Supplement for "Generalization Bounds for Domain Adaptation"

A   Relationship between $D_{\mathcal{F}}(S,T)$ and Other Quantities

By (5), the quantity $D_{\mathcal{F}}(S,T)$ can be equivalently rewritten as

$$D_{\mathcal{F}}(S,T) = \sup_{g \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell(g(\mathbf{x}^{(S)}), \mathbf{y}^{(S)}) - \mathrm{E}^{(T)} \ell(g(\mathbf{x}^{(T)}), \mathbf{y}^{(T)}) \right|$$

$$= \sup_{g \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell\big(g(\mathbf{x}^{(S)}), g_*^{(S)}(\mathbf{x}^{(S)})\big) - \mathrm{E}^{(T)} \ell\big(g(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})\big) \right|. \qquad (22)$$

Next, based on the equivalent form (22), we discuss the relationship between the quantity $D_{\mathcal{F}}(S,T)$ and other quantities including the $\mathcal{H}$-*divergence* and the *discrepancy distance* proposed in [13] and [20], respectively.

A.1 $\mathcal{H}$-Divergence and Discrepancy Distance

In classification tasks, setting ℓ as the absolute-value loss function ($\ell(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|$), Ben-David *et al.* [13] introduced a variant of the $\mathcal{H}$-*divergence*:

$$d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}^{(S)}, \mathcal{D}^{(T)}) = \sup_{g_1, g_2 \in \mathcal{H}} \left| \mathrm{E}^{(S)} \ell\big(g_1(\mathbf{x}^{(S)}), g_2(\mathbf{x}^{(S)})\big) - \mathrm{E}^{(T)} \ell\big(g_1(\mathbf{x}^{(T)}), g_2(\mathbf{x}^{(T)})\big) \right| \quad (23)$$

to achieve VC-dimension-based generalization bounds for domain adaptation under the condition of "λ-close": there exists a $\lambda > 0$ such that

$$\lambda \geq \inf_{g \in \mathcal{G}} \left\{ \int \ell(g(\mathbf{x}^{(S)}), \mathbf{y}^{(S)}) d\mathrm{P}(\mathbf{z}^{(S)}) + \int \ell(g(\mathbf{x}^{(T)}), \mathbf{y}^{(T)}) d\mathrm{P}(\mathbf{z}^{(T)}) \right\}.$$

In both classification and regression tasks, given a function class $\mathcal{G}$ and a loss function ℓ, Mansour *et al.* [20] defined the *discrepancy distance* as

$$\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)}) = \sup_{g_1, g_2 \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell\big(g_1(\mathbf{x}^{(S)}), g_2(\mathbf{x}^{(S)})\big) - \mathrm{E}^{(T)} \ell\big(g_1(\mathbf{x}^{(T)}), g_2(\mathbf{x}^{(T)})\big) \right|, \quad (24)$$

and then used this quantity to obtain generalization bounds based on Rademacher complexities for domain adaptation.

As mentioned by Mansour *et al.* [20], the quantities (23) and (24) match in the setting of classification tasks by setting ℓ as the absolute-value loss function, while the usage of (24) does not require the condition of "λ-close" but the usage of (23) does. Recalling Definition 3.1, since there is no limitation on the function class $\mathcal{F}$, the quantity $D_{\mathcal{F}}(S,T)$ can be used in both classification and regression tasks. Therefore, we only need to consider the relationship between the proposed quantity $D_{\mathcal{F}}(S,T)$ and the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$.

A.2   Relationship between $D_{\mathcal{F}}(S,T)$ and $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$

From Definition 3.1 and (22), we can find that the quantity $D_{\mathcal{F}}(S,T)$ directly measures the difference between two distributions of the domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$. However, as addressed in Section 2, if a domain $\mathcal{Z}^{(S)}$ differs from another domain $\mathcal{Z}^{(T)}$, there are three possibilities: $\mathcal{D}^{(S)}$ differs from $\mathcal{D}^{(T)}$, or $g_*^{(S)}$ differs from $g_*^{(T)}$, or both of them occur. Therefore, we need to consider two kinds of differences: the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$. Next, we will show that the integral probability metric $D_{\mathcal{F}}(S,T)$ can be bounded by the summation of two separate quantities that measure the difference between $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between $g_*^{(S)}$ and $g_*^{(T)}$, respectively.

As shown in (24), the quantity $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ measures the difference between the distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$. Next, we introduce another quantity to measure the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$:

10

Definition A.1 *Given a loss function ℓ and a function class $\mathcal{G}$, we define*

$$Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)}) := \sup_{g_1 \in \mathcal{G}} \left| \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(S)}(\mathbf{x}^{(T)})) \right|. \quad (25)$$

Note that if the loss function ℓ and the function class $\mathcal{G}$ are both non-trivial (*i.e.*, $\mathcal{F}$ is non-trivial), the quantity $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$ is a (semi)metric between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$. In fact, it is not hard to verify that $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$ satisfies the triangle inequality and is equal to *zero* if and only if $g_*^{(S)}$ and $g_*^{(T)}$ match.

By combining (22), (24) and (25), we have

$$\begin{aligned}
\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)}) &= \sup_{g_1, g_2 \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell(g_1(\mathbf{x}^{(S)}), g_2(\mathbf{x}^{(S)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_2(\mathbf{x}^{(T)})) \right| \\
&\geq \sup_{g_1 \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell(g_1(\mathbf{x}^{(S)}), g_*^{(S)}(\mathbf{x}^{(S)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(S)}(\mathbf{x}^{(T)})) \right| \\
&= \sup_{g_1 \in \mathcal{G}} \Big| \mathrm{E}^{(S)} \ell(g_1(\mathbf{x}^{(S)}), g_*^{(S)}(\mathbf{x}^{(S)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})) \\
&\quad + \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(S)}(\mathbf{x}^{(T)})) \Big| \\
&\geq \sup_{g_1 \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell(g_1(\mathbf{x}^{(S)}), g_*^{(S)}(\mathbf{x}^{(S)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})) \right| \\
&\quad - \sup_{g_1 \in \mathcal{G}} \left| \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})) - \mathrm{E}^{(T)} \ell(g_1(\mathbf{x}^{(T)}), g_*^{(S)}(\mathbf{x}^{(T)})) \right| \\
&= D_{\mathcal{F}}(S, T) - Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)}), \quad (26)
\end{aligned}$$

and thus

$$D_{\mathcal{F}}(S, T) \leq \mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)}) + Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)}). \quad (27)$$

Therefore, $D_{\mathcal{F}}(S, T)$ can be bounded by the summation of the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ and the quantity $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$, which measure the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$, respectively.

B Proof of Theorem 5.1

In order to achieve the proof, we need to develop the specific Hoeffding-type deviation inequality for multiple sources and the symmetrization inequality for domain adaptation with multiple sources.

B.1 Hoeffding-Type Deviation Inequality for Multiple Sources

Deviation (or concentration) inequalities play an essential role in obtaining the generalization bounds for a certain learning process. Generally, specific deviation inequalities need to be developed for different learning processes. There are many popular deviation and concentration inequalities, *e.g.*, Hoeffding's inequality, McDiarmid's inequality, Bennett's inequality, Bernstein's inequality and Talagrand's inequality. These results are all built under the assumption of same distribution, and thus they are not applicable (or at least cannot be directly applied) to the setting of multiple sources. Next, based on Hoeffding's inequality [21], we present a deviation inequality for multiple sources.

Theorem B.1 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a, b]$. Let $\mathbf{Z}_1^{N_k} = \{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}$ be the set of i.i.d. samples drawn from the source domain $\mathcal{Z}^{(S_k)} \subset \mathbb{R}^L$ $(1 \leq k \leq K)$. Given $\mathbf{w} = (w_1, \cdots, w_K) \in [0, 1]^K$ with $\sum_{k=1}^{K} w_k = 1$ and for any $f \in \mathcal{F}$, we define a function $F_{\mathbf{w}} : \mathbb{R}^{L \sum_{k=1}^{K} N_k} \to \mathbb{R}$ as*

$$F_{\mathbf{w}}\left(\{\mathbf{X}_1^{N_k}\}_{k=1}^{K} \right) := \sum_{k=1}^{K} w_k \left(\prod_{i \neq k} N_i \right) \sum_{n=1}^{N_k} f(\mathbf{x}_n^{(k)}), \quad (28)$$

11

where for any $1 \leq k \leq K$ and given $N_k \in \mathbb{N}$, the set $\mathbf{X}_1^{N_k}$ is denoted as

$$\mathbf{X}_1^{N_k} := \{\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)}, \cdots, \mathbf{x}_{N_k}^{(k)}\} \in (\mathbb{R}^L)^{N_k}.$$

Then, we have for any $\xi > 0$,

$$\Pr\left\{\left|\mathrm{E}^{(S)}F_\mathbf{w} - F_\mathbf{w}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)\right| > \xi\right\}$$

$$\leq 2\exp\left\{-\frac{2\xi^2}{(b-a)^2\left(\prod_{k=1}^K N_k\right)\left(\sum_{k=1}^K w_k^2\left(\prod_{i\neq k} N_i\right)\right)}\right\}, \tag{29}$$

where $\mathrm{E}^{(S)}$ stands for the expectation taken on all source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$.

