicitation. In *Conference on Learning Theory*, pp. 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.

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*, pp. 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*, pp. 5767–5777, 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 *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 177–194. ACM, 2018.

Yuqing Kong and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)*, 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*, pp. 251–264. Springer, 2016.

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

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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

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*, pp. 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. *PNAS: 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*, pp. 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.

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*, pp. 482–526, 2014.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 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*, pp. 5–20, 2008.

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

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.

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

A AUXILIARY ANALYSIS

A.1 AUXILIARY RESULTS ON SPARSITY CONTROL

As we mentioned in Section 4, the proof of Theorem 4.2 relies on auxiliary results, which we show in this section. We first state our main result, which acts as an oracle-type inequality on the rate of convergence of $\widehat{t}(x;p,q)$ towards $t^*(x;p,q) = f'(q(x)/p(x))$.

Theorem A.1. Given $0 < \varepsilon < 1$, for any sample size n satisfies that $n \gtrsim [\gamma + \gamma^{-1}\log(1/\varepsilon)]^2$, the error bound of the estimated $\widehat{t} \in \Phi_M(L,k,s)$ defined in (4.3) towards the ground truth $t^* = f'(q/p)$ satisfies

$$\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} \lesssim \min_{\widetilde{t}\in\Phi_M(L,k,s)} \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2}\log n + n^{-1/2}\Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1}\log(1/\varepsilon)\Big]$$

with probability at least $1 - \varepsilon \cdot \exp(-\gamma^2)$. Here γ takes the form $\gamma = s^{1/2}\log(V^2 L)$, where $V = \prod_{j=0}^{L+1}(k_j + 1)$.

We defer the proof of the theorem to Section B.5.

As a by-product, note that $t^* = f'(\theta^*) = f'(q/p)$, based on the error bound established in Theorem A.1, we immediately obtain the following results on the error bound of the density ratio $\widehat{\theta} = (f')^{-1}(\widehat{t})$ towards the true ratio $\theta^* = q/p$.

Corollary A.2. Given $0 < \varepsilon < 1$, for the sample size $n \gtrsim [\gamma + \gamma^{-1}\log(1/\varepsilon)]^2$, it holds with probability at least $1 - \varepsilon \cdot \exp(-\gamma^2)$ that

$$\|\widehat{\theta} - \theta^*\|_{L_2(\mathbb{P})} \lesssim \min_{\widetilde{t}\in\Phi_M(L,k,s)} \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2}\log n + n^{-1/2}\Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1}\log(1/\varepsilon)\Big].$$

Here γ takes the form $\gamma = s^{1/2}\log(V^2 L)$, where $V = \prod_{j=0}^{L+1}(k_j + 1)$.

Proof. Note that $(f')^{-1} = (f^\dagger)'$ and $f^\dagger$ has Lipschitz continuous gradient with parameter $1/\mu_0$ from Assumption 3.4 and Lemma D.6, we obtain the result immediately from Theorem A.1. □

A.2 ERROR BOUND USING NORM CONTROL

In this section, we consider using norm of the parameters (specifically speaking, the norm of W_j and v_j in (4.1)) to control the error bound, which is an alternative of the network model shown in (4.2). We only consider the generalization error bound in this setting; therefore, we assume that the ground truth $t^* = f'(q/p)$ locates within the target constraint set $\mathcal{D}$. We consider the family of L-layer neural networks with bounded spectral norm for weight matrices $W = \{W_j \in \mathbb{R}^{k_j \times k_{j-1}}\}_{j=1}^{L+1}$, where $k_0 = d$ and $k_{L+1} = 1$, and vector $v = \{v_j \in \mathbb{R}^{k_j}\}_{j=1}^{L}$, which is denoted as

$$\Phi_{\text{norm}} = \Phi_{\text{norm}}(L,k,A,B) = \Big\{\varphi(x;W,v) \in \Phi(L,k) : \|v_j\|_2 \le A_j \text{ for all } j \in [L], \\ \|W_j\|_2 \le B_j \text{ for all } j \in [L+1]\Big\}, \tag{A.1}$$

where $\sigma_{v_j}(x)$ is short for $\sigma(x - v_j)$ for any $j \in [L]$ as before. Then by this specific choice of $\mathcal{D} = \Phi_{\text{norm}}$, we write the below program

$$\widehat{t}(x;p,q) = \operatorname*{argmin}_{t\in\Phi_{\text{norm}}} \mathbb{E}_{\mathbb{P}_n}\Big\{f^\dagger\big[t(x)\big]\Big\} - \mathbb{E}_{\mathbb{Q}_n}\big[t(x)\big], \\ \widehat{D}_f(q\|p) = \mathbb{E}_{\mathbb{Q}_n}\big[\widehat{t}(x;p,q)\big] - \mathbb{E}_{\mathbb{P}_n}\Big\{f^\dagger\big[\widehat{t}(x;p,q)\big]\Big\}. \tag{A.2}$$

Based on this formulation, we derive the error bound on the estimated f-divergence in the following theorem. Before we state the theorem, we first define two parameters for the family of neural networks $\Phi_{\text{norm}}(L,k,A,B)$ as follows

$$\gamma_1 = B\prod_{j=1}^{L+1} B_j \cdot \sqrt{\sum_{j=0}^{L+1} k_j^2}, \qquad \gamma_2 = \frac{L\cdot\Big(\sqrt{\sum_{j=1}^{L+1} k_j^2 B_j^2} + \sum_{j=1}^{L} A_j\Big)}{\sum_{j=0}^{L+1} k_j^2 \cdot \min_j B_j^2}. \tag{A.3}$$

We proceed to state the theorem.

Theorem A.3. We assume that the ground truth $t^* \in \Phi_{\text{norm}}$. Then for any $0 < \varepsilon < 1$, with probability at least $1 - \varepsilon$, we have the following error bound on $\widehat{D}_f(q\|p)$:

$$\big|\widehat{D}_f(q\|p) - D_f(q\|p)\big| \lesssim \gamma_1 \cdot n^{-1/2} \log(\gamma_2 n) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \sqrt{\log(1/\varepsilon)}.$$

Here γ_1 and γ_2 are defined in (A.3).

We defer the proof to Section B.6.

The next theorem uses the error bound of the estimated f-divergence in Theorem A.3. Recall that in Section §A.2, we assume that the minimizer t^* to the population version problem (4.1) lies within the norm-controlled family of neural networks $\Phi_{\text{norm}}(L, k, A, B)$.

Theorem A.4. Recall that we defined the parameter γ_1 and γ_2 of the family of neural networks $\Phi_{\text{norm}}(L, k, A, B)$ in (A.3), the estimated distribution $\widehat{q}$ in (5.3), and the ground truth $q^* = p$. We denote the the L_2 covering number of the probability distribution function class $\mathcal{Q}$ as $N_2(\delta, \mathcal{Q})$, then for any $0 < \varepsilon < 1$, with probability at least $1 - \varepsilon$, we have

$$D_f(\widehat{q}\|p) \lesssim b_2(n, \gamma_1, \gamma_2) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \cdot \sqrt{\log\Big\{N_2\big[b_2(n, \gamma_1, \gamma_2), \mathcal{Q}\big]/\varepsilon\Big\}} + \min_{\widetilde{q} \in \mathcal{Q}} D_f(\widetilde{q}\|p),$$

where $b_2(n, \gamma_1, \gamma_2) = \gamma_1 n^{-1/2} \log(\gamma_2 n)$.

We defer the proof to Section B.7.

B Proofs of Theorems

B.1 Proof of Theorem 3.5

If the player truthfully reports, he will receive the following expected payment per sample i: with probability at least $1 - \delta(n)$,

$$\begin{aligned}
\mathbb{E}[S(r_i, \cdot)] &:= a - b\big(\mathbb{E}_{x \sim \mathbb{Q}_n}[\widehat{t}(x)] - \mathbb{E}_{x_i \sim \mathbb{P}_n}[f^\dagger(\widehat{t}(x_i))]\big) \\
&= a - b \cdot \widehat{D}_f(q||p) \\
&\geq a - b \cdot (D_f(p||p) + \epsilon(n)) \ \text{ (agent believes } p = q) \\
&= a - b\epsilon(n)
\end{aligned}$$

Similarly, any misreporting according to a distribution $\widetilde{p}$ with measure $\widetilde{\mathbb{P}}$ will lead to the following derivation with probability at least $1 - \delta$

$$\begin{aligned}
\mathbb{E}[S(r_i, \cdot)] &:= a - b\big(\mathbb{E}_{x \sim \mathbb{Q}_n}[\widehat{t}(x)] - \mathbb{E}_{x_i \sim \widetilde{\mathbb{P}}_n}[f^\dagger(\widehat{t}(x_i))]\big) \\
&= a - b \cdot \widehat{D}_f(q||\widetilde{p}) \\
&\leq a - b \cdot (D_f(p||\widetilde{p}) - \epsilon(n)) + \delta(n) \cdot \bar{S} \\
&\leq a + b\epsilon(n) \ \text{ (non-negativity of } D_f)
\end{aligned}$$

Combining above, and using union bound, leads to $(2\delta(n), 2b\epsilon(n))$-properness.

