Supplementary Materials

8 Auxiliary Results via Deep Neural Networks

8.1 Estimation via Deep Neural Networks

Since the most general estimator $D_f^{\natural}(q\|p)$ proposed in (4.2) requires solving an optimization problem over a function space, which is usually intractable, we introduce an estimator of the f-divergence $D_f(q\|p)$ using the family of deep neural networks in this section. We now define the family of deep neural networks as follows.

Definition 8.1. 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 deep neural networks is defined as

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

where $\sigma_v(x) = \max\{0, x - v\}$ is the ReLU activation function.

To avoid overfitting, the sparsity of the deep neural networks is a typical assumption in deep learning literature. In practice, such a sparsity property is achieved through certain techniques, e.g., dropout (Srivastava et al., 2014), or certain network architecture, e.g., convolutional neural network (Krizhevsky et al., 2012). We now define the family of sparse neural networks as follows,

$$\begin{aligned}\Phi_M(L, k, s) = \Big\{&\varphi(x; W, v) \in \Phi(L, d)\colon \|\varphi\|_\infty \le M, \|W_j\|_\infty \le 1 \text{ for } j \in [L+1], \\ &\|v_j\|_\infty \le 1 \text{ for } j \in [L], \ \sum_{j=1}^{L+1}\|W_j\|_0 + \sum_{j=1}^{L}\|v_j\|_0 \le s\Big\},\end{aligned} \tag{8.1}$$

where s is the sparsity. In contrast, another approach to avoid overfitting in deep learning literature is to control the norm of parameters (Li et al., 2018). See Section §8.4 for details.

Consider the following estimators via deep neural networks,

$$\begin{aligned}\widehat{t}(\cdot; p, q) &= \mathop{\mathrm{argmin}}_{t \in \Phi_M(L,k,s)} \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{x\sim\mathbb{Q}_n}[t(x)], \\ \widehat{D}_f(q\|p) &= \mathbb{E}_{x\sim\mathbb{Q}_n}[\widehat{t}(x; p, q)] - \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(\widehat{t}(x; p, q))].\end{aligned} \tag{8.2}$$

The following theorem characterizes the statistical rate of convergence of the estimator proposed in (8.2).

Theorem 8.2. Let $L = \mathcal{O}(\log n)$, $s = \mathcal{O}(N \log n)$, and $k = (d, d, \mathcal{O}(dN), \mathcal{O}(dN), \ldots, \mathcal{O}(dN), 1)$ in (8.1), where $N = n^{d/(2\beta+d)}$. Under Assumptions 3.1, 3.3, and 3.4, if $d < 2\beta$, then with probability at least $1 - \exp\{-n^{d/(2\beta+d)}\log^5 n\}$, we have

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

We defer the proof of the theorem in Section §10.4. By Theorem 8.2, the estimators in (8.2) achieve the optimal nonparametric rate of convergence (Stone, 1982) up to a logarithmic term. We can see that by setting $\gamma_\Phi = d/\beta$ in Theorem 4.3, we recover the result in Theorem 8.2. By (3.2) and Theorem 8.2, we have

$$\delta(n) = 1 - \exp\{-n^{d/(2\beta+d)}\log^5 n\}, \qquad \epsilon(n) = c \cdot n^{-\beta/(2\beta+d)}\log^{7/2} n,$$

where c is a positive absolute constant.

8.2 Reconstruction via Deep Neural Networks

To utilize the estimator $\widehat{D}_f(q\|p)$ proposed via deep neural networks in Section §8.1, we propose the following estimator,

$$\widehat{q} = \operatorname*{argmin}_{q\in\mathcal{Q}} \widehat{D}_f(q\|p), \tag{8.3}$$

where $\widehat{D}_f(q\|p)$ is given in (8.2).

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

Assumption 8.3. We have $N_2(\delta, \mathcal{Q}) = \mathcal{O}(\exp\{\delta^{-d/\beta}\})$.

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

Theorem 8.4. Under the same assumptions in Theorem 8.2, further if Assumption 8.3 holds, for sufficiently large sample size n, with probability at least $1 - 1/n$, we have

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

The proof of Theorem 8.4 is deferred in Section §10.6. We can see that by setting $\gamma_\Phi = d/\beta$ in Theorem 8.4, we recover the result in Theorem 5.2.

8.3 Auxiliary Results on Sparsity Control

In this section, we provide some auxiliary results on (8.2). We first state an oracle inequality showing the rate of convergence of $\widehat{t}(x; p, q)$.

Theorem 8.5. Given $0 < \varepsilon < 1$, for any sample size n satisfies that $n \gtrsim [\gamma + \gamma^{-1}\log(1/\varepsilon)]^2$, under Assumptions 3.1, 3.3, and 3.4, it holds that

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

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

We defer the proof of to Section §10.7.

As a by-product, note that $t^*(x; p, q) = f'(\theta^*(x; p, q)) = f'(q(x)/p(x))$, based on the error bound established in Theorem 8.5, we obtain the following result.

Corollary 8.6. Given $0 < \varepsilon < 1$, for the sample size $n \gtrsim [\gamma + \gamma^{-1}\log(1/\varepsilon)]^2$, under Assumptions 3.1, 3.3, and 3.4, it holds with probability at least $1 - \varepsilon\cdot\exp(-\gamma^2)$ that

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

Here $\gamma = s^{1/2}\log(V^2L)$ and $V = \prod_{j=0}^{L+1}(k_j+1)$.

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 12.6, we obtain the result from Theorem 8.5. □

8.4 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 (8.1)) to control the error bound, which is an alternative of the network model shown in (8.1). 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

$$\begin{aligned}\Phi_{\text{norm}} = \Phi_{\text{norm}}(L, k, A, B) = \{\varphi(x; W, v) \in \Phi(L, k) : \|v_j\|_2 \le A_j \text{ for all } j \in [L], \\ \|W_j\|_2 \le B_j \text{ for all } j \in [L+1]\},\end{aligned} \tag{8.4}$$

where $\sigma_{v_j}(x) = \max\{0, x - v_j\}$ for any $j \in [L]$. We write the following optimization problem,

$$\begin{aligned} \widehat{t}(x; p, q) &= \operatorname*{argmin}_{t \in \Phi_{\text{norm}}} \mathbb{E}_{x \sim \mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{x \sim \mathbb{Q}_n}[t(x)], \\ \widehat{D}_f(q\|p) &= \mathbb{E}_{x \sim \mathbb{Q}_n}[\widehat{t}(x; p, q)] - \mathbb{E}_{x \sim \mathbb{P}_n}[f^\dagger(\widehat{t}(x; p, q))]. \end{aligned} \tag{8.5}$$

Based on this formulation, we derive the error bound on the estimated f-divergence in the following theorem. We only consider the generalization error in this setting. Therefore, we assume that the ground truth $t^*(x; p, q) = f'(q(x)/p(x)) \in \Phi_{\text{norm}}$. 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,

$$\gamma_1 = B \prod_{j=1}^{L+1} B_j \cdot \sqrt{\sum_{j=0}^{L+1} k_j^2}, \qquad \gamma_2 = \frac{L \cdot (\sqrt{\sum_{j=1}^{L+1} k_j^2 B_j^2} + \sum_{j=1}^{L} A_j)}{\sum_{j=0}^{L+1} k_j^2 \cdot \min_j B_j^2} \cdot \sum_{j=1}^{L} A_j. \tag{8.6}$$

Now, we state the theorem.

Theorem 8.7. We assume that $t^*(x; p, q) \in \Phi_{\text{norm}}$. Then for any $0 < \varepsilon < 1$, with probability at least $1 - \varepsilon$, it holds that

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

where γ_1 and γ_2 are defined in (8.6).

We defer the proof to Section §10.8.