This result is an extension of the classical Hoeffding-type deviation inequality under the assumption of same distribution (*cf.* [2]). Compared to the classical result, the resultant deviation inequality (29) is suitable to the setting of multiple sources. These two inequalities coincide when there is only one source, *i.e.*, $K = 1$

The proof of Theorem B.1 is processed by a martingale method. Before the formal proof, we introduce some essential notations.

Let $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ be sample sets drawn from multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$, respectively. Define a random variable

$$S_n^{(k)} := \mathrm{E}^{(S)}\left\{F_\mathbf{w}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^n\right\}, \quad 1 \leq k \leq K, \ 0 \leq n \leq N_k, \tag{30}$$

where

$$\mathbf{Z}_1^n = \{\mathbf{z}_1^{(k)}, \mathbf{z}_2^{(k)}, \cdots, \mathbf{z}_n^{(k)}\} \subseteq \mathbf{Z}_1^{N_k}, \text{ and } \mathbf{Z}_1^0 = \varnothing.$$

It is clear that

$$S_0^{(1)} = \mathrm{E}^{(S)}F_\mathbf{w} \text{ and } S_{N_K}^{(K)} = F_\mathbf{w}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K),$$

where $\mathrm{E}^{(S)}$ stands for the expectation taken on all source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$.

Then, according to (28) and (30), we have for any $1 \leq k \leq K$ and $1 \leq n \leq N_k$:

$$\begin{aligned} S_n^{(k)} - S_{n-1}^{(k)} =& \mathrm{E}^{(S)}\left\{F_\mathbf{w}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^n\right\} \\ & - \mathrm{E}^{(S)}\left\{F_\mathbf{w}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^{n-1}\right\} \\ =& \mathrm{E}^{(S)}\left\{\sum_{k=1}^K w_k\left(\prod_{i\neq k} N_i\right)\sum_{n=1}^{N_k} f(\mathbf{z}_n^{(k)})|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^n\right\} \\ & - \mathrm{E}^{(S)}\left\{\sum_{k=1}^K w_k\left(\prod_{i\neq k} N_i\right)\sum_{n=1}^{N_k} f(\mathbf{z}_n^{(k)})|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^{n-1}\right\} \\ =& \sum_{l=1}^{k-1} w_l\left(\prod_{i\neq l} N_i\right)\sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) + w_k\left(\prod_{i\neq k} N_i\right)\sum_{j=1}^n f(\mathbf{z}_j^{(k)}) \\ & + \mathrm{E}^{(S)}\left\{\sum_{l=k+1}^K w_l\left(\prod_{i\neq l} N_i\right)\sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) + w_k\left(\prod_{i\neq k} N_i\right)\sum_{j=n+1}^{N_k} f(\mathbf{z}_j^{(k)})\right\} \\ & - \sum_{l=1}^{k-1} w_l\left(\prod_{i\neq l} N_i\right)\sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) - w_k\left(\prod_{i\neq k} N_i\right)\sum_{j=1}^{n-1} f(\mathbf{z}_j^{(k)}) \\ & - \mathrm{E}^{(S)}\left\{\sum_{l=k+1}^K w_l\left(\prod_{i\neq l} N_i\right)\sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) + w_k\left(\prod_{i\neq k} N_i\right)\sum_{j=n}^{N_k} f(\mathbf{z}_j^{(k)})\right\} \\ =& w_k\left(\prod_{i\neq k} N_i\right)\left(f(\mathbf{z}_n^{(k)}) - \mathrm{E}^{(S_k)}f\right). \end{aligned} \tag{31}$$

12

To prove Theorem B.1, we need the following inequality resulted from Hoeffding's lemma.

Lemma B.2 *Let f be a function with the range $[a, b]$. Then, the following holds for any $\alpha > 0$:*

$$\mathrm{E}\left\{e^{\alpha(f(\mathbf{z}^{(S)})-\mathrm{E}^{(S)}f)}\right\} \leq e^{\frac{\alpha^2(b-a)^2}{8}}. \tag{32}$$

Proof. We consider

$$(f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)}f)$$

as a random variable. Then, it is clear that

$$\mathrm{E}\{f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)}f\} = 0.$$

Since the value of $\mathrm{E}^{(S)}f$ is a constant denoted as e, we have

$$a - e \leq f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)}f \leq b - e.$$

According to Hoeffding's lemma, we then have

$$\mathrm{E}\left\{e^{\alpha(f(\mathbf{z}^{(S)})-\mathrm{E}^{(S)}f)}\right\} \leq e^{\frac{\alpha^2(b-a)^2}{8}}. \tag{33}$$

This completes the proof. ∎

We are now ready to prove Theorem B.1.

Proof of Theorem B.1. According to (28), (31), Lemma B.2, Markov's inequality, Jensen's inequality and the law of iterated expectation, we have for any $\alpha > 0$,

$$\Pr\left\{F_{\mathbf{w}}\left(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\right) - \mathrm{E}^{(S)}F_{\mathbf{w}} > \xi\right\}$$

$$\leq e^{-\alpha\xi}\mathrm{E}\left\{e^{\alpha\left(F_{\mathbf{w}}\left(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\right)-\mathrm{E}^{(S)}F_{\mathbf{w}}\right)}\right\}$$

$$= e^{-\alpha\xi}\mathrm{E}\left\{\mathrm{E}\left\{e^{\alpha\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)}\Big|\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_{K-1}},\mathbf{Z}_1^{N_K-1}\right\}\right\}$$

$$= e^{-\alpha\xi}\mathrm{E}\left\{e^{\alpha\left(\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)-\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)\right)}\mathrm{E}\left\{e^{\alpha\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)}\Big|\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_{K-1}},\mathbf{Z}_1^{N_K-1}\right\}\right\}$$

$$= e^{-\alpha\xi}\mathrm{E}\left\{e^{\alpha\left(\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)-\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)\right)}\mathrm{E}\left\{e^{\alpha w_K(\prod_{i\neq K}N_i)(f(\mathbf{z}_N^{(K)})-\mathrm{E}^{(S_K)}f)}\right\}\right\}$$

$$\leq e^{-\alpha\xi}\mathrm{E}\left\{e^{\alpha\left(\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)-\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)\right)}\right\}e^{\frac{\alpha^2w_K^2(\prod_{i\neq K}N_i)^2(b-a)^2}{8}}, \tag{34}$$

where $\mathbf{Z}_1^{N_K-1} := \{\mathbf{z}_1^{(K)},\cdots,\mathbf{z}_{N_K-1}^{(K)}\} \subset \mathbf{Z}_1^{N_K}$. Therefore, we have

$$\Pr\left\{F_{\mathbf{w}}\left(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\right) - \mathrm{E}^{(S)}F_{\mathbf{w}} > \xi\right\} \leq e^{\Phi(\alpha)-\alpha\xi}, \tag{35}$$

where

$$\Phi(\alpha) = \frac{\alpha^2(b-a)^2\left(\prod_{k=1}^K N_k\right)\left(\sum_{k=1}^K w_k^2\left(\prod_{i\neq k}N_i\right)\right)}{8}. \tag{36}$$

Similarly, we can obtain

$$\Pr\left\{\mathrm{E}^{(S)}F_{\mathbf{w}} - F_{\mathbf{w}}\left(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\right) > \xi\right\} \leq e^{\Phi(\alpha)-\alpha\xi}. \tag{37}$$

Note that $\Phi(\alpha) - \alpha\xi$ is a quadratic function with respect to $\alpha > 0$ and thus the minimum value "$\min_{\alpha>0}\{\Phi(\alpha) - \alpha\xi\}$" is achieved when

$$\alpha = \frac{4\xi}{(b-a)^2\left(\prod_{k=1}^K N_k\right)\left(\sum_{k=1}^K w_k^2\left(\prod_{i\neq k}N_i\right)\right)}.$$

13

By combining (35), (36) and (37), we arrive at

$$\Pr\left\{\left|\mathrm{E}^{(S)}F_{\mathbf{w}} - F_{\mathbf{w}}\big(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\big)\right| > \xi\right\}$$

$$\leq 2\exp\left\{-\frac{2\xi^2}{(b-a)^2\big(\prod_{k=1}^K N_k\big)\Big(\sum_{k=1}^K w_k^2\big(\prod_{i\neq k} N_i\big)\Big)}\right\}.$$

This completes the proof. ∎

In the following subsection, we present a symmetrization inequality for domain adaptation with multiple sources.

B.2 Symmetrization Inequality

Symmetrization inequalities are mainly used to replace the expected risk by an empirical risk computed on another sample set that is independent of the given sample set but has the same distribution. In this manner, the generalization bounds can be achieved based on some kinds of complexity measures, *e.g.*, the covering number and the VC dimension. However, the classical symmetrization results are built under the assumption of same distribution (*cf.* [2]). The symmetrization inequality for domain adaptation with multiple sources is presented in the following theorem:

Theorem B.3 *Assume that $\mathcal{F}$ is a function class with the range $[a,b]$. Let sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ and $\{\mathbf{Z'}_1^{N_k}\}_{k=1}^K$ be drawn from the source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$. Then, given an arbitrary $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and $\mathbf{w} \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$, we have for any $\big(\prod_{k=1}^K N_k\big) \geq \frac{8(b-a)^2}{(\xi')^2}$,*

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \xi\right\} \leq 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E'}_{\mathbf{w}}^{(S)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \frac{\xi'}{2}\right\}, \qquad (38)$$

where $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$.