B.2 Proof of Theorem 3.7

Consider an arbitrary agent i. Suppose every other agent truthfully reports.

$$\begin{aligned}
\mathbb{E}\big[S(r_i, \{r_j\}_{j \neq i})\big] &= a + b\big(\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n | r_i}[\widehat{t}(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \times \mathbb{Q}_n | r_i}\{f^\dagger(\widehat{t}(\mathbf{x}))\}\big) \\
&= a + b\mathbb{E}\big[\mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \oplus \mathbb{Q}_n | r_i}[\widehat{t}(x)] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_n \times \mathbb{Q}_n | r_i}\{f^\dagger(\widehat{t}(\mathbf{x}))\}\big]
\end{aligned}$$

Consider the divergence term $\mathbb{E}\big[\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n|r_i}[\widehat{t}(x)] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n|r_i}\{f^\dagger(\widehat{t}(\mathbf{x}))\}\big]$. Reporting a $r_i \sim \widetilde{\mathbb{P}} \neq \mathbb{P}$ (denoting its distribution as $\widetilde{p}$) leads to the following score

$$
\begin{aligned}
&\mathbb{E}_{r_i\sim\widetilde{\mathbb{P}}_n}[\mathbb{E}_{\mathbf{x}\sim\widetilde{\mathbb{P}}_n\oplus\mathbb{Q}_n|r_i}[\widehat{t}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\widetilde{\mathbb{P}}_n\times\mathbb{Q}_n|r_i}\{f^\dagger(\widehat{t}(\mathbf{x}))\}] \\
&= \mathbb{E}_{\mathbf{x}\sim\widetilde{\mathbb{P}}_n\oplus\mathbb{Q}_n}[\widehat{t}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\widetilde{\mathbb{P}}_n\times\mathbb{Q}_n}\{f^\dagger(\widehat{t}(\mathbf{x}))\} \text{ (tower property)} \\
&\leq \max_t \mathbb{E}_{\mathbf{x}\sim\widetilde{\mathbb{P}}_n\oplus\mathbb{Q}_n}[t(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\widetilde{\mathbb{P}}_n\times\mathbb{Q}_n}\{f^\dagger(t(\mathbf{x}))\} \text{ (max)} \\
&= \widehat{D}_f(\widetilde{p}\oplus q||\widetilde{p}\times q) \\
&\leq D_f(\widetilde{p}\oplus q||\widetilde{p}\times q) + \epsilon(n) \\
&= I_f(\widetilde{p};q) + \epsilon(n) \text{ (definition)} \\
&\leq I_f(p;q) + \epsilon(n) \text{ (data processing inequality (Kong \& Schoenebeck, 2019))}
\end{aligned}
$$

with probability at least $1-\delta(n)$ (the other $\delta(n)$ probability with maximum score $\bar{S}$).

Now we prove that truthful reporting leads at least

$$
I_f(p;q) - \epsilon(n)
$$

of the divergence term:

$$
\begin{aligned}
&\mathbb{E}_{x_i\sim\mathbb{P}_n}[\mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n|x_i}[\widehat{t}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n|x_i}\{f^\dagger(\widehat{t}(\mathbf{x}))\}] \\
&= \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\oplus\mathbb{Q}_n}[\widehat{t}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_n\times\mathbb{Q}_n}\{f^\dagger(\widehat{t}(\mathbf{x}))\} \text{ (tower property)} \\
&= \widehat{D}_f(p\oplus q||p\times q) \\
&\geq D_f(p\oplus q||p\times q) - \epsilon(n) \\
&= I_f(p;q) - \epsilon(n) \text{ (definition)}
\end{aligned}
$$

with probability at least $1-\delta(n)$ (the other $\delta(n)$ probability with score at least 0). Therefore the expected divergence terms differ at most by $2\epsilon(n)$ with probability at least $1-2\delta(n)$ (via union bound). The above combines to establish a $(2\delta(n), 2b\epsilon(n))$-BNE.

B.3 Proof of Theorem 4.2

Part 1. We proceed to prove the bound on $\|t^* - \widehat{t}\|_{L_2(\mathbb{P})}$. We first proceed to find some $\widetilde{t} \in \Phi_M(L,k,s)$. Note that the ground truth t^* is assumed to be on a finite support $D \subset [a,b]^d$. For simplicity, the notation that a vector plus or minus a scaler actually means that the scaler is operated to each coordinate of the vector. For example, for vector $x \in \mathbb{R}^d$ and a scaler $c \in \mathbb{R}$, we use $x - c$ to represent $(x_1 - c, \ldots, x_d - c)^\top$, where x_i is the i-th coordinate of the vector x. In order to invoke Theorem D.5, we denote $t'(y) = t^*[(b-a)y + a]$, then the support of t' actually lies in the unit cube $[0,1]^d$. We choose $L' = \mathcal{O}(\log n), s' = \mathcal{O}(N\log n), k' = (d, \mathcal{O}(dN), \mathcal{O}(dN), \ldots, \mathcal{O}(dN), 1)$, and $m' = \log n$, we then utilize Theorem D.5 to construct some $\widetilde{t}' \in \Phi_M(L', k', s')$ such that

$$
\|\widetilde{t}' - t'\|_{L^\infty([0,1]^d)} \lesssim N^{-\beta/d}.
$$

We further define $\widetilde{t}(x) = \widetilde{t}'[(x-a)/(b-a)]$, and claim that this function can also be written as a ReLU neural network in $\Phi_M(L,k,s)$. To see this, note that we only need to add one more layer in front of $\widetilde{t}'$, and this additional layer gives the linear transformation

$$
x \mapsto \frac{x}{b-a} - \frac{a}{b-a},
$$

where $x \in D \subset [a,b]^d$ is the input of the neural network $\widetilde{t}$. Then indeed the proposed function $\widetilde{t}$ can be written as a ReLU network in $\Phi_M(L,k,s)$. We fix this $\widetilde{t}$ and use Theorem A.1, then with probability at least $1 - \varepsilon\cdot\exp(-\gamma^2)$, we have

$$
\begin{aligned}
\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} &\lesssim \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2}\log n + n^{-1/2}\Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1}\log(1/\varepsilon)\Big] \\
&\lesssim N^{-\beta/d} + \gamma n^{-1/2}\log n + n^{-1/2}\Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1}\log(1/\varepsilon)\Big]. \qquad \text{(B.1)}
\end{aligned}
$$

Note that γ takes the form $\gamma = s^{1/2}\log(V^2 L)$, where $V = \mathcal{O}(d^L \cdot N^L)$ and L, s given above, we write $\gamma = \mathcal{O}(N^{1/2}\log^{5/2} n)$. Moreover, by the choice $N = n^{d/(2\beta+d)}$, combining (B.1) and taking $\varepsilon = 1/n$, we then conclude the first part of the theorem.

Part 2. For the second part, we denote by $\mathcal{L}(t) = \mathbb{E}_{x\sim\mathbb{Q}}[t(x)] - \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))]$ and $\widehat{\mathcal{L}}(t) = \mathbb{E}_{x\sim\mathbb{Q}_n}[t(x)] - \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))]$. Then from Assumption 3.4 and Lemma D.6, we know that $\widehat{\mathcal{L}}(\cdot)$ is strongly convex with a constant coefficient. Note that by triangular inequality, we have

$$|\widehat{D}_f(q\|p) - D_f(q\|p)| = |\widehat{\mathcal{L}}(\widehat{t}) - \mathcal{L}(t^*)| \le |\widehat{\mathcal{L}}(t^*) - \widehat{\mathcal{L}}(\widehat{t})| + |\widehat{\mathcal{L}}(t^*) - \mathcal{L}(t^*)| =: A_1 + A_2.$$

We proceed to bound A_1 and A_2.

Bound on A_1: Recall that $\widehat{\mathcal{L}}(\cdot)$ is strongly convex. Consequently, we have

$$A_1 \lesssim \|t^* - \widehat{t}\|^2_{L_2(\mathbb{P})} \lesssim n^{-\frac{\beta}{2\beta+d}}\log^{7/2} n,$$

with probability at least $1 - \exp(-n^{d/(2\beta+d)}\log^5 n)$, where the last inequality comes from **Part 1**.