The next theorem characterizes the rate of convergence of $\widehat{q} = \operatorname{argmin}_{q \in \mathcal{Q}} \widehat{D}_f(q\|p)$, where $\widehat{D}_f(q\|p)$ is proposed in (8.5).

Theorem 8.8. For any $0 < \varepsilon < 1$, with probability at least $1 - \varepsilon$, we have

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

where $b_2(n, \gamma_1, \gamma_2) = \gamma_1 n^{-1/2} \log(\gamma_2 n)$, and $N_2(\delta, \mathcal{Q})$ is the covering number of $\mathcal{Q}$.

We defer the proof to Section §10.9.

9 Exemplary $\widehat{t}$ and $f^\dagger$

As for experiments on MNIST and CIFAR-10, we choose to skip Step 1 in Algorithm 1 and 2 and instead adopt $\widehat{t}$ and $f^\dagger$ as suggested by (Nowozin et al., 2016). Exemplary $\widehat{t}$ and $f^\dagger$ are specified in Table 2.

Table 2: Exemplary $\widehat{t}$, $f^\dagger$.

Name	$D_f(\mathbb{P}\|\mathbb{Q})$	$\widehat{t}(v)$	$\text{dom}_{f^\dagger}$	$f^\dagger(u)$
Total Variation	$\int \frac{1}{2}\|p(z) - q(z)\|dz$	$\frac{1}{2}\tanh(v)$	$u \in [-\frac{1}{2}, \frac{1}{2}]$	u
Jenson-Shannon	$\int p(x) \log \frac{p(z)}{q(z)}$	$\log \frac{2}{1 + e^{-v}}$	$u < \log 2$	$-\log(2 - e^u)$
Squared Hellinger	$\int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz$	$1 - e^v$	$u < 1$	$\frac{u}{1-u}$
Pearson $\mathbf{x}$	$\int \frac{(q(z) - p(z))^2}{p(z)} dz$	v	$\mathbb{R}$	$\frac{1}{4}u^2 + u$
Neyman $\mathbf{x}$	$\int \frac{(p(z) - q(z))^2}{p(z)} dz$	$1 - e^v$	$u < 1$	$2 - 2\sqrt{1-u}$
KL	$\int p(z) \log \frac{p(z)}{q(z)} dx$	v	$\mathbb{R}$	e^{u-1}
Reverse KL	$\int q(z) \log \frac{q(z)}{p(z)} dz$	$-e^v$	$\mathbb{R}_-$	$-1 - \log(-u)$
Jeffrey	$\int (q(z) - p(z)) \log \frac{p(z)}{q(z)} dz$	v	$\mathbb{R}$	$W(e^{1-u}) + \frac{1}{W(e^{1-u})} + u - 2$

10 Proofs of Theorems

10.1 Proof of Theorem 3.5

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

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

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

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

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

10.2 Proof of Theorem 3.7

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

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

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

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

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.

10.3 Proof of Theorem 4.3

We first show the convergence of $t^\natural$, and then the convergence of $D_f^\natural(q\|p)$. For any real-valued function ϱ, we write $\mathbb{E}_\mathbb{P}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}}[\varrho(x)]$, $\mathbb{E}_\mathbb{Q}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{P}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}_n}[\varrho(x)]$, and $\mathbb{E}_{\mathbb{Q}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\varrho(x)]$ for notational convenience.

For any $\widetilde{t}\in\Phi$, we establish the following lemma.

Lemma 10.1. Under the assumptions stated in Theorem 4.3, it holds that

$$
\begin{aligned}
1/(4L_0)\cdot\|t^\natural - t^*\|^2_{L_2(\mathbb{P})} \le \{&\mathbb{E}_{\mathbb{Q}_n}[(t^\natural - t^*)/2] - \mathbb{E}_\mathbb{Q}[(t^\natural - t^*)/2]\} \\
&- \{\mathbb{E}_{\mathbb{P}_n}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)] - \mathbb{E}_\mathbb{P}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)]\}.
\end{aligned}
$$

Here μ_0 and L_0 are specified in Assumption 3.4.

We defer the proof to Section §11.1.

Note that by Lemma 10.1 and the fact that $f^\dagger$ is Lipschitz continuous, we have

$$
\begin{aligned}
\|t^\natural - t^*\|^2_{L_2(\mathbb{P})} \lesssim \{&\mathbb{E}_{\mathbb{Q}_n}[(t^\natural - t^*)/2] - \mathbb{E}_\mathbb{Q}[(t^\natural - t^*)/2]\} \\
&- \{\mathbb{E}_{\mathbb{P}_n}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)] - \mathbb{E}_\mathbb{P}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)]\}. \qquad (10.1)
\end{aligned}
$$

Further, to upper bound the RHS of (10.15), we establish the following lemma.

Lemma 10.2. 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 under the assumptions stated in Theorem 8.5, we have

$$
\mathbb{P}\bigg\{\sup_{t\colon\,\psi(t)\in\Psi}\frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t)-\psi(t^*)] - \mathbb{E}_\mathbb{P}[\psi(t)-\psi(t^*)]|}{n^{-2/(\gamma_\Phi+2)}} \ge c_2\bigg\} \le c_1\exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2),
$$

where c_1 and c_2 are positive absolute constants.

We defer the proof to Section §11.2.

Note that the results in Lemma 10.2 also apply to the distribution $\mathbb{Q}$, and by using the fact that the true density ratio $\theta^*(x;p,q) = q(x)/p(x)$ is bounded below and above, we know that $L_2(\mathbb{Q})$ is indeed equivalent to $L_2(\mathbb{P})$. We thus focus on $L_2(\mathbb{P})$ here. By (10.1), Lemma 10.2, and the Lipschitz property of $f^\dagger$ according to Lemma 12.6, with probability at least $1-c_1\exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2)$, we have

$$
\|t^\natural - t^*\|_{L_2(\mathbb{P})} \lesssim n^{-1/(\gamma_\Phi+2)}. \qquad (10.2)
$$

Note that we have

$$
\begin{aligned}
&|D_f^\natural(q\|p) - D_f(q\|p)| \\
&\quad\le |\mathbb{E}_{\mathbb{Q}_n}[t^\natural - t^*] - \mathbb{E}_\mathbb{Q}[t^\natural - t^*]| + |\mathbb{E}_{\mathbb{P}_n}[f^\dagger(t^\natural) - f^\dagger(t^*)] - \mathbb{E}_\mathbb{P}[f^\dagger(t^\natural) - f^\dagger(t^*)]| \\
&\qquad\quad + |\mathbb{E}_\mathbb{Q}[t^\natural - t^*] - \mathbb{E}_\mathbb{P}[f^\dagger(t^\natural) - f^\dagger(t^*)]| + |\mathbb{E}_{\mathbb{Q}_n}[t^*] - \mathbb{E}_\mathbb{Q}[t^*]| + |\mathbb{E}_{\mathbb{P}_n}[f^\dagger(t^*)] - \mathbb{E}_\mathbb{P}[f^\dagger(t^*)]| \\
&\quad= B_1 + B_2 + B_3 + B_4 + B_5. \qquad (10.3)
\end{aligned}
$$

We upper bound B_1, B_2, B_3, B_4, and B_5 in the sequel. First, by Lemma 10.2, with probability at least $1-c_1\exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2)$, we have

$$
B_1 \lesssim n^{-2/(\gamma_\Phi+2)}. \qquad (10.4)
$$

Similar upper bound also holds for B_2. Also, following from (10.2), with probability at least $1-c_1\exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2)$, we have

$$
B_3 \lesssim n^{-1/(\gamma_\Phi+2)}. \qquad (10.5)
$$

Meanwhile, by Hoeffding's inequality, with probability at least $1 - c_1 \exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2)$, we have

$$B_4 \lesssim n^{-1/(\gamma_\Phi+2)}. \tag{10.6}$$

Similar upper bound also holds for B_5. Now, combining (10.3), (10.4), (10.5), and (10.6), with probability at least $1 - c_1 \exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2)$, we have

$$|D^{\natural}_f(q\|p) - D_f(q\|p)| \lesssim n^{-1/(\gamma_\Phi+2)}.$$

We conclude the proof of Theorem 4.3.