This theorem shows that given $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, the probability of the event:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \xi$$

can be bounded by using the probability of the event:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E'}_{\mathbf{w}}^{(S)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \frac{\xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)}{2} \qquad (39)$$

that is only determined by the characteristics of the source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ when $\prod_{k=1}^K N_k \geq 8(b-a)^2/(\xi')^2$ with $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. Compared to the classical symmetrization result under the assumption of same distribution (*cf.* [2]), there is a discrepancy term $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ in the derived inequality. Especially, the two results will coincide when any source domain and the target domain match, *i.e.*, $D_{\mathcal{F}}(S_k,T) = 0$ holds for any $1 \leq k \leq K$. The following is the proof of Theorem B.3.

Proof of Theorem B.3. Let $\widehat{f}$ be the function achieving the supremum:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right|$$

with respect to the sample set $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$. According to (6), (8) and (12), we arrive at

$$\left|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right| = \left|\mathrm{E}^{(T)}\widehat{f} - \overline{\mathrm{E}}^{(S)}\widehat{f} + \overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + \left|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right|, \quad (40)$$

and thus,

$$\Pr\left\{\left|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right| > \xi\right\} \leq \Pr\left\{D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + \left|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right| > \xi\right\}, \qquad (41)$$

14

where the expectation $\overline{\mathrm{E}}^{(S)}\widehat{f}$ is defined as

$$\overline{\mathrm{E}}^{(S)}\widehat{f} := \sum_{k=1}^{K} w_k \mathrm{E}^{(S_k)}\widehat{f}. \tag{42}$$

Let

$$\xi' := \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S, T), \tag{43}$$

and denote $\wedge$ as the conjunction of two events. According to the triangle inequality, we have

$$\left(|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| - |\mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f} - \overline{\mathrm{E}}^{(S)}\widehat{f}| \right) \leq |\mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|,$$

and thus for any $\xi' > 0$,

$$\left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \xi'} \right) \left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f}| < \frac{\xi'}{2}} \right)$$

$$= \mathbf{1}_{\left\{ |\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \xi' \right\} \wedge \left\{ |\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f}| < \frac{\xi'}{2} \right\}}$$

$$\leq \mathbf{1}_{|\mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \frac{\xi'}{2}} .$$

Then, taking the expectation with respect to $\{\mathbf{Z}'_1^{N_k}\}_{k=1}^{K}$ gives

$$\left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \xi'} \right) \mathrm{Pr}' \left\{ |\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f}| < \frac{\xi'}{2} \right\}$$

$$\leq \mathrm{Pr}' \left\{ |\mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \frac{\xi'}{2} \right\}. \tag{44}$$

By Chebyshev's inequality, since $\{\mathbf{Z}'_1^{N_k}\}_{k=1}^{K}$ are the sets of i.i.d. samples drawn from the multiple sources $\{\mathscr{Z}^{(S_k)}\}_{k=1}^{K}$ respectively, we have for any $\xi' > 0$,

$$\mathrm{Pr}' \left\{ |\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f}| \geq \frac{\xi'}{2} \right\} \leq \mathrm{Pr}' \left\{ \sum_{k=1}^{K} \frac{w_k}{N_k} \sum_{n=1}^{N_k} |\mathrm{E}^{(S_k)}\widehat{f} - \widehat{f}(\mathbf{z}'_n^{(k)})| \geq \frac{\xi'}{2} \right\}$$

$$= \mathrm{Pr}' \left\{ \sum_{k=1}^{K} w_k \left(\prod_{i \neq k} N_i \right) \sum_{n=1}^{N_k} |\mathrm{E}^{(S_k)}\widehat{f} - \widehat{f}(\mathbf{z}'_n^{(k)})| \geq \frac{\xi' \prod_{k=1}^{K} N_k}{2} \right\}$$

$$\leq \frac{4\mathrm{E} \left\{ \sum_{k=1}^{K} w_k \left(\prod_{i \neq k} N_i \right) \sum_{n=1}^{N_k} |\mathrm{E}^{(S_k)}\widehat{f} - \widehat{f}(\mathbf{z}'_n^{(k)})|^2 \right\}}{\left(\prod_{k=1}^{K} N_k \right)^2 (\xi')^2}$$

$$= \frac{4\mathrm{E} \left\{ \sum_{k=1}^{K} w_k \left(\prod_{i \neq k} N_i \right) N_k (b-a)^2 \right\}}{\left(\prod_{k=1}^{K} N_k \right)^2 (\xi')^2}$$

$$= \frac{4\left(\prod_{k=1}^{K} N_k \right)(b-a)^2}{\left(\prod_{k=1}^{K} N_k \right)^2 (\xi')^2} = \frac{4(b-a)^2}{(\xi')^2 \left(\prod_{k=1}^{K} N_k \right)}. \tag{45}$$

Subsequently, according to (44) and (45), we have for any $\xi' > 0$,

$$\mathrm{Pr}' \left\{ |\mathrm{E}'_{\mathbf{w}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \frac{\xi'}{2} \right\} \geq \left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \xi'} \right) \left(1 - \frac{4(b-a)^2}{(\xi')^2 \left(\prod_{k=1}^{K} N_k \right)} \right). \tag{46}$$

By combining (41), (43) and (46), taking the expectation with respect to $\{\mathbf{Z}_1^{N_k}\}_{k=1}^{K}$ and letting

$$\frac{4(b-a)^2}{(\xi')^2 \left(\prod_{k=1}^{K} N_k \right)} \leq \frac{1}{2}$$

15

can lead to: for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$,

$$\Pr\left\{|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \xi\right\} \leq \Pr\left\{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \xi'\right\}$$

$$\leq 2\Pr\left\{|\mathrm{E}_{\mathbf{w}}'^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| > \frac{\xi'}{2}\right\} \tag{47}$$

with $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. This completes the proof. ∎

By using the resultant deviation inequality and the symmetrization inequality, we can achieve the proof of Theorem 5.1.

B.3 Proof of Theorem 5.1

Proof of Theorem 5.1. Consider ϵ as an independent Rademacher random variables, *i.e.*, an independent $\{-1,1\}$-valued random variable with equal probability of taking either value. Given sample sets $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$, denote for any $f \in \mathcal{F}$ and $1 \leq k \leq K$,

$$\overrightarrow{\epsilon}^{(k)} := (\epsilon_1^{(k)}, \cdots, \epsilon_{N_k}^{(k)}, -\epsilon_1^{(k)}, \cdots, -\epsilon_{N_k}^{(k)}) \in \{-1,1\}^{2N_k}, \tag{48}$$

and for any $f \in \mathcal{F}$,

$$\overrightarrow{f}(\mathbf{Z}_1^{2N_k}) := \left(f(\mathbf{z'}_1^{(k)}), \cdots, f(\mathbf{z'}_{N_k}^{(k)}), f(\mathbf{z}_1^{(k)}), \cdots, f(\mathbf{z}_{N_k}^{(k)})\right). \tag{49}$$

According to (6) and Theorem B.3, given an arbitrary $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and denoting $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, we have for any $\{N_k\}_{k=1}^K \in \mathbb{N}^K$ such that $(\prod_{k=1}^K N_k) \geq 8(b-a)^2/(\xi')^2$,

$$\Pr\left\{\sup_{f \in \mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \xi\right\}$$

$$\leq 2\Pr\left\{\sup_{f \in \mathcal{F}}\left|\mathrm{E}_{\mathbf{w}}'^{(S)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \frac{\xi'}{2}\right\} \qquad \text{(by Theorem B.3)}$$

$$= 2\Pr\left\{\sup_{f \in \mathcal{F}}\left|\sum_{k=1}^K \frac{w_k}{N_k}\sum_{n=1}^{N_k}\left(f(\mathbf{z'}_n^{(k)}) - f(\mathbf{z}_n^{(k)})\right)\right| > \frac{\xi'}{2}\right\}$$

$$= 2\Pr\left\{\sup_{f \in \mathcal{F}}\left|\sum_{k=1}^K \frac{w_k}{N_k}\sum_{n=1}^{N_k}\epsilon_n^{(k)}\left(f(\mathbf{z'}_n^{(k)}) - f(\mathbf{z}_n^{(k)})\right)\right| > \frac{\xi'}{2}\right\}$$

$$= 2\Pr\left\{\sup_{f \in \mathcal{F}}\left|\sum_{k=1}^K \frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)}, \overrightarrow{f}(\mathbf{Z}_1^{2N_k})\rangle\right| > \frac{\xi'}{4}\right\}. \qquad \text{(by (48) and (49))} \tag{50}$$