Bound on A_2: Note that both $t^*(\cdot)$ and $f^\dagger(t^*(\cdot))$ are bounded, then by Hoeffding's inequality, we obtain that

$$\mathbb{P}\{A_2 \le n^{-\frac{\beta}{2\beta+d}}\log^{7/2} n\} \ge 1 - \exp(-n^{(d-2\beta)/(2\beta+d)}\log^{14} n).$$

Therefore, by combining the above two bounds, we obtain that

$$|\widehat{D}_f(q\|p) - D_f(q\|p)| \lesssim n^{-\frac{\beta}{2\beta+d}}\log^{7/2} n$$

with probability at least $1 - \exp(-n^{(d-2\beta)/(2\beta+d)}\log^{14} n)$. This concludes the whole proof.

B.4 Proof of Theorem 5.3

We first need to bound the max deviation of the estimated f-divergence $\widehat{D}_f(q\|p)$ among all $q \in \mathcal{Q}$. The following lemma provides such a bound.

Lemma B.1. Assume that the density function $q \in \mathcal{Q}$. Then for any fixed density p, if the sample size n is sufficiently large, it holds that

$$\sup_{q\in\mathcal{Q}}\big|D_f(q\|p) - \widehat{D}_f(q\|p)\big| \lesssim n^{-\frac{2\beta}{2\beta+d}}\cdot\log^7 n$$

with probability at least $1 - 1/n$.

Proof. See Section §C.1 for a detailed proof. □

Now we turn to the proof of the theorem. Note that for any $\widetilde{q} \in \mathcal{Q}$, with probability at least $1 - 1/n$, we have

$$\begin{aligned} D_f(\widehat{q}\|p) &\le \big|D_f(\widehat{q}\|p) - \widehat{D}_f(\widehat{q}\|p)\big| + \widehat{D}_f(\widehat{q}\|p) \\ &\le \sup_{q\in\mathcal{Q}}\big|D_f(q\|p) - \widehat{D}_f(q\|p)\big| + \widehat{D}_f(\widetilde{q}\|p) \lesssim n^{-\frac{2\beta}{2\beta+d}}\cdot\log^7 n + D_f(\widetilde{q}\|p). \end{aligned} \tag{B.2}$$

Here in the second line we use the optimality of $\widehat{q}$ among all $\widetilde{q} \in \mathcal{Q}$ to the problem (5.3), while the last inequality uses Lemma B.1. Moreover, by taking the infimum of $\widetilde{q} \in \mathcal{Q}$ on both sides of (B.2), we obtain that

$$D_f(\widehat{q}\|p) \lesssim n^{-\frac{2\beta}{2\beta+d}}\cdot\log^7 n + \inf_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p).$$

This concludes the error bound.

B.5 PROOF OF THEOREM A.1

We write $\mathbb{E}_{\mathbb{P}}[f] = \mathbb{E}_{x\sim\mathbb{P}}[f(x)]$ for notational convenience. First we establish the following lemma.

Lemma B.2. The following inequality holds:

$$1/(4L_0)\cdot\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}^2 \le 1/\mu_0\cdot\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\cdot\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})} + \Big\{\mathbb{E}_{\mathbb{Q}_n}\big[(\widehat{t}-\widetilde{t})/2\big] - \mathbb{E}_{\mathbb{Q}}\big[(\widehat{t}-\widetilde{t})/2\big]\Big\} - \bigg(\mathbb{E}_{\mathbb{P}_n}\Big\{f^\dagger\big[(\widehat{t}+\widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\} - \mathbb{E}_{\mathbb{P}}\Big\{f^\dagger\big[(\widehat{t}+\widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\}\bigg)$$

Here μ_0 and L_0 are specified in Assumption 3.4.

Proof. See Section §C.2 for a detailed proof. □

Note that by the above Lemma B.2 and the fact that $f^\dagger$ is Lipschitz continuous, we have

$$\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}^2 \le \|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\cdot\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})} + \Big\{\mathbb{E}_{\mathbb{Q}_n}\big[(\widehat{t}-\widetilde{t})/2\big] - \mathbb{E}_{\mathbb{Q}}\big[(\widehat{t}-\widetilde{t})/2\big]\Big\} - \bigg(\mathbb{E}_{\mathbb{P}_n}\Big\{f^\dagger\big[(\widehat{t}+\widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\} - \mathbb{E}_{\mathbb{P}}\Big\{f^\dagger\big[(\widehat{t}+\widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\}\bigg)$$

Furthermore, to bound the RHS of the above inequality, we establish the following lemma concerning about the error bound on empirical process.

Lemma B.3. We assume that the function $\psi:\mathbb{R}\to\mathbb{R}$ is Lipschitz continuous and bounded such that $|\psi(x)|\le M_0$ for any $|x|\le M$. Then for any fixed $\widetilde{t}(x)\in\Phi_M$, $n\gtrsim[\gamma+\gamma^{-1}\log(1/\varepsilon)]^2$ and $0<\varepsilon<1$, we have the follows

$$\mathbb{P}\Bigg\{\sup_{t(x)\in\Phi_M}\frac{\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t)-\psi(\widetilde{t})\big]-\mathbb{E}_{\mathbb{P}}\big[\psi(t)-\psi(\widetilde{t})\big]\Big|}{\eta(n,\gamma,\varepsilon)\cdot\big\|\psi(t)-\psi(\widetilde{t})\big\|_{L_2(\mathbb{P})}\bigvee\lambda(n,\gamma,\varepsilon)}\le 16M_0\Bigg\}\ge 1-\varepsilon\cdot\exp(-\gamma^2),$$

where $\eta(n,\gamma,\varepsilon)=n^{-1/2}[\gamma\log n+\gamma^{-1}\log(1/\varepsilon)]$ and $\lambda(n,\gamma,\varepsilon)=n^{-1}[\gamma^2+\log(1/\varepsilon)]$. Here γ takes the form $\gamma=s^{1/2}\log(V^2L)$, where $V=\prod_{j=0}^{L+1}(k_j+1)$.

Proof. See Section §C.3 for a detailed proof. □

Note that the empirical process in the above Lemma B.3 also applies to the probability measure $\mathbb{Q}$, and by using the fact that the true density ratio $\theta^*=q/p$ is bounded below and above, we further know that $L_2(\mathbb{Q})$ is equivalent to $L_2(\mathbb{P})$. Therefore, by Lemma B.3 and the Lipschitz property of $f^\dagger$ according to Lemma D.6, with probability at least $1-\varepsilon\cdot\exp(-\gamma^2)$, we have the following bound

$$\begin{aligned}\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}^2 &\lesssim \|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\cdot\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})}\\ &\quad+\mathcal{O}\Big\{n^{-1/2}\big[\gamma\log n+\gamma^{-1}\log(1/\varepsilon)\big]\cdot\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\bigvee n^{-1}[\gamma^2+\log(1/\varepsilon)]\Big\},\end{aligned}\tag{B.3}$$

where we recall that the notation $\gamma=s^{1/2}\log(V^2L)$ is a parameter related with the family of neural networks Φ_M. We proceed to analyze the dominant part on the RHS of (B.3).

Case 1. If the term $\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\cdot\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})}$ dominates, then with probability at least $1-\varepsilon\cdot\exp(-\gamma^2)$

$$\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\lesssim\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})}.$$

Case 2. If the term $\mathcal{O}\{n^{-1/2}\big[\gamma\log n+\gamma^{-1}\log(1/\varepsilon)\big]\cdot\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\}$ dominates, then with probability at least $1-\varepsilon\cdot\exp(-\gamma^2)$

$$\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\lesssim n^{-1/2}\big[\gamma\log n+\gamma^{-1}\log(1/\varepsilon)\big].$$

Case 3. If the term $\mathcal{O}\{n^{-1}[\gamma^2+\log(1/\varepsilon)]\}$ dominates, then with probability at least $1-\varepsilon\cdot\exp(-\gamma^2)$

$$\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\lesssim n^{-1/2}\Big[\gamma+\sqrt{\log(1/\varepsilon)}\Big].$$

Therefore, by combining the above three cases, we have

$$\|\widehat{t} - \widetilde{t}\|_{L_2(\mathbb{P})} \lesssim \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2} \log n + n^{-1/2} \Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1} \log(1/\varepsilon) \Big].$$

Further the triangular inequality gives us

$$\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} \lesssim \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2} \log n + n^{-1/2} \Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1} \log(1/\varepsilon) \Big]$$

with probability at least $1 - \varepsilon \cdot \exp(-\gamma^2)$. Note that the above error bound holds for any $\widetilde{t} \in \Phi_M(L, k, s)$, especially for the special choice $\widetilde{t}$ such that it minimize $\|\widetilde{t} - t^*\|_{L_2(\mathbb{P})}$. Therefore, we have

$$\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} \lesssim \min_{\widetilde{t} \in \Phi_M(L,k,s)} \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2} \log n + n^{-1/2} \Big[\sqrt{\log(1/\varepsilon)} + \gamma^{-1} \log(1/\varepsilon) \Big]$$

with probability at least $1 - \varepsilon \cdot \exp(-\gamma^2)$. This concludes the proof of the theorem.