10.4 Proof of Theorem 8.2

Step 1. We upper bound $\|t^* - \widehat{t}\|_{L_2(\mathbb{P})}$ in the sequel. Note that $t^* \in \Omega \subset [a,b]^d$. To invoke Theorem 12.5, we denote by $t'(y) = t^*((b-a)y + a\mathbf{1}_d)$, where $\mathbf{1}_d = (1,1,\ldots,1)^\top \in \mathbb{R}^d$. Then the support of t' 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 12.5 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 $\widetilde{t}(\cdot) = \widetilde{t'} \circ \ell(\cdot)$, where $\ell(\cdot)$ is a linear mapping taking the following form,

$$\ell(x) = \frac{x}{b-a} - \frac{a}{b-a} \cdot \mathbf{1}_d.$$

To this end, we know that $\widetilde{t} \in \Phi_M(L,k,s)$, with parameters L, k, and s given in the statement of Theorem 8.2. We fix this $\widetilde{t}$ and invoke Theorem 8.5, then with probability at least $1 - \varepsilon \cdot \exp(-\gamma^2)$, we have

$$\begin{aligned}
\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} &\lesssim \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2} \log n + n^{-1/2}[\sqrt{\log(1/\varepsilon)} + \gamma^{-1} \log(1/\varepsilon)] \\
&\lesssim N^{-\beta/d} + \gamma n^{-1/2} \log n + n^{-1/2}[\sqrt{\log(1/\varepsilon)} + \gamma^{-1} \log(1/\varepsilon)]. \qquad (10.7)
\end{aligned}$$

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 in the statement of Theorem 8.2, it holds that $\gamma = \mathcal{O}(N^{1/2} \log^{5/2} n)$. Moreover, by the choice $N = n^{d/(2\beta+d)}$, combining (10.7) and taking $\varepsilon = 1/n$, we know that

$$\|\widehat{t} - t^*\|_{L_2(\mathbb{P})} \lesssim n^{-\beta/(2\beta+d)} \log^{7/2} n \tag{10.8}$$

with probability at least $1 - \exp\{-n^{d/(2\beta+d)} \log^5 n\}$.

Step 2. Note that we have

$$\begin{aligned}
&|\widehat{D}_f(q\|p) - D_f(q\|p)| \\
&\quad \le |\mathbb{E}_{\mathbb{Q}_n}[\widehat{t} - t^*] - \mathbb{E}_{\mathbb{Q}}[\widehat{t} - t^*]| + |\mathbb{E}_{\mathbb{P}_n}[f^\dagger(\widehat{t}) - f^\dagger(t^*)] - \mathbb{E}_{\mathbb{P}}[f^\dagger(\widehat{t}) - f^\dagger(t^*)]| \\
&\qquad + |\mathbb{E}_{\mathbb{Q}}[\widehat{t} - t^*] - \mathbb{E}_{\mathbb{P}}[f^\dagger(\widehat{t}) - f^\dagger(t^*)]| + |\mathbb{E}_{\mathbb{Q}_n}[t^*] - \mathbb{E}_{\mathbb{Q}}[t^*]| + |\mathbb{E}_{\mathbb{P}_n}[f^\dagger(t^*)] - \mathbb{E}_{\mathbb{P}}[f^\dagger(t^*)]| \\
&\quad = B_1 + B_2 + B_3 + B_4 + B_5. \qquad (10.9)
\end{aligned}$$

We upper bound B_1, B_2, B_3, B_4, and B_5 in the sequel. First, by Lemma 10.6, with probability at least $1 - \exp\{-n^{d/(2\beta+d)} \log^5 n\}$, we have

$$B_1 \lesssim n^{-2\beta/(2\beta+d)} \log^{7/2} n. \tag{10.10}$$

Similar upper bound also holds for B_2. Also, following from (10.8), with probability at least $1 - \exp\{-n^{d/(2\beta+d)} \log^5 n\}$, we have

$$B_3 \lesssim n^{-\beta/(2\beta+d)} \log^{7/2} n. \tag{10.11}$$

Meanwhile, by Hoeffding's inequality, with probability at least $1 - \exp(-n^{d/(2\beta+d)})$, we have

$$B_4 \lesssim n^{-\beta/(2\beta+d)}. \tag{10.12}$$

Similar upper bound also holds for B_5. Now, combining (10.9), (10.10), (10.11), and (10.12), with probability at least $1 - \exp\{-n^{d/(2\beta+d)} \log^5 n\}$, we have

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

We conclude the proof of Theorem 8.2.

10.5 Proof of Theorem 5.2

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

Lemma 10.3. Under the assumptions stated in Theorem 8.4, for any fixed density p, if the sample size n is sufficiently large, it holds that

$$\sup_{q\in\mathcal{Q}} |D_f(q\|p) - D^\natural_f(q\|p)| \lesssim n^{-1/(\gamma_\Phi+2)} \cdot \log n$$

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

We defer the proof to Section §11.3.

Now we turn to the proof of the theorem. We denote by $\widetilde{q}' = \operatorname{argmin}_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p)$, then with probability at least $1 - 1/n$, we have

$$\begin{aligned} D_f(q^\natural\|p) &\le |D_f(q^\natural\|p) - D^\natural_f(q^\natural\|p)| + D^\natural_f(q^\natural\|p) \\ &\le \sup_{q\in\mathcal{Q}} |D_f(q\|p) - D^\natural_f(q\|p)| + D^\natural_f(\widetilde{q}'\|p) \\ &\le \sup_{q\in\mathcal{Q}} |D_f(q\|p) - D^\natural_f(q\|p)| + |D^\natural_f(\widetilde{q}'\|p) - D_f(\widetilde{q}'\|p)| + D_f(\widetilde{q}'\|p) \\ &\lesssim n^{-1/(\gamma_\Phi+2)} \cdot \log n + D_f(\widetilde{q}'\|p). \end{aligned} \tag{10.13}$$

Here in the second inequality we use the optimality of $q^\natural$ over $\widetilde{q}' \in \mathcal{Q}$ to the problem (5.2), while the last inequality uses Lemma 10.3 and Theorem 4.3. Moreover, note that $D_f(\widetilde{q}'\|p) = \min_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p)$, combining (10.13), it holds that with probability at least $1 - 1/n$,

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

This concludes the proof of the theorem.

10.6 Proof of Theorem 8.4

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

Lemma 10.4. Under the assumptions stated in Theorem 8.4, for any fixed density p, if the sample size n is sufficiently large, it holds that

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

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

We defer the proof to Section §11.4.

Now we turn to the proof of the theorem. We denote by $\widetilde{q}' = \operatorname{argmin}_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p)$, then with probability at least $1 - 1/n$, we have

$$\begin{aligned} D_f(\widehat{q}\|p) &\le |D_f(\widehat{q}\|p) - \widehat{D}_f(\widehat{q}\|p)| + \widehat{D}_f(\widehat{q}\|p) \\ &\le \sup_{q\in\mathcal{Q}} |D_f(q\|p) - \widehat{D}_f(q\|p)| + \widehat{D}_f(\widetilde{q}'\|p) \\ &\le \sup_{q\in\mathcal{Q}} |D_f(q\|p) - \widehat{D}_f(q\|p)| + |\widehat{D}_f(\widetilde{q}'\|p) - D_f(\widetilde{q}'\|p)| + D_f(\widetilde{q}'\|p) \\ &\lesssim n^{-\beta/(d+2\beta)} \cdot \log^7 n + D_f(\widetilde{q}'\|p). \end{aligned} \tag{10.14}$$

Here in the second inequality we use the optimality of $\widehat{q}$ over $\widetilde{q}' \in \mathcal{Q}$ to the problem (8.3), while the last inequality uses Lemma 10.4 and Theorem 8.2. Moreover, note that $D_f(\widetilde{q}'\|p) = \min_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p)$, combining (10.14), it holds that with probability at least $1 - 1/n$,

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

This concludes the proof of the theorem.