Fix a realization of $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$ and let Λ be a $\xi'/8$-radius cover of $\mathcal{F}$ with respect to the $\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)$ norm. Since $\mathcal{F}$ is composed of the bounded functions with the range $[a,b]$, we assume that the same holds for any $h \in \Lambda$. If f_0 is the function that achieves the following supremum

$$\sup_{f \in \mathcal{F}}\left|\sum_{k=1}^K \frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)}, \overrightarrow{f}(\mathbf{Z}_1^{2N_k})\rangle\right| > \frac{\xi'}{4},$$

there must be an $h_0 \in \Lambda$ that satisfies

$$\sum_{k=1}^K \frac{w_k}{2N_k}\left(|f_0(\mathbf{z'}_n^{(k)}) - h_0(\mathbf{z'}_n^{(k)})| + |f_0(\mathbf{z}_n^{(k)}) - h_0(\mathbf{z}_n^{(k)})|\right) < \frac{\xi'}{8},$$

and meanwhile,

$$\left|\sum_{k=1}^K \frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)}, \overrightarrow{h_0}(\mathbf{Z}_1^{2N_k})\rangle\right| > \frac{\xi'}{8}.$$

16

Therefore, for the realization of $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$, we arrive at

$$\Pr\left\{\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{f}(\mathbf{Z}_1^{2N_k})\rangle\Big|>\frac{\xi'}{4}\right\}$$

$$\leq\Pr\left\{\sup_{h\in\Lambda}\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\rangle\Big|>\frac{\xi'}{8}\right\}. \tag{51}$$

Moreover, we denote the event

$$A:=\left\{\Pr\left\{\sup_{h\in\Lambda}\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\rangle\Big|>\frac{\xi'}{8}\right\}\right\},$$

and let $\mathbf{1}_A$ be the characteristic function of the event A. By Fubini's Theorem, we have

$$\Pr\{A\}=\mathrm{E}\left\{\mathrm{E}_{\overrightarrow{\epsilon}}\{\mathbf{1}_A\}\big|\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K\right\}$$

$$=\mathrm{E}\left\{\Pr\left\{\sup_{h\in\Lambda}\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\rangle\Big|>\frac{\xi'}{8}\right\}\Big|\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K\right\}. \tag{52}$$

Fix a realization of $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$ again. According to (48), (49) and Theorem B.1, we have

$$\Pr\left\{\sup_{h\in\Lambda}\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\rangle\Big|>\frac{\xi'}{8}\right\}$$

$$\leq|\Lambda|\max_{h\in\Lambda}\Pr\left\{\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\rangle\Big|>\frac{\xi'}{8}\right\}$$

$$=\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\max_{h\in\Lambda}\Pr\left\{\big|\mathrm{E}'^{(S)}_{\mathbf{w}}h-\mathrm{E}^{(S)}_{\mathbf{w}}h\big|>\frac{\xi'}{4}\right\}$$

$$\leq\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\max_{h\in\Lambda}\Pr\left\{\big|\overline{\mathrm{E}}^{(S)}h-\mathrm{E}'^{(S)}_{\mathbf{w}}h\big|+\big|\overline{\mathrm{E}}^{(S)}h-\mathrm{E}^{(S)}_{\mathbf{w}}h\big|>\frac{\xi'}{4}\right\}$$

$$\leq2\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\max_{h\in\Lambda}\Pr\left\{\big|\overline{\mathrm{E}}^{(S)}h-\mathrm{E}^{(S)}_{\mathbf{w}}h\big|>\frac{\xi'}{8}\right\}$$

$$\leq4\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\}, \tag{53}$$

where the expectation $\overline{\mathrm{E}}^{(S)}$ is defined in (42).

The combination of (50), (51) and (53) leads to the result: given an arbitrary $\xi>D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and for any $\big(\prod_{k=1}^K N_k\big)\geq8\,(b-a)^2/(\xi')^2$ with $\xi'=\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$,

$$\Pr\left\{\sup_{f\in\mathcal{F}}\big|\mathrm{E}^{(T)}f-\mathrm{E}^{(S)}_{\mathbf{w}}f\big|>\xi\right\}$$

$$\leq8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\}$$

$$\leq8\mathcal{N}_1^{\mathbf{w}}\left(\mathcal{F},\xi'/8,2\sum_{k=1}^K N_k\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\}. \tag{54}$$

17

According to (54), letting

$$\epsilon := 8\mathcal{N}_1^{\mathbf{w}}\left(\mathcal{F}, \xi'/8, 2\sum_{k=1}^{K} N_k\right) \exp\left\{-\frac{\left(\prod_{k=1}^{K} N_k\right)\left(\xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\right)^2}{32(b-a)^2\left(\sum_{k=1}^{K} w_k^2(\prod_{i\neq k} N_i)\right)}\right\},$$

we then arrive at with probability at least $1 - \epsilon$,

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}_{\mathbf{w}}^{(S)}f - \mathrm{E}^{(T)}f\right| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + \left(\frac{\ln\mathcal{N}_1^{\mathbf{w}}\left(\mathcal{F}, \xi'/8, 2\sum_{k=1}^{K} N_k\right) - \ln(\epsilon/8)}{\frac{\left(\prod_{k=1}^{K} N_k\right)}{32(b-a)^2\left(\sum_{k=1}^{K} w_k^2(\prod_{i\neq k} N_i)\right)}}\right)^{\frac{1}{2}},$$

where $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. This completes the proof. ∎

C Domain Adaptation Combining Source and Target Data

Here, we study the generalization bounds of the learning process for domain adaptation combining source and target data. We first introduce some notations to formalize the problem.

C.1 Problem Setup

Denote $\mathcal{Z}^{(S)} := \mathcal{X}^{(S)} \times \mathcal{Y}^{(S)} \subset \mathbb{R}^I \times \mathbb{R}^J$ and $\mathcal{Z}^{(T)} := \mathcal{X}^{(T)} \times \mathcal{Y}^{(T)} \subset \mathbb{R}^I \times \mathbb{R}^J$ as the source domain and the target domain, respectively. Let $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ stand for the distributions of the input spaces $\mathcal{X}^{(S)}$ and $\mathcal{X}^{(T)}$, respectively. Denote $g_*^{(S)} : \mathcal{X}^{(S)} \to \mathcal{Y}^{(S)}$ and $g_*^{(T)} : \mathcal{X}^{(T)} \to \mathcal{Y}^{(T)}$ as the labeling functions of $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. In the situation of domain adaptation combining source and target data (*cf.* [13, 18]), the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ differ from each other, or the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$ differ from each other, or both cases occur. There are some (but not enough) samples $\mathbf{Z}_1^{N_T} := \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ drawn from the target domain $\mathcal{Z}^{(T)}$ in addition to a large amount of samples $\mathbf{Z}_1^{N_S} := \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ drawn from the source domain $\mathcal{Z}^{(S)}$ with $N^{(T)} \ll N^{(S)}$. Given a $\tau \in [0,1)$, we denote $g_\tau \in \mathcal{G}$ as the function that minimizes the convex combination of the source and the target empirical risks over $\mathcal{G}$:

$$\mathrm{E}_\tau(\ell \circ g) := \tau \mathrm{E}_{N_T}^{(T)}(\ell \circ g) + (1-\tau)\mathrm{E}_{N_S}^{(S)}(\ell \circ g), \tag{55}$$

and it is expected that g_τ will perform well for any pair $\mathbf{z}^{(T)} = (\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) \in \mathcal{Z}^{(T)}$, *i.e.*, g_τ approximates the labeling function $g_*^{(T)}$ as precisely as possible.

As mentioned in [13,18], setting τ involves a tradeoff between the source data that are sufficient but not accurate and the target data that are accurate but not sufficient. Especially, setting $\tau = 0$ provides a learning process of the basic domain adaptation with one single source (*cf.* [19]).

Similar to the situation of domain adaptation with multiple sources, two types of quantities: $\mathrm{E}^{(T)}(\ell \circ g_\tau) - \mathrm{E}_\tau(\ell \circ g_\tau)$ and $\mathrm{E}^{(T)}(\ell \circ g_\tau) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*)$ also play an essential role in analyzing the asymptotic behavior of the learning process for domain adaptation combining source and target data. By the similar way of (3), we need to consider the supremum

$$\sup_{g\in\mathcal{G}}\left|\mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_\tau(\ell \circ g)\right|, \tag{56}$$

which is the so-called generalization bound of the learning process for domain adaptation combining source and target data. Following the notation of (6), the generalization bound (56) can be equivalently rewritten as

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_\tau f\right|.$$

18

C.2 The Uniform Entropy Number

In the situation of domain adaptation combining source and target data, we have to introduce a variant of the ℓ_1 norm on $\mathcal{F}$. Let $\mathbf{Z}_1^{N_S} = \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} = \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be two sets of samples drawn from the domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Moreover, let $\mathbf{Z'}_1^{N_S}$ and $\overline{\mathbf{Z}}'_1^{N_T}$ be the ghost sample sets of $\mathbf{Z}_1^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T}$, respectively. Denote $\mathbf{Z}_1^{2N_S} := \{\mathbf{Z}_1^{N_S}, \mathbf{Z'}_1^{N_S}\}$ and $\overline{\mathbf{Z}}_1^{2N_T} := \{\overline{\mathbf{Z}}_1^{N_T}, \overline{\mathbf{Z}}'_1^{N_T}\}$, respectively.