B.6 Proof of Theorem A.3

We follow the proof in Li et al. (2018). We denote our loss function in (A.2) to be $\mathcal{L}[t(x)] = f^\dagger[t(x^{\mathrm{I}})] - t(x^{\mathrm{II}})$, where x^{I} follows the distribution $\mathbb{P}$, while x^{II} follows $\mathbb{Q}$. To prove the theorem, we first link the generalization error in our theorem to the empirical Rademacher complexity (ERC). Then given the data $\{x_i\}_{i=1}^n$, the ERC related with the class $\mathcal{L}(\Phi_{\text{norm}})$ is defined to be

$$\mathfrak{R}_n\big[\mathcal{L}(\Phi_{\text{norm}})\big] = \mathbb{E}_\varepsilon \Bigg\{ \sup_{\varphi \in \Phi_{\text{norm}}} \bigg| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot \mathcal{L}\big[\varphi(x_i; W, v)\big] \bigg| \big\{x_i\big\}_{i=1}^n \Bigg\}, \tag{B.4}$$

where ε_i's are i.i.d. Rademacher random variables, i.e., $\mathbb{P}[\varepsilon_i = 1] = \mathbb{P}[\varepsilon_i = -1] = 1/2$. Here the expectation in (B.4) is taken over the Rademacher random variables $\{\varepsilon_i\}_{i \in [n]}$.

To see the importance of the ERC, we introduce the following Lemma B.4 (Mohri et al., 2018), which links the ERC to the generalization error bound, and therefore we derive the desired error bound of estimated f-divergence $\widehat{D}_f(q\|p)$ towards the true divergence $D_f(q\|p)$ in our case.

Lemma B.4. Assume that $\sup_{\varphi \in \Phi_{\text{norm}}} |\mathcal{L}(\varphi)| \leq M_1$, then for any $\varepsilon > 0$, with probability at least $1 - \varepsilon$, we have

$$\sup_{\varphi \in \Phi_{\text{norm}}} \Big\{ \mathbb{E}_x \Big\{ \mathcal{L}\big[\varphi(x; W, v)\big] \Big\} - \frac{1}{n} \sum_{i=1}^n \mathcal{L}\big[\varphi(x_i; W, v)\big] \Big\} \lesssim \mathfrak{R}_n\big[\mathcal{L}(\Phi_{\text{norm}})\big] + M_1 \cdot n^{-1/2} \sqrt{\log(1/\varepsilon)},$$

where the expectation is taken over $x^{\mathrm{I}} \sim \mathbb{P}$ and $x^{\mathrm{II}} \sim \mathbb{Q}$.

Equipped with the above lemma, to derive the final error bound on the generalization, we need the following lemma to provide an error bound on the ERC proposed in (B.4).

Lemma B.5. Let $\mathcal{L}$ be a Lipschitz continuous loss function and Φ_{norm} be a family of L-layer networks defined in (A.1). Moreover, we assume that the input $x \in \mathbb{R}^d$ is bounded such that $\|x\|_2 \leq B$. Then the ERC defined in (B.4) satisfies that

$$\mathfrak{R}_n\big[\mathcal{L}(\Phi_{\text{norm}})\big] \lesssim \gamma_1 \cdot n^{-1/2} \log(\gamma_2 n),$$

where γ_1 and γ_2 are given in (A.3).

Proof. See §C.4 for a detailed proof. □

Now we proceed to prove the theorem. Recall that as we mentioned before, the ground truth $t^* \in \Phi_{\text{norm}}$. For notational convenience, we further denote that

$$\widehat{t} \in \operatorname*{argmin}_{t \in \Phi_{\text{norm}}} \widehat{H}(t) = \mathbb{E}_{\mathbb{P}_n} \Big\{ f^\dagger\big[t(x)\big] \Big\} - \mathbb{E}_{\mathbb{Q}_n}\big[t(x)\big],$$

and the corresponding population version

$$t^* \in \operatorname*{argmin}_t H(t) = \mathbb{E}_{\mathbb{P}}\Big\{f^\dagger\big[t(x)\big]\Big\} - \mathbb{E}_{\mathbb{Q}}\big[t(x)\big].$$

Then $\mathbb{E}[\widehat{H}(t)] = H(t)$. We proceed to bound $|\widehat{D}_f(q\|p) - D_f(q\|p)| = |\widehat{H}(\widehat{t}) - H(t^*)|$. Note that if $\widehat{H}(\widehat{t}) \geq H(t^*)$, then we have

$$0 \leq \widehat{H}(\widehat{t}) - H(t^*) \leq \widehat{H}(t^*) - H(t^*), \tag{B.5}$$

where the second inequality follows that fact that $\widehat{t}$ is the minimizer of $\widehat{H}$. On the other hand, if $\widehat{H}(\widehat{t}) \leq H(t^*)$, then we have

$$0 \geq \widehat{H}(\widehat{t}) - H(t^*) \geq \widehat{H}(\widehat{t}) - H(\widehat{t}), \tag{B.6}$$

where the second inequality follows that fact that t^* is the minimizer of H. Therefore, by considering both (B.5) and (B.6), and the fact that $\mathcal{L}(\varphi) \lesssim \prod_{j=1}^{L+1} B_j$ for any $\varphi \in \Phi_{\text{norm}}$, we deduce that

$$\big|\widehat{H}(\widehat{t}) - H(t^*)\big| \leq \sup_{t\in\Phi_{\text{norm}}} \big|\widehat{H}(t) - H(t)\big| \lesssim \mathfrak{R}_n\big[\mathcal{L}(\Phi_{\text{norm}})\big] + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2}\sqrt{\log(1/\varepsilon)} \tag{B.7}$$

with probability at least $1-\varepsilon$. Here the second inequality follows directly from Lemma B.4. We further plug the result from Lemma B.5 into (B.7), then we deduce that with probability at least $1-\varepsilon$, the bound of estimated f-divergence towards the truth divergence satisfies

$$\begin{aligned}
&\big|\widehat{H}(\widehat{t}) - H(t^*)\big| = \big|\widehat{D}_f(q\|p) - D_f(q\|p)\big| \\
&\lesssim \frac{B}{\sqrt{n}} \cdot \prod_{j=1}^{L+1} B_j \cdot \sqrt{\sum_{j=0}^{L+1} k_j^2 \cdot \log\left[\frac{\sqrt{nL}\cdot\Big(\sqrt{\sum_{j=1}^{L+1} k_j^2 B_j^2} + \sum_{j=1}^{L} A_j\Big)}{\sqrt{\sum_{j=0}^{L+1} k_j^2}\cdot \min_j B_j}\right]} + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2}\sqrt{\log(1/\varepsilon)}.
\end{aligned}$$

Recall the definition of γ_1 and γ_2 in (A.3), then we concludes the theorem.

B.7 Proof of Theorem A.4

We first need to bound the max deviation of the estimated f-divergence $\widehat{D}_f(q\|p)$ among all $q \in \mathcal{Q}$. We utilize the following lemma to provide such a bound.

Lemma B.6. Assume that the distribution q is in the set $\mathcal{Q}$, and we denote its L_2 covering number as $N_2(\delta, \mathcal{Q})$. Then for any target distribution p, we have

$$\max_{q\in\mathcal{Q}}\Big|D_f(q\|p) - \widehat{D}_f(q\|p)\Big| \lesssim b_2(n,\gamma_1,\gamma_2) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \cdot \sqrt{\log\Big\{N_2\big[b_2(n,\gamma_1,\gamma_2),\mathcal{Q}\big]/\varepsilon\Big\}}$$

with probability at least $1-\varepsilon$. Here $b_2(n,\gamma_1,\gamma_2) = \gamma_1 n^{-1/2}\log(\gamma_2 n)$ and c is a positive absolute constant.