10.7 Proof of Theorem 8.5

For any real-valued function ϱ, we write $\mathbb{E}_{\mathbb{P}}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{Q}}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{P}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}_n}[\varrho(x)]$, and $\mathbb{E}_{\mathbb{Q}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\varrho(x)]$ for notational convenience.

For any $\widetilde{t} \in \Phi_M(L, k, s)$, we establish the following lemma.

Lemma 10.5. Under the assumptions stated in Theorem 8.5, it holds that

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

Here μ_0 and L_0 are specified in Assumption 3.4.

The proof of Lemma 10.5 is deferred to Section §11.5.

Note that by Lemma 10.5 and the fact that $f^\dagger$ is Lipschitz continuous, we have

$$\begin{aligned} \|\widehat{t}-\widetilde{t}\|^2_{L_2(\mathbb{P})} \lesssim \|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\cdot\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})} + \{\mathbb{E}_{\mathbb{Q}_n}[(\widehat{t}-\widetilde{t})/2] - \mathbb{E}_{\mathbb{Q}}[(\widehat{t}-\widetilde{t})/2]\} \\ - \{\mathbb{E}_{\mathbb{P}_n}[f^\dagger((\widehat{t}+\widetilde{t})/2) - f^\dagger(\widetilde{t})] - \mathbb{E}_{\mathbb{P}}[f^\dagger((\widehat{t}+\widetilde{t})/2) - f^\dagger(\widetilde{t})]\}. \end{aligned} \tag{10.15}$$

Furthermore, to bound the RHS of the above inequality, we establish the following lemma.

Lemma 10.6. 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 under the assumptions stated in Theorem 8.5, 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

$$\mathbb{P}\left\{\sup_{t(\cdot)\in\Phi_M(L,k,s)} \frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t)-\psi(\widetilde{t})] - \mathbb{E}_{\mathbb{P}}[\psi(t)-\psi(\widetilde{t})]|}{\max\{\eta(n,\gamma,\varepsilon)\cdot\|\psi(t)-\psi(\widetilde{t})\|_{L_2(\mathbb{P})}, \lambda(n,\gamma,\varepsilon)\}} \le 16M_0\right\} \ge 1-\varepsilon\cdot\exp(-\gamma^2),$$

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 $V = \prod_{j=0}^{L+1}(k_j+1)$.

We defer the proof to Section §11.6.

Note that the results in Lemma 10.6 also apply to the distribution $\mathbb{Q}$, and by using the fact that the true density ratio $\theta^*(x;p,q) = q(x)/p(x)$ is bounded below and above, we know that $L_2(\mathbb{Q})$ is indeed equivalent to $L_2(\mathbb{P})$. We thus focus on $L_2(\mathbb{P})$ here. By (10.15), Lemma 10.6, and the Lipschitz property of $f^\dagger$ according to Lemma 12.6, with probability at least $1-\varepsilon\cdot\exp(-\gamma^2)$, we have the following bound

$$\begin{aligned} \|\widehat{t}-\widetilde{t}\|^2_{L_2(\mathbb{P})} \lesssim \|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})}\cdot\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})} \\ + \mathcal{O}(n^{-1/2}[\gamma\log n + \gamma^{-1}\log(1/\varepsilon)]\cdot\|\widehat{t}-\widetilde{t}\|_{L_2(\mathbb{P})} \vee n^{-1}[\gamma^2+\log(1/\varepsilon)]), \end{aligned} \tag{10.16}$$

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 (10.16).

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

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

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

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

Therefore, by combining the above three cases, we have

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

Further combining the triangle inequality, we have

$$\|\widehat{t}-t^*\|_{L_2(\mathbb{P})} \lesssim \|\widetilde{t}-t^*\|_{L_2(\mathbb{P})} + \gamma n^{-1/2}\log n + n^{-1/2}[\sqrt{\log(1/\varepsilon)} + \gamma^{-1}\log(1/\varepsilon)] \tag{10.17}$$

with probability at least $1-\varepsilon\cdot\exp(-\gamma^2)$. Note that (10.17) holds for any $\widetilde{t}\in\Phi_M(L,k,s)$, especially for the choice $\widetilde{t}$ which minimizes $\|\widetilde{t}-t^*\|_{L_2(\mathbb{P})}$. Therefore, we have

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

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

10.8 Proof of Theorem 8.7

We follow the proof in Li et al. (2018). We denote by the loss function in (8.5) as $\mathcal{L}[t(x)] = f^\dagger(t(x^{\mathrm{I}})) - t(x^{\mathrm{II}})$, where x^{I} follows the distribution $\mathbb{P}$ and x^{II} follows $\mathbb{Q}$. To prove the theorem, we first link the generalization error in our theorem to the empirical Rademacher complexity (ERC). Given the data $\{x_i\}_{i=1}^n$, the ERC related with the class $\mathcal{L}(\Phi_{\mathrm{norm}})$ is defined as

$$\mathfrak{R}_n[\mathcal{L}(\Phi_{\mathrm{norm}})] = \mathbb{E}_\varepsilon\Biggl[\sup_{\varphi\in\Phi_{\mathrm{norm}}}\Bigl|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\cdot\mathcal{L}[\varphi(x_i;W,v)]\Bigr|\;\Bigg|\;\{x_i\}_{i=1}^n\Biggr], \tag{10.18}$$

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 $\mathbb{E}_\varepsilon(\cdot)$ is taken over the Rademacher random variables $\{\varepsilon_i\}_{i\in[n]}$.

We introduce the following lemma, which links the ERC to the generalization error bound.

Lemma 10.7 ((Mohri et al., 2018))**.** Assume that $\sup_{\varphi\in\Phi_{\mathrm{norm}}}|\mathcal{L}(\varphi)|\le M_1$, then for any $\varepsilon>0$, with probability at least $1-\varepsilon$, we have

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

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

Equipped with Lemma 10.7, we only need to bound the ERC defined in (10.18).

Lemma 10.8. Let $\mathcal{L}$ be a Lipschitz continuous loss function and Φ_{norm} be the family of networks defined in (8.4). We assume that the input $x\in\mathbb{R}^d$ is bounded such that $\|x\|_2\le B$. Then it holds that

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

where γ_1 and γ_2 are given in (8.6).

We defer the proof to Section §11.7.

Now we proceed to prove the theorem. Recall that we assume that $t^*\in\Phi_{\mathrm{norm}}$. For notational convenience, we denote by

$$\widehat{H}(t) = \mathbb{E}_{x\sim\mathbb{P}_n}[f^\dagger(t(x))] - \mathbb{E}_{x\sim\mathbb{Q}_n}[t(x)], \qquad H(t) = \mathbb{E}_{x\sim\mathbb{P}}[f^\dagger(t(x))] - \mathbb{E}_{x\sim\mathbb{Q}}[t(x)].$$

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 $\widehat{H}(\widehat{t})\ge H(t^*)$, then we have

$$0 \le \widehat{H}(\widehat{t}) - H(t^*) \le \widehat{H}(t^*) - H(t^*), \tag{10.19}$$

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

$$0 \ge \widehat{H}(\widehat{t}) - H(t^*) \ge \widehat{H}(\widehat{t}) - H(\widehat{t}), \tag{10.20}$$

where the second inequality follows that fact that t^* is the minimizer of $H(\cdot)$. Therefore, by (10.19), (10.20), and the fact that $\mathcal{L}(\varphi) \lesssim \prod_{j=1}^{L+1} B_j$ for any $\varphi \in \Phi_{\rm norm}$, we deduce that

$$|\widehat{H}(\widehat{t}) - H(t^*)| \le \sup_{t \in \Phi_{\rm norm}} |\widehat{H}(t) - H(t)| \lesssim \mathfrak{R}_n[\mathcal{L}(\Phi_{\rm norm})] + \prod_{j=1}^{L+1} B_j \cdot n^{-1/2}\sqrt{\log(1/\varepsilon)} \tag{10.21}$$

with probability at least $1-\varepsilon$. Here the second inequality follows from Lemma 10.7. By plugging the result from Lemma 10.8 into (10.21), we deduce that with probability at least $1-\varepsilon$, it holds that

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

This concludes the proof of the theorem.