Given an $f \in \mathcal{F}$, we define for any $\tau \in [0, 1)$,

$$\|f\|_{\ell_1^\tau(\{\mathbf{Z}_1^{2N_S}, \overline{\mathbf{Z}}_1^{2N_T}\})} := \frac{\tau}{N_T} \sum_{n=1}^{N_T} \left(|f(\mathbf{z}_n^{(T)})| + |f(\mathbf{z'}_n^{(T)})|\right) + \frac{1-\tau}{N_S} \sum_{n=1}^{N_S} \left(|f(\mathbf{z}_n^{(S)})| + |f(\mathbf{z'}_n^{(S)})|\right). \tag{57}$$

Note that the variant ℓ_1^τ ($\tau \in [0, 1)$) of the ℓ_1 norm is still a norm on the functional space, which can be easily verified by using the definition of norm, so we omit it here. Then, the uniform entropy number of $\mathcal{F}$ with respect to the $\ell_1^\tau(\{\mathbf{Z}_1^{2N_S}, \overline{\mathbf{Z}}_1^{2N_T}\})$ is defined as

$$\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi, 2(N_S + N_T)) := \sup_{\mathbf{Z}_1^{2N_S}, \overline{\mathbf{Z}}_1^{2N_T}} \ln \mathcal{N}\left(\mathcal{F}, \xi, \ell_1^\tau(\{\mathbf{Z}_1^{2N_S}, \overline{\mathbf{Z}}_1^{2N_T}\})\right), \tag{58}$$

where $\mathcal{N}\left(\mathcal{F}, \xi, \ell_1^\tau(\{\mathbf{Z}_1^{2N_S}, \overline{\mathbf{Z}}_1^{2N_T}\})\right)$ is the covering number of the function class $\mathcal{F}$ with respect to the norm $\ell_1^\tau(\{\mathbf{Z}_1^{2N_S}, \overline{\mathbf{Z}}_1^{2N_T}\})$.

C.3 Generalization Bounds

The following theorem provides a generalization bound for domain adaptation combining source and target data based on the uniform entropy number with respect to the norm ℓ_1^τ (*cf.* (58)).

Theorem C.1 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a, b]$. Let $\mathbf{Z}_1^{N_S} = \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} = \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be two sets of i.i.d. samples drawn from domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Then, for any $\tau \in [0, 1)$ and given an arbitrary $\xi > (1 - \tau)D_\mathcal{F}(S, T)$, we have for any $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$, with probability at least $1 - \epsilon$,*

$$\sup_{f \in \mathcal{F}} \left|\mathrm{E}_\tau f - \mathrm{E}^{(T)} f\right| \leq (1 - \tau)D_\mathcal{F}(S, T) + \left(\frac{\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T)) - \ln(\epsilon/8)}{\frac{N_S N_T}{32(b-a)^2((1-\tau)^2 N_T + \tau^2 N_S)}}\right)^{\frac{1}{2}}, \tag{59}$$

where $D_\mathcal{F}(S, T)$ is defined in (8) and $\xi' := \xi - (1 - \tau)D_\mathcal{F}(S, T)$.

Similar to the situation of domain adaptation with multiple sources, the proof of this theorem is processed by using a specific Hoeffding-type deviation inequality and a symmetrization inequality for domain adaptation combining source and target data (*cf.* Subsection C.6).

Compared to the classical result (13) under the assumption of same distribution, the expression of the generalization bound (59) contains a discrepancy term $(1 - \tau)D_\mathcal{F}(S, T)$ that is determined by two factors: the combination coefficient τ and the integral probability metric $D_\mathcal{F}(S, T)$. The two bounds coincide when the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.*, $D_\mathcal{F}(S, T) = 0$.

Note that this result exhibits a tradeoff between the sample numbers N_S and N_T. Although the tradeoff has been mentioned in some previous works (*cf.* [13, 18]), the following subsection will show a rigorous theoretical analysis of the tradeoff based on our resultant generalization bound for domain adaptation combining source and target data.

C.4 Asymptotic Convergence

Based on Theorem C.1, we can directly obtain the following result pointing out that the asymptotic convergence of the learning process for domain adaptation combining source and target data is affected by the uniform entropy number $\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T))$, the combination coefficient τ and the integral probability metric $D_\mathcal{F}(S, T)$.

19

Theorem C.2 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a, b]$. Given a $\tau \in [0, 1)$, if the following condition holds:*

$$\lim_{N_S \to +\infty} \frac{\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T))}{\frac{N_S N_T}{((1-\tau)^2 N_T + \tau^2 N_S)}} < +\infty \qquad (60)$$

with $\xi' := \xi - (1-\tau)D_\mathcal{F}(S, T)$, then we have for any $\xi > (1-\tau)D_\mathcal{F}(S, T)$,

$$\lim_{N_S \to +\infty} \Pr \left\{ \sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_\tau f \right| > \xi \right\} = 0. \qquad (61)$$

As shown in Theorem C.2, if the choice of $\tau \in [0, 1)$ and the uniform entropy number $\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T))$ satisfy the condition (60), the probability of the event "$\sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_\tau f \right| > \xi$" will converge to *zero* for any $\xi > (1 - \tau)D_\mathcal{F}(S, T)$, when N_S goes to *infinity*. This is partially in accordance with the classical result under the assumption of same distributions given by the combination of Theorem 2.3 and Definition 2.5 of [22].

Note that in the learning process for domain adaptation combining source and target data, the uniform convergence of the empirical risk on the source domain to the expected risk on the target domain may not hold, because the limit (61) does not hold for any $\xi > 0$ but for any $\xi > (1 - \tau)D_\mathcal{F}(S, T)$. By contrast, the limit (61) holds for all $\xi > 0$ in the learning process under the assumption of same distribution, if the condition (16) is satisfied. The two conditions coincide when the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.,* $D_\mathcal{F}(S, T) = 0$.

Additionally, we consider the choice of τ that is an essential factor to the asymptotic convergence of the learning process and is associated with the tradeoff between the sample numbers N_S and N_T. Recalling (59), if we fix the value of $\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T))$, setting $\tau = \frac{N_T}{N_T + N_S}$ minimizes the second term of the right-hand side of (59) and then we arrive at

$$\sup_{f \in \mathcal{F}} \left| \mathrm{E}_\tau f - \mathrm{E}^{(T)} f \right| \leq \frac{N_S D_\mathcal{F}(S, T)}{N_S + N_T} + \left(\frac{\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T)) - \ln(\epsilon/8)}{\frac{N_S + N_T}{32(b-a)^2}} \right)^{\frac{1}{2}}, \quad (62)$$

which implies that setting $\tau = \frac{N_T}{N_T + N_S}$ can result in the fastest rate of convergence, while it can also cause the relatively larger discrepancy between the empirical risk $\mathrm{E}_\tau f$ and the expected risk $\mathrm{E}^{(T)} f$, because the situation of domain adaptation is set up in the condition that $N_T \ll N_S$, which implies that $\frac{N_S}{N_S + N_T} \approx 1$. It is noteworthy that the value $\tau = \frac{N_T}{N_T + N_S}$ has been mentioned in the section of "Experimental Results" in [18]. Here, we show a rigorous theoretical analysis of this value and the related numerical experiment also supports this point (*cf.* Fig. 2).

Similar to the situation of domain adaptation with multiple sources, by setting $\tau = \frac{N_T}{N_T + N_S}$ and ignoring the discrepancy term $N_S D_\mathcal{F}(S, T)/(N_S + N_T)$, the learning process for domain adaptation combining source and target data has the same rate of convergence as that of the learning process under the assumption of same distribution [*cf.* (13) and (62)].

C.5 Numerical Experiments

We have performed some numerical experiments to verify the theoretical analysis of the asymptotic convergence of the learning process for domain adaptation combining source and target data.

In this situation, the samples $\{(\mathbf{x}_n^{(T)}, y_n^{(T)})\}_{n=1}^{N_T}$ ($N_T = 4000$) of the target domain $\mathcal{Z}^{(T)}$ are generated in the aforementioned way (*cf.* (18)). We randomly pick $N_T' = 100$ samples from them to form the objective function and the rest $N_T'' = 3900$ are used to test.

Similarly, the samples $\{(\mathbf{x}_n^{(S)}, y_n^{(S)})\}_{n=1}^{N_S}$ ($N_S = 4000$) of the source domain $\mathcal{Z}^{(S)}$ are generated as follows: for any $1 \leq n \leq N_S$,

$$y_n^{(S)} = \langle \mathbf{x}_n^{(S)}, \beta \rangle + R, \qquad (63)$$

where $\mathbf{x}_n^{(S)} \sim N(1, 2)$, $\beta \sim N(1, 5)$ and $R \sim N(0, 0.5)$.