Proof. See Section §C.5 for a detailed proof. □

Now we turn to the proof of the theorem. Note that for any $\widetilde{q} \in \mathcal{Q}$, with probability at least $1-\varepsilon$, we have

$$\begin{aligned}
D_f(\widehat{q}\|p) &\leq \big|D_f(\widehat{q}\|p) - \widehat{D}_f(\widehat{q}\|p)\big| + \widehat{D}_f(\widehat{q}\|p) \\
&\leq \max_{q\in\mathcal{Q}}\Big|D_f(q\|p) - \widehat{D}_f(q\|p)\Big| + \widehat{D}_f(\widetilde{q}\|p) \\
&\lesssim b_2(n,\gamma_1,\gamma_2) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \cdot \sqrt{\log\Big\{N_2\big[b_2(n,\gamma_1,\gamma_2),\mathcal{Q}\big]/\varepsilon\Big\}} + D_f(\widetilde{q}\|p),
\end{aligned} \tag{B.8}$$

where we recall that the ground truth t^* to the problem (5.2) takes the form $t^* = t^*(x; p, q) = f^\dagger(q/p)$. What's more, in the second inequality we use the optimality of $\widehat{q}$ among all $\widetilde{q} \in \mathcal{Q}$ to the problem (5.3), while the last line uses Lemma B.6. Moreover, note that (B.8) holds for all $\widetilde{q} \in \mathcal{Q}$, then by fixing $\widetilde{q} \in \mathcal{Q}$ to be the one such that $D_f(\widetilde{q}\|p)$ attains its minimum in the above (B.8), we obtain that

$$D_f(\widehat{q}\|p) \lesssim b_2(n, \gamma_1, \gamma_2) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \sqrt{\log\Big\{N_2\big[b_2(n, \gamma_1, \gamma_2), \mathcal{Q}\big]/\varepsilon\Big\}} + \min_{\widetilde{q} \in \mathcal{Q}} D_f(\widetilde{q}\|p).$$

This concludes the error bound of $\widehat{q}$ towards the desired density $q^* = p$.

C Lemmas and Proofs

C.1 Proof of Lemma B.1

Recall that the covering number of $\mathcal{Q}$ is $N_2(\delta, \mathcal{Q})$, we thus assume that there exists $q_1, \ldots, q_{N_2(\delta,\mathcal{Q})} \in \mathcal{Q}$ such that for any $q \in \mathcal{Q}$, there exists some q_k, where $1 \le k \le N_2(\delta, \mathcal{Q})$, so that $\|q - q_k\|_2 \le \delta$. Moreover, by taking $\delta = \delta_n = n^{-2\beta/(2\beta+d)}$ and union bound, we have

$$\begin{aligned}
&\mathbb{P}\bigg[\sup_{q\in\mathcal{Q}}\big|D_f(q\|p) - \widehat{D}_f(q\|p)\big| \ge c_1 \cdot n^{-\frac{2\beta}{2\beta+d}} \cdot \log^7 n\bigg] \\
&\quad\le \sum_{k=1}^{N_2(\delta_n,\mathcal{Q})} \mathbb{P}\Big[\big|D_f(q_k\|p) - \widehat{D}_f(q_k\|p)\big| \ge c_1 \cdot n^{-\frac{2\beta}{2\beta+d}} \cdot \log^7 n\Big] \\
&\quad\le N_2(\delta_n, \mathcal{Q}) \cdot \exp(-n^{\frac{d-2\beta}{2\beta+d}} \cdot \log^{14} n),
\end{aligned}$$

where the last line comes from Theorem 4.2. Combining Assumption 5.2, when n is sufficiently large, it holds that

$$\mathbb{P}\bigg[\sup_{q\in\mathcal{Q}}\big|D_f(q\|p) - \widehat{D}_f(q\|p)\big| \ge c_1 \cdot n^{-\frac{2\beta}{2\beta+d}} \cdot \log^7 n\bigg] \le 1/n,$$

which concludes the lemma.

C.2 Proof of Lemma B.2

By the definition of $\widehat{t}$, we have

$$\mathbb{E}_{\mathbb{P}_n}\big[f^\dagger(\widehat{t})\big] - \mathbb{E}_{\mathbb{Q}_n}[\widehat{t}] \le \mathbb{E}_{\mathbb{P}_n}\big[f^\dagger(\widetilde{t})\big] - \mathbb{E}_{\mathbb{Q}_n}[\widetilde{t}].$$

Note that the functional $G(t) = \mathbb{E}_{\mathbb{P}_n}[f^\dagger(t)] - \mathbb{E}_{\mathbb{Q}_n}[t]$ is convex, we then have

$$G\bigg(\frac{\widehat{t}+\widetilde{t}}{2}\bigg) - G(\widetilde{t}) \le \frac{G(\widehat{t}) - G(\widetilde{t})}{2} \le 0.$$

By re-arranging terms, we have

$$\begin{aligned}
&\bigg(\mathbb{E}_{\mathbb{P}_n}\Big\{f^\dagger\big[(\widehat{t}+\widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\} - \mathbb{E}_{\mathbb{P}}\Big[f^\dagger\big((\widehat{t}+\widetilde{t})/2\big) - f^\dagger(\widetilde{t})\Big]\bigg) - \Big\{\mathbb{E}_{\mathbb{Q}_n}\big[(\widehat{t}-\widetilde{t})/2\big] - \mathbb{E}_{\mathbb{Q}}\big[(\widehat{t}-\widetilde{t})/2\big]\Big\} \\
&\quad\le \mathbb{E}_{\mathbb{Q}}\big[(\widehat{t}-\widetilde{t})/2\big] - \mathbb{E}_{\mathbb{P}}\Big\{f^\dagger\big[(\widehat{t}+\widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\}. \qquad \text{(C.1)}
\end{aligned}$$

We denote by

$$B_f(\widetilde{t}, t) = \mathbb{E}_{\mathbb{P}}\big[f^\dagger(t) - f^\dagger(\widetilde{t})\big] - \mathbb{E}_{\mathbb{Q}}[t - \widetilde{t}]. \qquad \text{(C.2)}$$

then the RHS of (C.1) is exactly $-B_f[\widetilde{t}, (\widehat{t}+\widetilde{t})/2]$. We proceed to establish the lower bound of $B_f(\widetilde{t}, t)$ using $L_2(\mathbb{P})$ norm. Note that the ground truth t^* takes the form $t^* = f'(q/p)$, due to the

fact that $(f^\dagger)' \circ (f')(x) = x$, we know that $q/p = \partial f^\dagger(t^*)/\partial t$. Then by substituting the expectation taken over $\mathbb{Q}$ in (C.2) using the above relationship, we have

$$\begin{aligned} B_f(\widetilde{t}, t) &= \mathbb{E}_{\mathbb{P}}\Big[f^\dagger(t) - f^\dagger(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(t^*) \cdot (t - \widetilde{t})\Big] \\ &= \mathbb{E}_{\mathbb{P}}\Big[f^\dagger(t) - f^\dagger(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(\widetilde{t}) \cdot (t - \widetilde{t})\Big] + \mathbb{E}_{\mathbb{P}}\Big\{\Big[\frac{\partial f^\dagger}{\partial t}(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(t^*)\Big] \cdot (t - \widetilde{t})\Big\} \quad \text{(C.3)} \\ &= A_1 + A_2. \end{aligned}$$

Note that by Assumption 3.4 and Lemma D.6, we know that the Fenchel duality $f^\dagger$ is strongly convex with parameter $1/L_0$. This gives that

$$f^\dagger\big[t(x)\big] - f^\dagger\big[\widetilde{t}(x)\big] - \frac{\partial f^\dagger}{\partial t}\big[\widetilde{t}(x)\big] \cdot \big[t(x) - \widetilde{t}(x)\big] \geq 1/L_0 \cdot \big|t(x) - \widetilde{t}(x)\big|^2$$

for any x. Consequently we have

$$A_1 \geq 1/L_0 \cdot \|t - \widetilde{t}\|_{L_2(\mathbb{P})}^2.$$

Moreover, by Cauchy Schwarz inequality, the term A_2 in (C.3) is bounded as follows

$$A_2 \geq -\sqrt{\mathbb{E}_{\mathbb{P}}\Big\{\Big[\frac{\partial f^\dagger}{\partial t}(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(t^*)\Big]^2\Big\}} \cdot \sqrt{\mathbb{E}_{\mathbb{P}}\big[(t - \widetilde{t})^2\big]}.$$

Again, by Assumption 3.4 and Lemma D.6, we know that the Fenchel duality $f^\dagger$ has $1/\mu_0$-Lipschitz gradient, which gives that

$$\Big|\frac{\partial f^\dagger}{\partial t}\big[\widetilde{t}(x)\big] - \frac{\partial f^\dagger}{\partial t}\big[t^*(x)\big]\Big| \leq 1/\mu_0 \cdot \big|\widetilde{t}(x) - t^*(x)\big|$$

for any x. By this, we further bound A_2:

$$A_2 \geq -1/\mu_0 \cdot \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} \cdot \|t - \widetilde{t}\|_{L_2(\mathbb{P})}.$$

Combining both the the bound of A_1 and A_2, we have

$$B_f(\widetilde{t}, t) \geq 1/L_0 \cdot \|t - \widetilde{t}\|_{L_2(\mathbb{P})}^2 - 1/\mu_0 \cdot \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} \cdot \|t - \widetilde{t}\|_{L_2(\mathbb{P})}.$$

By this, together with (C.1), we conclude that

$$\begin{aligned} 1/(4L_0) \cdot \|\widehat{t} - \widetilde{t}\|_{L_2(\mathbb{P})}^2 \leq 1/\mu_0 \cdot \|\widehat{t} - \widetilde{t}\|_{L_2(\mathbb{P})} \cdot \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \Big\{\mathbb{E}_{\mathbb{Q}_n}\big[(\widehat{t} - \widetilde{t})/2\big] - \mathbb{E}_{\mathbb{Q}}\big[(\widehat{t} - \widetilde{t})/2\big]\Big\} \\ - \Big(\mathbb{E}_{\mathbb{P}_n}\Big\{f^\dagger\big[(\widehat{t} + \widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\} - \mathbb{E}_{\mathbb{P}}\Big\{f^\dagger\big[(\widehat{t} + \widetilde{t})/2\big] - f^\dagger(\widetilde{t})\Big\}\Big) \end{aligned}$$

This is exactly what we need.