10.9 Proof of Theorem 8.8

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 10.9. 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

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

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.

We defer the proof to Section §11.8.

Now we turn to the proof of the theorem. We denote by $\widetilde{q}' = \operatorname{argmin}_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p)$. Then with probability at least $1-\varepsilon$, we have

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

where we use the optimality of $\widehat{q}$ among all $\widetilde{q} \in \mathcal{Q}$ to the problem (8.3) in the second inequality, and we uses Lemma 10.9 and Theorem 8.2 in the last line. Moreover, note that $D_f(\widetilde{q}'\|p) = \min_{\widetilde{q}\in\mathcal{Q}} D_f(\widetilde{q}\|p)$, we obtain that

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

This concludes the proof of the theorem.

11 Lemmas and Proofs

11.1 Proof of Lemma 10.1

For any real-valued function ϱ, we write $\mathbb{E}_{\mathbb{P}}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{Q}}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{P}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}_n}[\varrho(x)]$, and $\mathbb{E}_{\mathbb{Q}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\varrho(x)]$ for notational convenience.

By the definition of $t^\natural$ in (4.2), we have

$$\mathbb{E}_{\mathbb{P}_n}[f^\dagger(t^\natural)] - \mathbb{E}_{\mathbb{Q}_n}(t^\natural) \leq \mathbb{E}_{\mathbb{P}_n}[f^\dagger(t^*)] - \mathbb{E}_{\mathbb{Q}_n}(t^*).$$

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

$$G(\frac{t^\natural + t^*}{2}) - G(t^*) \leq \frac{G(t^\natural) - G(t^*)}{2} \leq 0.$$

By re-arranging terms, we have

$$\begin{aligned}\{\mathbb{E}_{\mathbb{P}_n}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)] - \mathbb{E}_{\mathbb{P}}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)]\} - \{\mathbb{E}_{\mathbb{Q}_n}[(t^\natural - t^*)/2] - \mathbb{E}_{\mathbb{Q}}[(t^\natural - t^*)/2]\} \\ \leq \mathbb{E}_{\mathbb{Q}}[(t^\natural - t^*)/2] - \mathbb{E}_{\mathbb{P}}[f^\dagger((t^\natural + t^*)/2) - f^\dagger(t^*)]. \qquad (11.1)\end{aligned}$$

We denote by

$$B_f(t^*, t) = \mathbb{E}_{\mathbb{P}}[f^\dagger(t) - f^\dagger(t^*)] - \mathbb{E}_{\mathbb{Q}}(t - t^*). \qquad (11.2)$$

then the RHS of (11.1) is exactly $-B_f(t^*, (t^\natural + t^*)/2)$. We proceed to establish the lower bound of $B_f(t^*, t)$ using $L_2(\mathbb{P})$ norm. From $t^*(x; p, q) = f'(q(x)/p(x))$ and $(f^\dagger)' \circ (f')(x) = x$, we know that $q/p = \partial f^\dagger(t^*)/\partial t$. Then by substituting the second term on the RHS of (11.2) using the above relationship, we have

$$B_f(t^*, t) = \mathbb{E}_{\mathbb{P}}\left[f^\dagger(t) - f^\dagger(t^*) - \frac{\partial f^\dagger}{\partial t}(t^*) \cdot (t - t^*)\right]$$

Note that by Assumption 3.4 and Lemma 12.6, we know that the Fenchel duality $f^\dagger$ is strongly convex with parameter $1/L_0$. This gives that

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

for any x. Consequently, it holds that

$$B_f(t^*, t) \geq 1/L_0 \cdot \|t - t^*\|^2_{L_2(\mathbb{P})}. \qquad (11.3)$$

By (11.3), we conclude that

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

This concludes the proof of the lemma.

11.2 Proof of Lemma 10.2

For any real-valued function ϱ, we write $\mathbb{E}_{\mathbb{P}}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{Q}}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{P}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}_n}[\varrho(x)]$, and $\mathbb{E}_{\mathbb{Q}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\varrho(x)]$ for notational convenience.

We first introduce the following concepts. For any $K > 0$, the Bernstein difference $\rho^2_{K,\mathbb{P}}(t)$ of $t(\cdot)$ with respect to the distribution $\mathbb{P}$ is defined to be

$$\rho^2_{K,\mathbb{P}}(t) = 2K^2 \cdot \mathbb{E}_{\mathbb{P}}[\exp(|t|/K) - 1 - |t|/K].$$

Correspondingly, we denote by $\mathcal{H}_{K,B}$ the generalized entropy with bracketing induced by the 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

$$\Psi = \psi(\Phi) = \{\psi(t) : t(x) \in \Phi\}.$$

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

$$\Psi(\delta) = \{\psi(t) \in \Psi : \|\psi(t) - \psi(t^*)\|_{L_2(\mathbb{P})} \le \delta\},$$
$$\Psi'(\delta) = \{\Delta\psi(t) = \psi(t) - \psi(t^*) : \psi(t) \in \Psi(\delta)\}.$$

Note that $\sup_{\Delta\psi(t)\in\Psi'(\delta)} \|\Delta\psi(t)\|_\infty \le 2M_0$ and $\sup_{\Delta\psi(t)\in\Psi'(\delta)} \|\Delta\psi(t)\|_\infty \le \delta$, by Lemma 12.4 we have

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

To invoke Theorem 12.3 for $\mathcal{G} = \Psi'(\delta)$, we pick $K = 8M_0$. By the fact that $\sup_{\Delta\psi(t)\in\Psi'(\delta)} \|\Delta\psi(t)\|_\infty \le 2M_0$, Lemma 12.1, Assumption 4.2, and the fact that ψ is Lipschitz continuous, we have

$$\mathcal{H}_{8M_0,B}(u, \Psi'(\delta), \mathbb{P}) \le H_{2,B}(\sqrt{2}u, \Psi'(\delta), \mathbb{P}) \le u^{-\gamma_\Phi}$$

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

$$\int_0^R \mathcal{H}^{1/2}_{8M_0,B}(u, \Psi'(\delta), \mathbb{P})\mathrm{d}u \le \frac{2}{2-\gamma_\Phi} R^{-\gamma_\Phi/2+1}.$$