20

We also use the method of Least Square Regression to minimize the empirical risk

$$\mathrm{E}_\tau(\ell \circ g) = \frac{1-\tau}{N_S} \sum_{n=1}^{N_S} \ell(g(\mathbf{x}_n^{(S)}), y_n^{(S)}) + \frac{\tau}{N_T'} \sum_{n=1}^{N_T'} \ell(g(\mathbf{x}_n^{(T)}), y_n^{(T)})$$

for different combination coefficients $\tau \in \{0.1, 0.3, 0.5, 0.9\}$ and then compute the discrepancy $|E_\tau f - E_{N_T''}^{(T)} f|$ for each N_S. Since $N_S \gg N_T'$, the initial N_S is set to be 200. Each test is repeated 100 times and the final result is the average of the 100 results. After each test, we increment N_S by 200 until $N_S = 4000$. The experiment results are shown in Fig. 2.
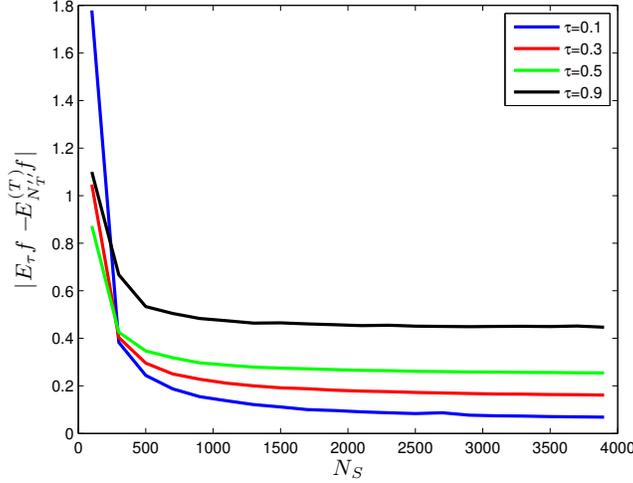
Figure 2: Domain Adaptation Combining Source and Target Data

Fig. (2) shows that for any choice of $\tau \in \{0.1, 0.3, 0.5, 0.9\}$, the curve of $|E_\tau f - E_{N_T''}^{(T)}|$ is decreasing as N_S increases. This is in accordance with our results of the asymptotic convergence of the learning process for domain adaptation combining source and target data (*cf.* Theorems C.1 and C.2).

At the end of Subsection C.4, we have theoretically analyzed how the choice of τ affects the rate of convergence of the learning process for domain adaptation combining source and target data. Our numerical experiments support this theoretical finding as well. In fact, as shown in Fig. 2, when $\tau \approx N_T'/(N_S + N_T')$, the discrepancy $|E_\tau^{(S)} f - E_{N_T''}^{(T)} f|$ has the fastest rate of convergence, and the rate becomes slower as τ is further away from $N_T'/(N_S + N_T')$. Thus, this is in accordance with the theoretical analysis of the asymptotic convergence presented above.

C.6 Proof of Theorem C.1

Here, we provide the proof of Theorem C.1. Similar to the situation of domain adaptation with multiple sources, we need to develop the related Hoeffding-type deviation inequality and the symmetrization inequality for domain adaptation combining source and target data.

C.6.1 Hoeffding-Type Deviation Inequality

Based on Hoeffding's inequality [21], we derive a deviation inequality for the combination of source and target domains.

Theorem C.3 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a, b]$. Let $\mathbf{Z}_1^{N_S} := \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} := \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be sets of i.i.d. samples drawn from the source domain $\mathcal{Z}^{(S)} \subset \mathbb{R}^L$ and the target domain $\mathcal{Z}^{(T)} \subset \mathbb{R}^L$, respectively. For any $\tau \in [0, 1)$,*

21

define a function $F_\tau : \mathbb{R}^{L(N_S+N_T)} \to \mathbb{R}$ as

$$F_\tau\left(\mathbf{X}_1^{N_T}, \mathbf{Y}_1^{N_S}\right) := \tau N_S \sum_{n=1}^{N_T} f(\mathbf{x}_n) + (1-\tau)N_T \sum_{n=1}^{N_S} f(\mathbf{y}_n), \tag{64}$$

where

$$\mathbf{X}_1^{N_T} := \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_T}\} \in (\mathbb{R}^L)^{N_T}; \ \ \mathbf{Y}_1^{N_S} := \{\mathbf{y}_1, \cdots, \mathbf{y}_{N_S}\} \in (\mathbb{R}^L)^{N_S}.$$

Then, we have for any $\tau \in [0,1)$ and any $\xi > 0$,

$$\Pr\left\{\left|F_\tau\left(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\right) - \mathrm{E}^{(*)}F_\tau\right| > \xi\right\}$$

$$\leq 2\exp\left\{-\frac{2\xi^2}{(b-a)^2 N_S N_T\left((1-\tau)^2 N_T + \tau^2 N_S\right)}\right\}, \tag{65}$$

where the expectation $\mathrm{E}^{(*)}$ is taken on both of the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$.

In this theorem, we present a deviation inequality for the combination of source and target domains, which is an extension of the classical Hoeffding-type deviation inequality under the assumption of same distribution (*cf.* [2]). Compare to the classical result, the deviation inequality (65) allows the samples to be drawn from two different domains.

The proof of Theorem C.3 is also processed by a martingale method. Before the formal proof, we introduce some essential notations.

For any $\tau \in [0,1)$, we denote

$$F_S(\mathbf{Z}_1^{N_S}) := (1-\tau)N_T \sum_{n=1}^{N_S} f(\mathbf{z}_n^{(S)}); \ \ F_T(\overline{\mathbf{Z}}_1^{N_T}) := \tau N_S \sum_{n=1}^{N_T} f(\mathbf{z}_n^{(T)}). \tag{66}$$

Recalling (64), it is evident that $F_\tau\left(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\right) = F_S(\mathbf{Z}_1^{N_S}) + F_T(\overline{\mathbf{Z}}_1^{N_T})$. We then define two random variables:

$$S_n := \mathrm{E}^{(S)}\left\{F_S(\mathbf{Z}_1^{N_S})|\mathbf{Z}_1^n\right\}, \ 0 \leq n \leq N_S;$$

$$T_n := \mathrm{E}^{(T)}\left\{F_T(\overline{\mathbf{Z}}_1^{N_T})|\overline{\mathbf{Z}}_1^n\right\}, \ 0 \leq n \leq N_T, \tag{67}$$

where

$$\mathbf{Z}_1^n = \{\mathbf{z}_1^{(S)}, \cdots, \mathbf{z}_n^{(S)}\} \subseteq \mathbf{Z}_1^{N_S} \ \text{ with } \ \mathbf{Z}_1^0 := \varnothing;$$

$$\overline{\mathbf{Z}}_1^n = \{\mathbf{z}_1^{(T)}, \cdots, \mathbf{z}_n^{(T)}\} \subseteq \overline{\mathbf{Z}}_1^{N_T} \ \text{ with } \ \overline{\mathbf{Z}}_1^0 := \varnothing.$$

It is clear that $S_0 = \mathrm{E}^{(S)}F_S$; $S_{N_S} = F_S(\mathbf{Z}_1^{N_S})$ and $T_0 = \mathrm{E}^{(T)}F_T$; $T_{N_T} = F_T(\overline{\mathbf{Z}}_1^{N_T})$.

According to (64) and (67), we have for any $1 \leq n \leq N_S$ and any $\tau \in [0,1)$,

$$S_n - S_{n-1}$$

$$= \mathrm{E}^{(S)}\left\{F_S(\mathbf{Z}_1^{N_S})|\mathbf{Z}_1^n\right\} - \mathrm{E}^{(S)}\left\{F_S(\mathbf{Z}_1^{N_S})|\mathbf{Z}_1^{n-1}\right\}$$

$$= \mathrm{E}^{(S)}\left\{(1-\tau)N_T \sum_{n=1}^{N_S} f(\mathbf{z}_n^{(S)})\Big|\mathbf{Z}_1^n\right\} - \mathrm{E}^{(S)}\left\{(1-\tau)N_T \sum_{n=1}^{N_S} f(\mathbf{z}_n^{(S)})\Big|\mathbf{Z}_1^{n-1}\right\}$$

$$= (1-\tau)N_T \sum_{m=1}^n f(\mathbf{z}_m^{(S)}) + \mathrm{E}^{(S)}\left\{(1-\tau)N_T \sum_{m=n+1}^{N_S} f(\mathbf{z}_m^{(S)})\right\}$$

$$- \left((1-\tau)N_T \sum_{m=1}^{n-1} f(\mathbf{z}_m^{(S)}) + \mathrm{E}^{(S)}\left\{(1-\tau)N_T \sum_{m=n}^{N_S} f(\mathbf{z}_m^{(S)})\right\}\right)$$

$$= (1-\tau)N_T\left(f(\mathbf{z}_n^{(S)}) - \mathrm{E}^{(S)}f\right). \tag{68}$$

22

Similarly, we also have for any $1 \leq n \leq N_T$,

$$T_n - T_{n-1} = \tau N_S \left(f(\mathbf{z}_n^{(T)}) - \mathrm{E}^{(T)} f \right). \tag{69}$$

We are now ready to prove Theorem C.3.