C.3 Proof of Lemma B.3

We need the following notations. For any $K > 0$, the Bernstein difference of $t(x)$ with respect to the measure $\mathbb{P}$ is defined to be

$$\rho_{K,\mathbb{P}}^2(t) = 2K^2 \cdot \mathbb{E}_{\mathbb{P}}\Big\{\exp\big(|t(x)|/K\big) - 1 - |t(x)|/K\Big\}.$$

Correspondingly, we denote by $\mathcal{H}_{K,B}$ the generalized entropy with bracketing induced by Bernstein difference $\rho_{K,\mathbb{P}}$. We denote by $H_{s,B}$ the entropy with bracketing induced by L_s norm, H_s the entropy induced by L_s norm, $H_{L_s(\mathbb{P}),B}$ the entropy with bracketing induced by $L_s(\mathbb{P})$ norm, and $H_{L_s(\mathbb{P})}$ the regular entropy induced by $L_s(\mathbb{P})$ norm.

We consider the space

$$\Psi_M = \psi(\Phi_M) = \big\{\psi(t) : t(x) \in \Phi_M\big\}.$$

For any $\delta > 0$, we denote the following space

$$\Psi_M(\delta) = \Big\{\psi(t) \in \Psi_M : \big\|\psi(t) - \psi(\widetilde{t})\big\|_{L_2(\mathbb{P})} \le \delta\Big\},$$
$$\Psi'_M(\delta) = \Big\{\Delta\psi(t) = \psi(t) - \psi(\widetilde{t}) : \psi(t) \in \Psi_M(\delta)\Big\}.$$

Note that $\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \|\Delta\psi(t)\|_\infty \le 2M_0$ and $\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \|\Delta\psi(t)\|_\infty \le \delta$, by Lemma D.4 we have

$$\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \rho_{8M_0,\mathbb{P}}\big[\Delta\psi(t)\big] \le \sqrt{2}\delta.$$

To invoke Theorem D.3 for $\mathcal{G} = \Psi'_M(\delta)$, we consider $K = 8M_0, R = \sqrt{2}\delta$. Note that given $\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \|\Delta\psi(t)\|_\infty \le 2M_0$, by Lemma D.1 and Lemma D.2, and the fact that ψ is Lipschitz continuous, we have

$$\mathcal{H}_{8M_0,B}\big(u, \Psi'_M(\delta), \mathbb{P}\big) \le H_\infty\big(u/(2\sqrt{2}), \Psi'_M(\delta)\big) \le 2(s+1)\log\big(4\sqrt{2}u^{-1}(L+1)V^2\big)$$

for any $u > 0$. Then, by algebra, we have the follows

$$\int_0^R \mathcal{H}^{1/2}_{8M_0,B}\big(u, \Psi'_M(\delta), \mathbb{P}\big)\,\mathrm{d}u \le 3s^{1/2}\delta\cdot\log(8V^2L/\delta).$$

For any $0 < \varepsilon < 1$, we take $C = 1$, and a, C_1 and C_0 in Theorem D.3 to be

$$a = 8M_0\log\big(\exp(\gamma^2)/\varepsilon\big)\gamma^{-1}\cdot\delta,\ C_0 = 6M_0\gamma^{-1}\sqrt{\log\big(\exp(\gamma^2)/\varepsilon\big)},\ C_1 = 33M_0^2\gamma^{-2}\log\big(\exp(\gamma^2)/\varepsilon\big).$$

Here we recall that $\gamma = s^{1/2}\log(V^2L)$. Then it is straightforward to check that our choice above satisfies the conditions in Theorem D.3 for any δ such that $\delta \ge \gamma n^{-1/2}$, when n is sufficiently large such that $n \gtrsim [\gamma + \gamma^{-1}\log(1/\varepsilon)]^2$. Consequently, by Theorem D.3, for $\delta \ge \gamma n^{-1/2}$, we have

$$\begin{aligned}
&\mathbb{P}\bigg\{\sup_{t(x)\in\Phi_M(\delta)}\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big]\Big| \ge 8M_0\log(\exp(\gamma^2)/\varepsilon)\gamma^{-1}\cdot\delta\cdot n^{-1/2}\bigg\}\\
&\quad= \mathbb{P}\bigg\{\sup_{\Delta\psi(t)\in\Psi'_M(\delta)}\Big|\mathbb{E}_{\mathbb{P}_n}\big[\Delta\psi(t)\big] - \mathbb{E}_{\mathbb{P}}\big[\Delta\psi(t)\big]\Big| \ge 8M_0\log(\exp(\gamma^2)/\varepsilon)\gamma^{-1}\cdot\delta\cdot n^{-1/2}\bigg\}\\
&\quad\le \varepsilon\cdot\exp(-\gamma^2).
\end{aligned}$$

By taking $\delta = \delta_n = \gamma n^{-1/2}$, we have

$$\mathbb{P}\bigg\{\sup_{t(x)\in\Phi_M(\delta)}\frac{\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big]\Big|}{n^{-1}[\gamma^2 + \log(1/\varepsilon)]} \le 8M_0\bigg\} \ge 1 - \varepsilon\cdot\exp(-\gamma^2) \tag{C.4}$$

On the other hand, we denote that $S = \min\{s > 1 : 2^{-s}(2M_0) < \delta_n\} = \mathcal{O}(\log(\gamma^{-1}n^{1/2}))$. For notational simplicity, we denote the set

$$A_s = \Big\{\psi(t) \in \Psi_M : \psi(t) \in \Psi_M(2^{-s+2}M_0), \psi(t) \notin \Psi_M(2^{-s+1}M_0)\Big\}. \tag{C.5}$$

Then by the peeling device, we have the following

$$\begin{aligned}
&\mathbb{P}\bigg\{\sup_{\psi(t)\in\Psi_M,\psi(t)\notin\Psi_M(\delta_n)}\frac{\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big]\Big|}{\big\|\psi(t) - \psi(\widetilde{t})\big\|_{L_2(\mathbb{P})}\cdot T(n,\gamma,\varepsilon)} \ge 16M_0\bigg\}\\
\le&\sum_{s=1}^{S}\mathbb{P}\bigg\{\sup_{\psi(t)\in A_s}\frac{\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big]\Big|}{2^{-s+1}M_0} \ge 16M_0\cdot T(n,\gamma,\varepsilon)\bigg\}\\
\le&\sum_{s=1}^{S}\mathbb{P}\bigg\{\sup_{\psi(t)\in A_s}\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big]\Big| \ge 8M_0\cdot(2^{-s+2}M_0)\cdot T(n,\gamma,\varepsilon)\bigg\}\\
\le&\sum_{s=1}^{S}\mathbb{P}\bigg\{\sup_{\psi(t)\in\Psi_M(2^{-s+2}M_0)}\Big|\mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big]\Big| \ge 8M_0\cdot(2^{-s+2}M_0)\cdot T(n,\gamma,\varepsilon)\bigg\}\\
\le&S\cdot\varepsilon\cdot\exp(-\gamma^2)/\log(\gamma^{-1}n^{1/2}) = c\cdot\varepsilon\cdot\exp(-\gamma^2),
\end{aligned}$$

where c is a positive absolute constant, and for notational convenience we denote by $T(n, \gamma, \varepsilon) = \gamma^{-1} \cdot n^{-1/2} \log(\log(\gamma^{-1} n^{1/2}) \exp(\gamma^2)/\varepsilon)$. Here in the second line, we use the fact that for any $\psi(t) \in A_s$, we have $\|\psi(t) - \psi(\widetilde{t})\|_{L_2(\mathbb{Q})} \geq 2^{-s+1} M_0$ by the definition of A_s in (C.5); in the forth line, we use the argument that since $A_s \subseteq \Psi_M(2^{-s+2} M_0)$, the probability of supremum taken over $\Psi_M(2^{-s+2} M_0)$ is larger than the one over A_s; in the last line we again invoke Theorem D.3. Consequently, this gives us

$$\mathbb{P}\Bigg\{ \sup_{\substack{\psi(t) \in \Psi_M \\ \psi(t) \notin \Psi_M(\delta_n)}} \frac{\Big| \mathbb{E}_{\mathbb{P}_n}\big[\psi(t) - \psi(\widetilde{t})\big] - \mathbb{E}_{\mathbb{P}}\big[\psi(t) - \psi(\widetilde{t})\big] \Big|}{\big\|\psi(t) - \psi(\widetilde{t})\big\|_{L_2(\mathbb{P})} \cdot n^{-1/2}\big[\gamma \log n + \gamma^{-1} \log(1/\varepsilon)\big]} \leq 16 M_0 \Bigg\} \geq 1 - \varepsilon \cdot \exp(-\gamma^2). \tag{C.6}$$

Combining (C.4) and (C.6), we conclude the lemma.