We take $C = 1$, and a, C_1 and C_0 in Theorem 12.3 to be

$$a = C_1\sqrt{n}R^2/K, \qquad C_0 = 2C^2C_2 \vee 2C, \qquad C_1 = C_0C_2,$$

where C_2 is a sufficiently large constant. Then it is straightforward to check that our choice above satisfies the conditions in Theorem 12.3 for any δ such that $\delta \ge n^{-1/(\gamma_\Phi+2)}$, when n is sufficiently large. With $\delta_n = n^{-1/(\gamma_\Phi+2)}$, we have

$$\begin{aligned}
&\mathbb{P}\biggl\{\sup_{t:\,\psi(t)\in\Psi,\psi(t)\notin\Psi(\delta_n)} \frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t)-\psi(t^*)] - \mathbb{E}_{\mathbb{P}}[\psi(t)-\psi(t^*)]|}{n^{-2/(\gamma_\Phi+2)}} \ge C_1/K\biggr\} \\
&\quad\le \mathbb{P}\biggl\{\sup_{t:\,\psi(t)\in\Psi,\psi(t)\notin\Psi(\delta_n)} \frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t)-\psi(t^*)] - \mathbb{E}_{\mathbb{P}}[\psi(t)-\psi(t^*)]|}{\|\psi(t)-\psi(t^*)\|^2_{L_2(\mathbb{P})}} \ge C_1/K\biggr\} \\
&\quad\le \sum_{s=0}^{S} \mathbb{P}\biggl\{\sup_{t:\,\psi(t)\in\Psi,\psi(t)\in\Psi(2^{s+1}\delta_n)} |\mathbb{E}_{\mathbb{P}_n}[\psi(t)-\psi(t^*)] - \mathbb{E}_{\mathbb{P}}[\psi(t)-\psi(t^*)]| \ge C_1/K \cdot (2^s\delta_n)^2\biggr\} \\
&\quad\le \sum_{s=0}^{S} C\exp\biggl(-\frac{C_1^2/K^2 \cdot 2^{2s} \cdot n^{\gamma_\Phi/(2+\gamma_\Phi)}}{C^2(C_1+1)}\biggr) \le c_1\exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2),
\end{aligned}$$

for some constant $c_1 > 0$. Here in the last line, we invoke Theorem 12.3 with $R = 2^s\delta_n$. Therefore, we have

$$\mathbb{P}\biggl\{\sup_{t:\,\psi(t)\in\Psi} \frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t)-\psi(t^*)] - \mathbb{E}_{\mathbb{P}}[\psi(t)-\psi(t^*)]|}{n^{-2/(\gamma_\Phi+2)}} \ge C_1/K\biggr\} \le c_1\exp(-n^{\gamma_\Phi/(2+\gamma_\Phi)}/c_1^2).$$

We conclude the proof of Lemma 10.2.

11.3 Proof of Lemma 10.3

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 $\delta = \delta_n = n^{-1/(\gamma_\Phi+2)}$ and union bound, we have

$$\begin{aligned}
&\mathbb{P}[\sup_{q\in\mathcal{Q}} |D_f(q\|p) - D^\natural_f(q\|p)| \ge c_1 \cdot n^{-1/(\gamma_\Phi+2)} \cdot \log n] \\
&\qquad\le \sum_{k=1}^{N_2(\delta_n,\mathcal{Q})} \mathbb{P}[|D_f(q_k\|p) - D^\natural_f(q_k\|p)| \ge c_1 \cdot n^{-1/(\gamma_\Phi+2)} \cdot \log n] \\
&\qquad\le N_2(\delta_n, \mathcal{Q}) \cdot \exp(-n^{\gamma_\Phi/(\gamma_\Phi+2)} \cdot \log n),
\end{aligned}$$

where the last line comes from Theorem 4.3. Combining Assumption 5.1, when n is sufficiently large, it holds that

$$\mathbb{P}[\sup_{q\in\mathcal{Q}} |D_f(q\|p) - D^\natural_f(q\|p)| \ge c_1 \cdot n^{-1/(\gamma_\Phi+2)} \cdot \log n] \le 1/n,$$

which concludes the proof of the lemma.

11.4 Proof of Lemma 10.4

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 $\delta = \delta_n = n^{-\beta/(d+2\beta)}$ and union bound, we have

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

where the last line comes from Theorem 8.2. Combining Assumption 8.3, when n is sufficiently large, it holds that

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

which concludes the proof of the lemma.

11.5 Proof of Lemma 10.5

For any real-valued function ϱ, we write $\mathbb{E}_{\mathbb{P}}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{Q}}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{P}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}_n}[\varrho(x)]$, and $\mathbb{E}_{\mathbb{Q}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\varrho(x)]$ for notational convenience.

By the definition of $\widehat{t}$ in (8.2), we have

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

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

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

By re-arranging terms, we have

$$\begin{aligned}
&\{\mathbb{E}_{\mathbb{P}_n}[f^\dagger((\widehat{t}+\widetilde{t})/2) - f^\dagger(\widetilde{t})] - \mathbb{E}_{\mathbb{P}}[f^\dagger((\widehat{t}+\widetilde{t})/2) - f^\dagger(\widetilde{t})]\} - \{\mathbb{E}_{\mathbb{Q}_n}[(\widehat{t}-\widetilde{t})/2] - \mathbb{E}_{\mathbb{Q}}[(\widehat{t}-\widetilde{t})/2]\} \\
&\quad\le \mathbb{E}_{\mathbb{Q}}[(\widehat{t}-\widetilde{t})/2] - \mathbb{E}_{\mathbb{P}}[f^\dagger((\widehat{t}+\widetilde{t})/2) - f^\dagger(\widetilde{t})]. \qquad (11.4)
\end{aligned}$$

We denote by

$$B_f(\widetilde{t}, t) = \mathbb{E}_{\mathbb{P}}[f^\dagger(t) - f^\dagger(\widetilde{t})] - \mathbb{E}_{\mathbb{Q}}(t - \widetilde{t}). \qquad (11.5)$$

then the RHS of (11.4) 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. From $t^*(x; p, q) = f'(q(x)/p(x))$ and $(f^\dagger)' \circ (f')(x) = x$, we know that $q/p = \partial f^\dagger(t^*)/\partial t$. Then by substituting the second term on the RHS of (11.5) using the above relationship, we have

$$\begin{aligned}
B_f(\widetilde{t}, t) &= \mathbb{E}_{\mathbb{P}}\biggl[f^\dagger(t) - f^\dagger(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(t^*) \cdot (t - \widetilde{t})\biggr] \\
&= \mathbb{E}_{\mathbb{P}}\biggl[f^\dagger(t) - f^\dagger(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(\widetilde{t}) \cdot (t - \widetilde{t})\biggr] + \mathbb{E}_{\mathbb{P}}\biggl\{\biggl[\frac{\partial f^\dagger}{\partial t}(\widetilde{t}) - \frac{\partial f^\dagger}{\partial t}(t^*)\biggr] \cdot (t - \widetilde{t})\biggr\} \\
&= A_1 + A_2. \qquad (11.6)
\end{aligned}$$

We lower bound A_1 and A_2 in the sequel.

Bound on A_1. Note that by Assumption 3.4 and Lemma 12.6, we know that the Fenchel duality $f^\dagger$ is strongly convex with parameter $1/L_0$. This gives that

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

for any x. Consequently, it holds that

$$A_1 \geq 1/L_0 \cdot \|t - \widetilde{t}\|^2_{L_2(\mathbb{P})}. \tag{11.7}$$

Bound on A_2. By Cauchy-Schwarz inequality, it holds that

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

Again, by Assumption 3.4 and Lemma 12.6, we know that the Fenchel duality $f^\dagger$ has $1/\mu_0$-Lipschitz gradient, which gives that

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

for any x. By this, the term A_2 is lower bounded:

$$A_2 \geq -1/\mu_0 \cdot \|\widetilde{t} - t^*\|_{L_2(\mathbb{P})} \cdot \|t - \widetilde{t}\|_{L_2(\mathbb{P})}. \tag{11.8}$$

Plugging (11.7) and (11.8) into (11.6), we have

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

By this, together with (11.4), we conclude that

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

This concludes the proof of the lemma.

11.6 Proof of Lemma 10.6

For any real-valued function ϱ, we write $\mathbb{E}_{\mathbb{P}}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{Q}}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}}[\varrho(x)]$, $\mathbb{E}_{\mathbb{P}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{P}_n}[\varrho(x)]$, and $\mathbb{E}_{\mathbb{Q}_n}(\varrho) = \mathbb{E}_{x\sim\mathbb{Q}_n}[\varrho(x)]$ for notational convenience.