Proof of Theorem C.3. According to (64) and (66), we have

$$\begin{aligned}
F_\tau(\mathbf{Z}_1^N) - \mathrm{E}^{(*)} F_\tau &= F_S(\mathbf{Z}_1^{N_S}) + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(*)} \{F_S + F_T\} \\
&= F_S(\mathbf{Z}_1^{N_S}) - \mathrm{E}^{(S)} F_S + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(T)} F_T.
\end{aligned} \tag{70}$$

According to Lemma B.2, (68), (69), (70), Markov's inequality, Jensen's inequality and the law of iterated expectation, we have for any $\alpha > 0$ and any $\tau \in [0, 1)$,

$$\begin{aligned}
&\Pr\left\{ F_\tau(\mathbf{Z}_1^N) - \mathrm{E}^{(*)} F_\tau > \xi \right\} \\
&= \Pr\left\{ F_S(\mathbf{Z}_1^{N_S}) - \mathrm{E}^{(S)} F_S + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(T)} F_T > \xi \right\} \\
&\leq \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(F_S(\mathbf{Z}_1^{N_S}) - \mathrm{E}^{(S)} F_S + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(T)} F_T \right)} \right\} \\
&= \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S}(S_n - S_{n-1}) + \sum_{n=1}^{N_T}(T_n - T_{n-1}) \right)} \right\} \\
&= \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S}(S_n - S_{n-1}) + \sum_{n=1}^{N_T}(T_n - T_{n-1}) \right)} \Big| \mathbf{Z}_1^{N_S - 1} \right\} \right\} \\
&= \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S - 1}(S_n - S_{n-1}) + \sum_{n=1}^{N_T}(T_n - T_{n-1}) \right)} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(S_{N_S} - S_{N_S-1} \right)} \Big| \mathbf{Z}_1^{N_S - 1} \right\} \right\} \\
&\leq \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S - 1}(S_n - S_{n-1}) + \sum_{n=1}^{N_T}(T_n - T_{n-1}) \right)} \right\} \mathrm{e}^{\frac{(1-\tau)^2 N_T^2 \alpha^2 (b-a)^2}{8}} \\
&= \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S - 1}(S_n - S_{n-1}) + \sum_{n=1}^{N_T - 1}(T_n - T_{n-1}) \right)} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(T_{N_T} - T_{N_T-1} \right)} \Big| \overline{\mathbf{Z}}_1^{N_T - 1} \right\} \right\} \\
&\quad \times \mathrm{e}^{\frac{(1-\tau)^2 N_T^2 \alpha^2 (b-a)^2}{8}} \\
&\leq \mathrm{e}^{-\alpha\xi} \mathrm{E}\left\{ \mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S - 1}(S_n - S_{n-1}) + \sum_{n=1}^{N_T - 1}(T_n - T_{n-1}) \right)} \right\} \mathrm{e}^{\frac{\tau^2 N_S^2 \alpha^2 (b-a)^2}{8}} \\
&\quad \times \mathrm{e}^{\frac{(1-\tau)^2 N_T^2 \alpha^2 (b-a)^2}{8}}.
\end{aligned} \tag{71}$$

Then, we have

$$\Pr\left\{ F_\tau\left(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T} \right) - \mathrm{E}^{(*)} F_\tau > \xi \right\} \leq \mathrm{e}^{\Phi(\alpha) - \alpha\xi}, \tag{72}$$

where

$$\Phi(\alpha) = \frac{\alpha^2 (1-\tau)^2 (b-a)^2 N_S N_T^2}{8} + \frac{\alpha^2 \tau^2 (b-a)^2 N_S^2 N_T}{8}. \tag{73}$$

Similarly, we can arrive at

$$\Pr\left\{ \mathrm{E}^{(*)} F_\tau - F_\tau\left(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T} \right) > \xi \right\} \leq \mathrm{e}^{\Phi(\alpha) - \alpha\xi}. \tag{74}$$

Note that "$\Phi(\alpha) - \alpha\xi$" is a quadratic function with respect to $\alpha > 0$ and thus the minimum value

$$\min_{\alpha > 0} \{ \Phi(\alpha) - \alpha\xi \}$$

is achieved when

$$\alpha = \frac{4\xi}{(b-a)^2 N_S N_T \left((1-\tau)^2 N_T + \tau^2 N_S \right)}.$$

By combining (72), (73) and (74), we arrive at

$$\Pr\left\{ |F_\tau\left(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T} \right) - \mathrm{E}^{(*)} F_\tau| > \xi \right\} \leq 2 \exp\left\{ -\frac{2\xi^2}{(b-a)^2 N_S N_T \left((1-\tau)^2 N_T + \tau^2 N_S \right)} \right\}.$$

This completes the proof. ∎

23

C.6.2 Symmetrization Inequality

In the following theorem, we present the symmetrization inequality for domain adaptation combining source and target data.

Theorem C.4 *Assume that $\mathcal{F}$ is a function class with the range $[a, b]$. Let $\mathbf{Z}_1^{N_S}$ and $\mathbf{Z}'_1^{N_S}$ be drawn from the source domain $\mathcal{Z}^{(S)}$, and $\overline{\mathbf{Z}}_1^{N_T}$ and $\overline{\mathbf{Z}}'_1^{N_T}$ be drawn from the target domain $\mathcal{Z}^{(T)}$. Then, for any $\tau \in [0, 1)$ and given an arbitrary $\xi > (1 - \tau)D_{\mathcal{F}}(S, T)$, we have for any $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$,*

$$\Pr\left\{ \sup_{f \in \mathcal{F}} \left| E^{(T)}f - E_\tau f \right| > \xi \right\} \leq 2\Pr\left\{ \sup_{f \in \mathcal{F}} \left| E'_\tau f - E_\tau f \right| > \frac{\xi'}{2} \right\} \tag{75}$$

with $\xi' = \xi - (1 - \tau)D_{\mathcal{F}}(S, T)$.

This theorem shows that for any $\xi > (1 - \tau)D_{\mathcal{F}}(S, T)$, the probability of the event:

$$\sup_{f \in \mathcal{F}} \left| E^{(T)}f - E_\tau f \right| > \xi$$

can be bounded by using the probability of the event:

$$\sup_{f \in \mathcal{F}} \left| E'_\tau f - E_\tau f \right| > \frac{\xi'}{2}$$

that is only determined by the samples drawn from the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$, when $N_S N_T \geq 8(b-a)^2/(\xi')^2$. Compared to the classical symmetrization result under the assumption of same distribution (*cf.* [2]), there is a discrepancy term $(1 - \tau)D_{\mathcal{F}}(S, T)$. The two results will coincide when the source and the target domains match, *i.e.*, $D_{\mathcal{F}}(S, T) = 0$. The following is the proof of Theorem C.4.

Proof of Theorem C.4. Let $\widehat{f}$ be the function achieving the supremum:

$$\sup_{f \in \mathcal{F}} |E^{(T)}f - E_\tau f|$$

with respect to $\mathbf{Z}_1^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T}$. According to (8), and (55), we arrive at

$$\begin{aligned}
\left| E^{(T)}\widehat{f} - E_\tau \widehat{f} \right| &= \left| \tau E^{(T)}\widehat{f} + (1-\tau)E^{(T)}\widehat{f} - (1-\tau)E^{(S)}\widehat{f} + (1-\tau)E^{(S)}\widehat{f} - E_\tau \widehat{f} \right| \\
&= \left| \tau(E^{(T)}\widehat{f} - E_{N_T}^{(T)}\widehat{f}) + (1-\tau)(E^{(T)}\widehat{f} - E^{(S)}\widehat{f}) + (1-\tau)(E^{(S)}\widehat{f} - E_{N_S}^{(S)}\widehat{f}) \right| \\
&\leq (1-\tau)D_{\mathcal{F}}(S, T) + \left| \tau(E^{(T)}\widehat{f} - E_{N_T}^{(T)}\widehat{f}) + (1-\tau)(E^{(S)}\widehat{f} - E_{N_S}^{(S)}\widehat{f}) \right|, \tag{76}
\end{aligned}$$

and thus

$$\begin{aligned}
&\Pr\left\{ \left| E^{(T)}\widehat{f} - E_\tau \widehat{f} \right| > \xi \right\} \\
&\leq \Pr\left\{ (1-\tau)D_{\mathcal{F}}(S, T) + \left| \tau(E^{(T)}\widehat{f} - E_{N_T}^{(T)}\widehat{f}) + (1-\tau)(E^{(S)}\widehat{f} - E_{N_S}^{(S)}\widehat{f}) \right| > \xi \right\}, \tag{77}
\end{aligned}$$

where

$$E_{N_T}^{(T)}\widehat{f} := \frac{1}{N_T} \sum_{n=1}^{N_T} \widehat{f}(\mathbf{z}_n^{(T)});$$

$$E_{N_S}^{(S)}\widehat{f} := \frac{1}{N_S} \sum_{n=1}^{N_S} \widehat{f}(\mathbf{z}_n^{(S)}). \tag{78}$$

Let

$$\xi' = \xi - (1 - \tau)D_{\mathcal{F}}(S, T) \tag{79}$$

24

and denote $\wedge$ as the conjunction of two events. According to the triangle inequality, we have

$$\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})|>\xi'\right\}}\right)\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})|<\frac{\xi'}{2}\right\}}\right)$$

$$=\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})|>\xi'\right\}\wedge\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})|<\frac{\xi'}{2}\right\}}$$

$$\leq\mathbf{1}_{\left\{|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})|>\frac{\xi'}{2}\right\}}.$$