C.4 Proof of Lemma B.5

The proof of the theorem utilizes following two lemmas. The first lemma characterizes the Lipschitz property of $\varphi(x; W, v)$ in the input x.

Lemma C.1. Given W and v, then for any $\varphi(\cdot; W, v) \in \Phi_{\text{norm}}$ and $x_1, x_2 \in \mathbb{R}^d$, we have

$$\big\|\varphi(x_1; W, v) - \varphi(x_2; W, v)\big\|_2 \leq \|x_1 - x_2\|_2 \cdot \prod_{j=1}^{L+1} B_j.$$

Proof. See Section §C.6 for a detailed proof. □

The following lemma characterizes the Lipschitz property of $\varphi(x; W, v)$ in the network parameter pair (W, v).

Lemma C.2. Given any bounded $x \in \mathbb{R}^d$ such that $\|x\|_2 \leq B$, then for any weights $W^1 = \{W_j^1\}_{j=1}^{L+1}, W^2 = \{W_j^2\}_{j=1}^{L+1}, v^1 = \{v_j^1\}_{j=1}^{L}, v^2 = \{v_j^2\}_{j=1}^{L}$, and functions $\varphi(\cdot, W^1, v^1), \varphi(\cdot, W^2, v^2) \in \Phi_{\text{norm}}$, we have

$$\big\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\big\| \leq \frac{B\sqrt{2L+1} \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \sqrt{\sum_{j=1}^{L+1} \|W_j^1 - W_j^2\|_{\mathrm{F}}^2 + \sum_{j=1}^{L} \|v_j^1 - v_j^2\|_2^2}.$$

Proof. See Section §C.7 for a detailed proof. □

We now turn to the proof of Lemma B.5. Note that by Lemma C.2, we know that $\varphi(x; W, v)$ is L_w-Lipschitz in the parameter $(W, v) \in \mathbb{R}^b$, where the dimension b takes the form

$$b = \sum_{j=1}^{L+1} k_j k_{j-1} + \sum_{j=1}^{L} k_j \leq \sum_{j=0}^{L+1} (k_j + 1)^2, \tag{C.7}$$

and the Lipschitz constant L_w satisfies

$$L_w = \frac{B\sqrt{2L+1} \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j}. \tag{C.8}$$

In addition, we know that the covering number of $\mathcal{W} = \{(W, v) \in \mathbb{R}^b : \sum_{j=1}^{L+1} \|W_j\|_{\mathrm{F}} + \sum_{j=1}^{L} \|v_j\|_2 \leq K\}$, where

$$K = \sqrt{\sum_{j=1}^{L+1} k_j^2 B_j^2} + \sum_{j=1}^{L} A_j, \tag{C.9}$$

satisfies

$$N(\mathcal{W}, \delta) \leq \left(\frac{3K}{\delta}\right)^b.$$

By the above facts, we deduce that the covering number of $\mathcal{L}(\Phi_{\text{norm}})$ satisfies

$$N\big[\mathcal{L}(\Phi_{\text{norm}}), \delta\big] \leq \left(\frac{c_1 K L_w}{\delta}\right)^b,$$

for some positive absolute constant c_1. Then by Dudley entropy integral bound on the ERC, we know that

$$\mathfrak{R}_n\big[\mathcal{L}(\Phi_{\text{norm}})\big] \leq \inf_{\tau>0} \tau + \frac{1}{\sqrt{n}} \int_\tau^\vartheta \sqrt{\log N\big[\mathcal{L}(\Phi_{\text{norm}}), \delta\big]}\, \mathrm{d}\delta, \tag{C.10}$$

where $\vartheta = \sup_{g(\cdot;W,v)\in\mathcal{L}(\Phi_{\text{norm}}), x\in\mathbb{R}^d} \big|g(x; W, v)\big|$. Moreover, from Lemma C.1 and the fact that the loss function is Lipschitz continuous, we have

$$\vartheta \leq c_2 \cdot B \cdot \prod_{j=1}^{L+1} B_j \tag{C.11}$$

for some positive absolute constant c_2. Therefore, by calculations, we derive from (C.10) that

$$\mathfrak{R}_n\big[\mathcal{L}(\Phi_{\text{norm}})\big] = \mathcal{O}\left(\frac{\vartheta}{\sqrt{n}} \cdot \sqrt{b \cdot \log \frac{K L_w \sqrt{n}}{\vartheta \sqrt{b}}}\right),$$

then we conclude the lemma simply by plugging in (C.7), (C.8), (C.9) and (C.11), and using the definition of γ_1 and γ_2 in (A.3).

C.5 Proof of Lemma B.6

Remember that the covering number of $\mathcal{Q}$ is $N_2(\delta, \mathcal{Q})$, we assume that there exists $q_1, \ldots, q_{N_2(\delta,\mathcal{Q})} \in \mathcal{Q}$ such that for any $q \in \mathcal{Q}$, there exists some q_k, where $1 \leq k \leq N_2(\delta, \mathcal{Q})$, so that $\|q - q_k\|_2 \leq \delta$. Moreover, by taking $\delta = \gamma_1 n^{-1/2} \log(\gamma_2 n) = b_2(n, \gamma_1, \gamma_2)$ and $N_2 = N_2[b_2(n, \gamma_1, \gamma_2), \mathcal{Q}]$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{q\in\mathcal{Q}} \Big|D_f(q\|p) - \widehat{D}_f(q\|p)\Big| \geq c \cdot \Big[b_2(n, \gamma_1, \gamma_2) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \cdot \sqrt{\log(N_2/\varepsilon)}\Big]\right\} \\
&\quad\leq \sum_{k=1}^{N_2} \mathbb{P}\left\{\Big|D_f(q\|p) - \widehat{D}_f(q\|p)\Big| \geq c \cdot \Big[b_2(n, \gamma_1, \gamma_2) + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2} \cdot \sqrt{\log(N_2/\varepsilon)}\Big]\right\} \\
&\quad\leq N_2 \cdot \varepsilon / N_2 = \varepsilon,
\end{aligned}$$

where the second line comes from union bound, and the last line comes from Theorem A.3. By this, we conclude the proof.

C.6 Proof of Lemma C.1

The proof follows by applying the Lipschitz property and bounded spectral norm of W_j recursively:

$$\begin{aligned}
\big\|\varphi(x_1; W, v) - \varphi(x_2; W, v)\big\|_2 &= \Big\|W_{L+1}\big(\sigma_{v_L} \cdots W_2 \sigma_{v_1} W_1 x_1 - \sigma_{v_L} \cdots W_2 \sigma_{v_1} W_1 x_2\big)\Big\|_2 \\
&\leq \|W_{L+1}\|_2 \cdot \Big\|\sigma_{v_L}\big(W_L \cdots W_2 \sigma_{v_1} W_1 x_1 - W_L \cdots W_2 \sigma_{v_1} W_1 x_2\big)\Big\|_2 \\
&\leq B_{L+1} \cdot \Big\|W_L \cdots W_2 \sigma_{v_1} W_1 x_1 - W_L \cdots W_2 \sigma_{v_1} W_1 x_2\Big\|_2 \\
&\leq \cdots \leq \prod_{j=1}^{L+1} B_j \cdot \|x_1 - x_2\|_2.
\end{aligned}$$

Here in the third line we uses the fact that $\|W_j\|_2 \leq B_j$ and the 1-Lipschitz property of $\sigma_{v_j}(\cdot)$, and in the last line we recursively apply the same argument as in the above lines. This concludes the proof.