We first introduce the following concepts. For any $K > 0$, the Bernstein difference $\rho^2_{K,\mathbb{P}}(t)$ of $t(\cdot)$ with respect to the distribution $\mathbb{P}$ is defined to be

$$\rho^2_{K,\mathbb{P}}(t) = 2K^2 \cdot \mathbb{E}_{\mathbb{P}}[\exp(|t|/K) - 1 - |t|/K].$$

Correspondingly, we denote by $\mathcal{H}_{K,B}$ the generalized entropy with bracketing induced by the 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.

Since we focus on fixed L, k, and s, we denote by $\Phi_M = \Phi_M(L, k, s)$ for notational convenience. We consider the space

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

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

$$\begin{aligned} \Psi_M(\delta) &= \{\psi(t) \in \Psi_M : \|\psi(t) - \psi(\widetilde{t})\|_{L_2(\mathbb{P})} \leq \delta\}, \\ \Psi'_M(\delta) &= \{\Delta\psi(t) = \psi(t) - \psi(\widetilde{t}) : \psi(t) \in \Psi_M(\delta)\}. \end{aligned}$$

Note that $\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \|\Delta\psi(t)\|_\infty \leq 2M_0$ and $\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \|\Delta\psi(t)\|_\infty \leq \delta$, by Lemma 12.4 we have

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

To invoke Theorem 12.3 for $\mathcal{G} = \Psi'_M(\delta)$, we pick $K = 8M_0$ and $R = \sqrt{2}\delta$. Note that from the fact that $\sup_{\Delta\psi(t)\in\Psi'_M(\delta)} \|\Delta\psi(t)\|_\infty \leq 2M_0$, by Lemma 12.1, Lemma 12.2, and the fact that ψ is Lipschitz continuous, we have

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

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

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

For any $0 < \varepsilon < 1$, we take $C = 1$, and a, C_1 and C_0 in Theorem 12.3 to be

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

Here $\gamma = s^{1/2}\log(V^2L)$. Then it is straightforward to check that our choice above satisfies the conditions in Theorem 12.3 for any δ such that $\delta \geq \gamma n^{-1/2}$, when n is sufficiently large such that $n \gtrsim [\gamma + \gamma^{-1}\log(1/\varepsilon)]^2$. Consequently, by Theorem 12.3, for $\delta \geq \gamma n^{-1/2}$, we have

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

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

$$\mathbb{P}\bigg\{\sup_{t(x)\in\Phi_M(\delta)} \frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t) - \psi(\widetilde{t})] - \mathbb{E}_{\mathbb{P}}[\psi(t) - \psi(\widetilde{t})]|}{n^{-1}[\gamma^2 + \log(1/\varepsilon)]} \leq 8M_0\bigg\} \geq 1 - \varepsilon\cdot\exp(-\gamma^2). \tag{11.9}$$

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 convenience, we denote the set

$$A_s = \{\psi(t) \in \Psi_M : \psi(t) \in \Psi_M(2^{-s+2}M_0), \psi(t) \notin \Psi_M(2^{-s+1}M_0)\}. \tag{11.10}$$

Then by the peeling device, we have the following

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

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 (11.10); 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 invoke Theorem 12.3. Consequently, this gives us

$$\mathbb{P}\bigg\{\sup_{\substack{\psi(t)\in\Psi_M \\ \psi(t)\notin\Psi_M(\delta_n)}} \frac{|\mathbb{E}_{\mathbb{P}_n}[\psi(t) - \psi(\widetilde{t})] - \mathbb{E}_{\mathbb{P}}[\psi(t) - \psi(\widetilde{t})]|}{\|\psi(t) - \psi(\widetilde{t})\|_{L_2(\mathbb{P})}\cdot n^{-1/2}[\gamma\log n + \gamma^{-1}\log(1/\varepsilon)]} \leq 16M_0\bigg\} \geq 1 - \varepsilon\cdot\exp(-\gamma^2). \tag{11.11}$$

Combining (11.9) and (11.11), we finish the proof of the lemma.

11.7 Proof of Lemma 10.8

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 11.1. 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

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

We defer the proof to Section §11.9.

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

Lemma 11.2. 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

$$\begin{aligned}&\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\| \\ &\quad \leq \frac{B\sqrt{2L+1} \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \sum_{j=1}^{L} A_j \cdot \sqrt{\sum_{j=1}^{L+1} \|W_j^1 - W_j^2\|_{\text{F}}^2 + \sum_{j=1}^{L} \|v_j^1 - v_j^2\|_2^2}.\end{aligned}$$

We defer the proof to Section §11.10.

We now turn to the proof of Lemma 10.8. Note that by Lemma 11.2, 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

$$b = \sum_{j=1}^{L+1} k_j k_{j-1} + \sum_{j=1}^{L} k_j \leq \sum_{j=0}^{L+1} (k_j + 1)^2, \tag{11.12}$$

and the Lipschitz constant L_w satisfies

$$L_w = \frac{B\sqrt{2L+1} \cdot \prod_{j=1}^{L+1} B_j}{\min_j B_j} \cdot \sum_{j=1}^{L} A_j. \tag{11.13}$$

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

$$K = \sqrt{\sum_{j=1}^{L+1} k_j^2 B_j^2} + \sum_{j=1}^{L} A_j, \tag{11.14}$$

satisfies

$$N(\mathcal{W}, \delta) \leq (3K\delta^{-1})^b.$$

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

$$N[\mathcal{L}(\Phi_{\text{norm}}), \delta] \leq (c_1 K L_w \delta^{-1})^b,$$

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

$$\mathfrak{R}_n[\mathcal{L}(\Phi_{\text{norm}})] \leq \inf_{\tau > 0} \tau + \frac{1}{\sqrt{n}} \int_{\tau}^{\vartheta} \sqrt{\log N[\mathcal{L}(\Phi_{\text{norm}}), \delta]} \mathrm{d}\delta, \tag{11.15}$$

where $\vartheta = \sup_{g(\cdot; W, v) \in \mathcal{L}(\Phi_{\text{norm}}), x \in \mathbb{R}^d} |g(x; W, v)|$. Moreover, from Lemma 11.1 and the fact that the loss function is Lipschitz continuous, we have

$$\vartheta \leq c_2 \cdot B \cdot \prod_{j=1}^{L+1} B_j \tag{11.16}$$

for some positive absolute constant c_2. Therefore, by calculations, we derive from (11.15) that

$$\mathfrak{R}_n[\mathcal{L}(\Phi_{\rm norm})] = \mathcal{O}\biggl(\frac{\vartheta}{\sqrt{n}} \cdot \sqrt{b \cdot \log \frac{KL_w\sqrt{n}}{\vartheta\sqrt{b}}}\biggr),$$

then we conclude the proof of the lemma by plugging in (11.12), (11.13), (11.14), and (11.16), and using the definition of γ_1 and γ_2 in (8.6).

11.8 Proof of Lemma 10.9

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 $1 \le k \le N_2(\delta, \mathcal{Q})$, so that $\|q - q_k\|_2 \le \delta$. 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

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

where the second line comes from union bound, and the last line comes from Theorem 8.7. By this, we conclude the proof of the lemma.

11.9 Proof of Lemma 11.1

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

$$\begin{aligned}
\|\varphi(x_1; W, v) - \varphi(x_2; W, v)\|_2 &= \|W_{L+1}(\sigma_{v_L} \cdots W_2\sigma_{v_1}W_1x_1 - \sigma_{v_L} \cdots W_2\sigma_{v_1}W_1x_2)\|_2 \\
&\le \|W_{L+1}\|_2 \cdot \|\sigma_{v_L}(W_L \cdots W_2\sigma_{v_1}W_1x_1 - W_L \cdots W_2\sigma_{v_1}W_1x_2)\|_2 \\
&\le B_{L+1} \cdot \|W_L \cdots W_2\sigma_{v_1}W_1x_1 - W_L \cdots W_2\sigma_{v_1}W_1x_2\|_2 \\
&\le \cdots \le \prod_{j=1}^{L+1} B_j \cdot \|x_1 - x_2\|_2.
\end{aligned}$$

Here in the third line we uses the fact that $\|W_j\|_2 \le B_j$ 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 of the lemma.

11.10 Proof of Lemma 11.2

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

$$\varphi(x; W, v) = W_{L+1}\sigma_{v_L}W_L \cdots \sigma_{v_1}W_1x.$$

For notational convenience, we denote by $\varphi_j^i(x) = \sigma_{v_j^i}(W_j^i x)$ 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

$$\begin{aligned}
\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\|_2 &= \|W_{L+1}^1\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - W_{L+1}^2\varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\|_2 \\
&\le \|W_{L+1}^1\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - W_{L+1}^2\varphi_L^1 \circ \cdots \circ \varphi_1^1(x)\|_2 \\
&\qquad + \|W_{L+1}^2\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - W_{L+1}^2\varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\|_2 \\
&\le \|W_{L+1}^1 - W_{L+1}^2\|_{\rm F} \cdot \|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x)\|_2 \\
&\qquad + B_{L+1} \cdot \|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\|_2. \qquad (11.17)
\end{aligned}$$

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

$$\begin{aligned}\|\varphi_\ell^i \circ \cdots \circ \varphi_1^i(x)\|_2 &\le \|W_\ell^i \varphi_{\ell-1}^i \circ \cdots \circ \varphi_1^i(x)\|_2 + \|v_\ell^i\|_2 \\ &\le B_\ell \cdot \|\varphi_{\ell-1}^i \circ \cdots \circ \varphi_1^i(x)\|_2 + A_\ell \\ &\le \|x\|_2 \cdot \prod_{j=1}^{\ell} B_j + \sum_{j=1}^{\ell} A_j \prod_{i=j+1}^{\ell} B_i, \end{aligned} \tag{11.18}$$

where the first inequality comes from the triangle inequality, and the second inequality comes from the bounded spectral norm of W_j^i, while the last inequality simply applies the previous arguments recursively. Therefore, combining (11.17), we have

$$\begin{aligned}\|\varphi(x, W^1, v^1) - \varphi(x, W^2, v^2)\|_2 \le \bigg(B \cdot \prod_{j=1}^{L} B_j + \sum_{j=1}^{L} A_j \prod_{i=j+1}^{L} B_i\bigg) \cdot \|W_{L+1}^1 - W_{L+1}^2\|_F \\ + B_{L+1} \cdot \|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\|_2. \end{aligned} \tag{11.19}$$

Similarly, by triangular inequality, we have

$$\begin{aligned}&\|\varphi_L^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \cdots \circ \varphi_1^2(x)\|_2 \\ &\quad \le \|\varphi_L^1 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x)\|_2 \\ &\qquad\quad + \|\varphi_L^2 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \varphi_{L-1}^2 \circ \cdots \circ \varphi_1^2(x)\|_2 \\ &\quad \le \|\varphi_L^1 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_L^2 \circ \varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x)\|_2 \\ &\qquad\quad + B_L \cdot \|\varphi_{L-1}^1 \circ \cdots \circ \varphi_1^1(x) - \varphi_{L-1}^2 \circ \cdots \circ \varphi_1^2(x)\|_2, \end{aligned} \tag{11.20}$$

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

$$\begin{aligned}\|\varphi_L^1(y) - \varphi_L^2(y)\|_2 &= \|\sigma(W_L^1 y - v_L^1) - \sigma(W_L^2 y - v_L^2)\}\|_2 \\ &\le \|v_L^1 - v_L^2\|_2 + \|W_L^1 - W_L^2\|_{\mathrm{F}} \cdot \|y\|_2, \end{aligned}$$

where the inequality comes from the 1-Lipschitz property of $\sigma(\cdot)$. Moreover, combining (11.18), it holds that

$$\|\varphi_L^1(y) - \varphi_L^2(y)\|_2 \le \|v_L^1 - v_L^2\|_2 + \|W_L^1 - W_L^2\|_{\mathrm{F}} \cdot \bigg(B \cdot \prod_{j=1}^{L-1} B_j + \sum_{j=1}^{L-1} A_j \prod_{i=j+1}^{L-1} B_i\bigg). \tag{11.21}$$

By (11.20) and (11.21), we have

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

Here in the second inequality we recursively apply the previous arguments. Further combining (11.19), we obtain that

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

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

12 Auxiliary Results

Lemma 12.1. 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 $1 \le q < \infty$, and $\mathbb{Q}$ a distribution, 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 van de Geer and van de Geer (2000) for a detailed proof. □

Lemma 12.2. The entropy of the neural network set defined in (8.1) satisfies

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

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

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

Theorem 12.3. Assume that $\sup_{g \in \mathcal{G}} \rho_K(g) \le R$. Take a, C, C_0, and 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 $C_0^2 \ge C^2(C_1 + 1)$. It holds that

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

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

Lemma 12.4. 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 and van de Geer (2000) for a detailed proof. □

Theorem 12.5. 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 12.6. 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. □

13 Experiment details

To evaluate the performance of our mechanism on the MNIST and CIFAR-10 test dataset, we first observe that for high-dimensional data, the optimization task in step 1 may fail to converge to the global (or a high-quality local) optimum. Adopting a fixed form of $\widehat{t}$ can still guarantee incentive properties of our mechanism and also consumes less time. Thus, we skip Step 1 in Algorithms 1, 2, and instead adopt $\widehat{t}$ from the existing literature.

13.1 Evaluation with ground-truth verification

To estimate distributions w.r.t. images, we borrow a practical trick as implemented in Nowozin et al. (2016): let's denote a public discriminator as D which has been pre-trained on corresponding training dataset. Given a batch of clean (ground-truth) images $\{x_i\}_{i=1}^n$, agent **A**'s corresponding untruthful reports $\{\widetilde{x}_i\}_{i=1}^n$, the score of **A**'s reports is calculated by:

$$S(\{\widetilde{x}_i\}_{i=1}^n, \{x_i\}_{i=1}^n) = a - \frac{b}{n} \cdot \sum_{i=1}^n \Big[\widehat{t}(D(x_i)) - f^\dagger(\widehat{t}(D(\widetilde{x}_i)))\Big]$$

13.2 Evaluation without ground-truth verification

Suppose we have access to a batch of peer reported images $\{\bar{x}_i\}_{i=1}^n$, agent **A**'s corresponding untruthful reports $\{\widetilde{x}_i\}_{i=1}^n$. For $\mathbb{P}_n = \{\widetilde{x}_i\}_{i=1}^n$, $\mathbb{Q}_n = \{\bar{x}_i\}_{i=1}^n$, we use $(D(\widetilde{x}_i) + D(\bar{x}_i))/2$ to estimate the distribution $x \sim \mathbb{P} \oplus \mathbb{Q}$, and $D(\widetilde{x}_i) \cdot D(\bar{x}_i)$ is the estimation of $x \sim \mathbb{P} \times \mathbb{Q}$. The score of **A**'s reports is calculated by:

$$S(\{\widetilde{x}_i\}_{i=1}^n, \{\bar{x}_i\}_{i=1}^n) = a + \frac{b}{n} \cdot \sum_{i=1}^n \Big[\widehat{t}\Big(\frac{D(\widetilde{x}_i) + D(\bar{x}_i)}{2}\Big) - f^\dagger(\widehat{t}(D(\widetilde{x}_i) \cdot D(\bar{x}_i)))\Big]$$

13.3 Computing infrastructure

In our experiments, we use a GPU cluster (8 TITAN V GPUs and 16 GeForce GTX 1080 GPUs) for training and evaluation.