C.7 PROOF OF LEMMA C.2

Remember that $\varphi(x; W, v)$ takes the form

$$\varphi(x; W, v) = W_{L+1}\sigma_{v_L}W_L \cdots \sigma_{v_1}W_1 x.$$

For the sake of notational simplicity, we further denote $\varphi_j^i(x) = \sigma_{v_j^i}W_j^i x$ for $i = 1, 2$. By this, $\varphi(x; W, v)$ has the form $\varphi(x; W^i, v^i) = W_{L+1}^i\varphi_L^i \circ \cdots \circ \varphi_1^i(x)$. First, note that for any W^1, W^2, v^1 and v^2, by triangular inequality, we have

$$\begin{aligned}
\big\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\big\|_2 &= \big\|W_{L+1}^1\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - W_{L+1}^2\varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2 \\
&\le \big\|W_{L+1}^1\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - W_{L+1}^2\varphi_L^1 \circ \cdots \circ \varphi_1^1(x)\big\|_2 \\
&\qquad + \big\|W_{L+1}^2\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - W_{L+1}^2\varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2 \\
&\le \|W_{L+1}^1 - W_{L+1}^2\|_{\mathrm{F}} \cdot \big\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x)\big\|_2 \\
&\qquad + B_{L+1} \cdot \big\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2. \qquad \text{(C.12)}
\end{aligned}$$

Moreover, note that for any $\ell \in [L]$, we have the follows bound on $\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x)\|_2$:

$$\begin{aligned}
\big\|\varphi_\ell^i \circ \cdots \circ \varphi_1^i(x)\big\|_2 &\le \big\|W_\ell^i\varphi_{\ell-1}^i \circ \cdots \circ \varphi_1^i(x)\big\|_2 \\
&\le B_\ell \cdot \big\|\varphi_{\ell-1}^i \circ \cdots \circ \varphi_1^i(x)\big\|_2 \le \|x\|_2 \cdot \prod_{j=1}^{\ell} B_j,
\end{aligned}$$

where the first inequality we use the 1-Lipschitz property of the ReLU activator, and the second inequality uses the bounded spectral norm of W_j^i, while the last inequality simply applies the previous arguments recursively. Therefore, combining (C.12), we have

$$\begin{aligned}
\big\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\big\|_2 &\le B \cdot \prod_{j=1}^{L} B_j \cdot \|W_{L+1}^1 - W_{L+1}^2\|_F \\
&\qquad + B_{L+1} \cdot \big\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2. \qquad \text{(C.13)}
\end{aligned}$$

Similarly, by triangular inequality, we have

$$\begin{aligned}
&\big\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2 \\
&\quad\le \big\|\varphi_L^1 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x)\big\|_2 \\
&\qquad + \big\|\varphi_L^2 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \varphi_{L-1}^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2 \\
&\quad\le \big\|\varphi_L^1 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x)\big\|_2 \\
&\qquad + B_L \cdot \big\|\varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_{L-1}^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2, \qquad \text{(C.14)}
\end{aligned}$$

where the second inequality uses the bounded spectral norm of W_L and 1-Lipschitz property of $\sigma_{v_L}(\cdot)$. For notational convenience, we further denote $y = \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x)$, then

$$\big\|\varphi_L^1(y) - \varphi_L^2(y)\big\|_2 = \Big\|\max\big\{W_L^1 y - v_L^1, 0\big\} - \max\big\{W_L^2 y - v_L^2, 0\big\}\Big\|_2 \le \|v_L^1 - v_L^2\|_2. \qquad \text{(C.15)}$$

By (C.14) and (C.15), we have

$$\begin{aligned}
&\big\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2 \\
&\quad\le \big\|v_L^1 - v_L^2\big\|_2 + B_L \cdot \big\|\varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_{L-1}^2 \circ \cdots \circ \varphi_1^2(x)\big\|_2 \\
&\quad\le \sum_{j=1}^{L} \prod_{i=j+1}^{L} B_i \cdot \big\|v_j^1 - v_j^2\big\|_2 + \frac{B \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \sum_{j=1}^{L} \|W_j^1 - W_j^2\|_{\mathrm{F}} \\
&\quad\le \frac{B \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \sum_{j=1}^{L} \Big(\big\|v_j^1 - v_j^2\big\|_2 + \|W_j^1 - W_j^2\|_{\mathrm{F}}\Big).
\end{aligned}$$

Here in the third line we recursively apply the previous arguments. Further combining (C.13), we obtain that

$$\begin{aligned}
&\big\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\big\|_2 \\
&\quad\le \frac{B \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \left[\sum_{j=1}^{L+1} \big\|W_j^1 - W_j^2\big\|_{\mathrm{F}} + \sum_{j=1}^{L} \big\|v_j^1 - v_j^2\big\|_2\right] \\
&\quad\le \frac{B\sqrt{2L+1} \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \sqrt{\sum_{j=1}^{L+1} \|W_j^1 - W_j^2\|_{\mathrm{F}}^2 + \sum_{j=1}^{L} \|v_j^1 - v_j^2\|_2^2},
\end{aligned}$$

where we use Cauchy-Schwarz inequality in the last line. This concludes the proof of the lemma.

D Auxiliary Results

Lemma D.1. The following statements for entropy hold.

1. Suppose that $\sup_{g\in\mathcal{G}} \|g\|_\infty \le M$, then
$$\mathcal{H}_{4M,B}(\sqrt{2}\delta, \mathcal{G}, \mathbb{Q}) \le H_{2,B}(\delta, \mathcal{G}, \mathbb{Q})$$
for any $\delta > 0$.
2. For all $1 \le q < \infty$, and $\mathbb{Q}$ a probability measure, we have
$$H_{p,B}(\delta, \mathcal{G}, \mathbb{Q}) \le H_\infty(\delta/2, \mathcal{G}),$$
for any $\delta > 0$. Here H_∞ is the entropy induced by infinity norm.
3. Based on the above two statements, suppose that $\sup_{g\in\mathcal{G}} \|g\|_\infty \le M$, we have
$$\mathcal{H}_{4M,B}(\sqrt{2}\cdot\delta, \mathcal{G}, \mathbb{Q}) \le H_\infty(\delta/2, \mathcal{G}),$$
by taking $p = 2$.

Proof. See Lemma 5.10 in (van de Geer & van de Geer, 2000) for a detailed proof. □

Lemma D.2. The entropy of the neural network set defined in (4.1) satisfies

$$H_\infty\big[\delta, \Phi_M(L, p, s)\big] \le (s+1)\log\big[2\delta^{-1}(L+1)V^2\big],$$

where $V = \prod_{l=0}^{L+1}(p_l + 1)$.

Proof. See (Schmidt-Hieber, 2017) for a detailed proof. □

Theorem D.3. Let the space $\mathcal{G}$ satisfy $\sup_{g\in\mathcal{G}} \rho_K(g) \le R$. Take a, C, C_0, C_1 satisfying that $a \le C_1\sqrt{n}R^2/K, a \le 8\sqrt{n}R, a \ge C_0 \cdot [\int_0^R H_{K,B}^{1/2}(u, \mathcal{G}, \mathbb{P})du \vee R]$ and $C_0^2 \ge C^2(C_1+1)$. Then

$$\mathbb{P}\left\{\sup_{g\in\mathcal{G}}\big|\mathbb{E}_{\mathbb{P}_n}[g] - \mathbb{E}_{\mathbb{P}}[g]\big| \ge a \cdot n^{-1/2}\right\} \le C\exp\left[-\frac{a^2}{C^2(C_1+1)R^2}\right].$$

Proof. See Theorem 5.11 in (van de Geer & van de Geer, 2000) for a detailed proof. □

Lemma D.4. Suppose that $\|g\|_\infty \le K$, and $\|g\| \le R$, then $\rho_{2K,\mathbb{P}}^2(g) \le 2R^2$. Moreover, for any $K' \ge K$, we have $\rho_{2K',\mathbb{P}}^2(g) \le 2R^2$.

Proof. See (van de Geer & van de Geer, 2000) for a detailed proof. □

Theorem D.5. For any function f in the Hölder ball $\mathcal{C}_d^\beta([0,1]^d, K)$ and any integers $m \geq 1$ and $N \geq (\beta+1)^d \vee (K+1)$, there exists a network $\widetilde{f} \in \Phi(L, (d, 12dN, \ldots, 12dN, 1), s)$ with number of layers $L = 8+(m+5)(1+\lceil \log_2 d \rceil)$ and number of parameters $s \leq 94d^2(\beta+1)^{2d}N(m+6)(1+\lceil \log_2 d \rceil)$, such that

$$\|\widetilde{f} - f\|_{L^\infty([0,1]^d)} \leq (2K+1)3^{d+1}N2^{-m} + K2^\beta N^{-\beta/d}.$$

Proof. See (Schmidt-Hieber, 2017) for a detailed proof. □

Lemma D.6. If the function f is strongly convex with parameter $\mu_0 > 0$ and has Lipschitz continuous gradient with parameter $L_0 > 0$, then the Fenchel duality $f^\dagger$ of f is $1/L_0$-strongly convex and has $1/\mu_0$-Lipschitz continuous gradient (therefore, $f^\dagger$ itself is Lipschitz continuous).

Proof. See (Zhou, 2018) for a detailed proof. □