Then, taking the expectation with respect to $\mathbf{Z}'^{N_S}_1$ and $\overline{\mathbf{Z}}'^{N_T}_1$ gives

$$\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})|>\xi'\right\}}\right)$$

$$\times \Pr'\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})\right|<\frac{\xi'}{2}\right\}$$

$$\leq \Pr'\left\{\left|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})\right|>\frac{\xi'}{2}\right\}. \tag{80}$$

By Chebyshev's inequality, since $\mathbf{Z}'^{N_S}_1 = \{\mathbf{z}'^{(S)}_n\}^{N_S}_{n=1}$ and $\overline{\mathbf{Z}}'^{N_T}_1 = \{\mathbf{z}'^{(T)}_n\}^{N_T}_{n=1}$ are sets of i.i.d. samples drawn from the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ respectively, we have for any $\xi' > 0$ and any $\tau \in [0, 1)$,

$$\Pr'\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})\right|\geq\frac{\xi'}{2}\right\}$$

$$\leq\Pr'\left\{\frac{\tau}{N_T}\sum_{n=1}^{N_T}|\mathrm{E}^{(T)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(T)}_n)|+\frac{1-\tau}{N_S}\sum_{n=1}^{N_S}|\mathrm{E}^{(S)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(S)}_n)|\geq\frac{\xi'}{2}\right\}$$

$$\leq\frac{4\mathrm{E}\left\{\tau N_S N_T(\mathrm{E}^{(T)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(T)}))^2+(1-\tau)N_S N_T(\mathrm{E}^{(S)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(S)}))^2\right\}}{N_S^2 N_T^2(\xi')^2}$$

$$\leq\frac{4\mathrm{E}\left\{\tau N_S N_T(b-a)^2+(1-\tau)N_S N_T(b-a)^2\right\}}{N_S^2 N_T^2(\xi')^2}$$

$$=\frac{4(b-a)^2}{N_S N_T(\xi')^2}, \tag{81}$$

where $\mathbf{z}'^{(S)}$ and $\mathbf{z}'^{(T)}$ stand for the ghost random variables taking values from the domain source $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$, respectively.

Subsequently, according to (80) and (81), we have for any $\xi' > 0$,

$$\Pr'\left\{\left|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})\right|>\frac{\xi'}{2}\right\}$$

$$\geq\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})|>\xi'\right\}}\right)\left(1-\frac{4(b-a)^2}{N_S N_T(\xi')^2}\right). \tag{82}$$

According to (77), (79) and (82), by letting

$$\frac{4(b-a)^2}{N_S N_T(\xi')^2}\leq\frac{1}{2},$$

and taking the expectation with respect to $\mathbf{Z}^{N_S}_1$ and $\overline{\mathbf{Z}}^{N_T}_1$, we have for any $\xi' > 0$,

$$\Pr\left\{|\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}_\tau\widehat{f}|>\xi\right\}$$

$$\leq\Pr\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})\right|>\xi'\right\}$$

$$\leq2\Pr\left\{\left|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f}-\mathrm{E}^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f}-\mathrm{E}^{(S)}_{N_S}\widehat{f})\right|>\frac{\xi'}{2}\right\}$$

$$=2\Pr\left\{\left|\mathrm{E}'_\tau\widehat{f}-\mathrm{E}_\tau\widehat{f}\right|>\frac{\xi'}{2}\right\} \tag{83}$$

25

with $\xi' = \xi - (1-\tau)D_{\mathcal{F}}(S,T)$. This completes the proof. ∎

We are now ready to prove Theorem C.1.

C.6.3 Proof of Theorem C.1

Proof of Theorem C.1. Consider $\{\epsilon_n\}_{n=1}^N$ as independent Rademacher random variables, *i.e.*, independent $\{\pm 1\}$-valued random variables with equal probability of taking either value. Given $\{\epsilon_n\}_{n=1}^{N_S}$, $\{\epsilon_n\}_{n=1}^{N_T}$, $\mathbf{Z}_1^{2N_S}$ and $\overline{\mathbf{Z}}_1^{2N_T}$, denote

$$\overrightarrow{\epsilon}_S := (\epsilon_1, \cdots, \epsilon_{N_S}, -\epsilon_1, \cdots, -\epsilon_{N_S}) \in \{\pm 1\}^{2N_S};$$
$$\overrightarrow{\epsilon}_T := (\epsilon_1, \cdots, \epsilon_{N_T}, -\epsilon_1, \cdots, -\epsilon_{N_T}) \in \{\pm 1\}^{2N_T}, \tag{84}$$

and for any $f \in \mathcal{F}$,

$$\overrightarrow{f}(\mathbf{Z}_1^{2N_S}) := (f(\mathbf{z}_1'), \cdots, f(\mathbf{z}_{N_S}'), f(\mathbf{z}_1), \cdots, f(\mathbf{z}_{N_S})) \in [a,b]^{2N_S};$$
$$\overrightarrow{f}(\mathbf{Z}_1^{2N_T}) := (f(\mathbf{z}_1'), \cdots, f(\mathbf{z}_{N_T}'), f(\mathbf{z}_1), \cdots, f(\mathbf{z}_{N_T})) \in [a,b]^{2N_T}. \tag{85}$$

We also denote

$$\mathbf{Z} := \left\{ \overline{\mathbf{Z}}_1^{2N_T}, \mathbf{Z}_1^{2N_S} \right\} \in \left(\mathcal{Z}^{(T)} \right)^{2N_T} \times \left(\mathcal{Z}^{(S)} \right)^{2N_S};$$
$$\overrightarrow{\epsilon} := \Big(\underbrace{\overrightarrow{\epsilon}_T, \cdots, \overrightarrow{\epsilon}_T}_{N_S}, \underbrace{\overrightarrow{\epsilon}_S, \cdots, \overrightarrow{\epsilon}_S}_{N_T} \Big) \in \{\pm 1\}^{4N_S N_T};$$
$$\overrightarrow{f}(\mathbf{Z}) := \Big(\underbrace{\overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T}), \cdots, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T})}_{N_S}, \underbrace{\overrightarrow{f}(\mathbf{Z}_1^{2N_S}), \cdots, \overrightarrow{f}(\mathbf{Z}_1^{2N_S})}_{N_T} \Big) \in [a,b]^{4N_S N_T}. \tag{86}$$

According to (6), (79) and Theorem C.4, for any $\tau \in [0,1)$ and given an arbitrary $\xi > (1-\tau)D_{\mathcal{F}}(S,T)$, we have for any $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - (1-\tau)D_{\mathcal{F}}(S,T)$,

$$\Pr\left\{ \sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_\tau f \right| > \xi \right\}$$

$$\leq 2\Pr\left\{ \sup_{f \in \mathcal{F}} \left| \mathrm{E}'_\tau f - \mathrm{E}_\tau f \right| > \frac{\xi'}{2} \right\} \qquad \text{(by Theorem C.4)}$$

$$= 2\Pr\left\{ \sup_{f \in \mathcal{F}} \left| \frac{\tau}{N_T} \sum_{n=1}^{N_T} \left(f(\mathbf{z}_n'^{(T)}) - f(\mathbf{z}_n^{(T)}) \right) + \frac{1-\tau}{N_S} \sum_{n=1}^{N_S} \left(f(\mathbf{z}_n'^{(S)}) - f(\mathbf{z}_n^{(S)}) \right) \right| > \frac{\xi'}{2} \right\}$$

$$= 2\Pr\left\{ \sup_{f \in \mathcal{F}} \left| \frac{\tau}{N_T} \sum_{n=1}^{N_T} \epsilon_n \left(f(\mathbf{z}_n'^{(T)}) - f(\mathbf{z}_n^{(T)}) \right) + \frac{1-\tau}{N_S} \sum_{n=1}^{N_S} \epsilon_n \left(f(\mathbf{z}_n'^{(S)}) - f(\mathbf{z}_n^{(S)}) \right) \right| > \frac{\xi'}{2} \right\}$$

$$= 2\Pr\left\{ \sup_{f \in \mathcal{F}} \left| \frac{\tau}{2N_T} \langle \overrightarrow{\epsilon}_T, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T}) \rangle + \frac{1-\tau}{2N_S} \langle \overrightarrow{\epsilon}_S, \overrightarrow{f}(\mathbf{Z}_1^{2N_S}) \rangle \right| > \frac{\xi'}{4} \right\}. \tag{87}$$

Given a $\tau \in [0,1)$, fix a realization of $\mathbf{Z}$ and let Λ be a $\xi'/8$-radius cover of $\mathcal{F}$ with respect to the $\ell_1^\tau(\mathbf{Z})$ norm. Since $\mathcal{F}$ is composed of the bounded functions with the range $[a,b]$, we assume that the same holds for any $h \in \Lambda$. If f_0 is the function that achieves the following supremum

$$\sup_{f \in \mathcal{F}} \left| \frac{\tau}{2N_T} \langle \overrightarrow{\epsilon}_T, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T}) \rangle + \frac{1-\tau}{2N_S} \langle \overrightarrow{\epsilon}_S, \overrightarrow{f}(\mathbf{Z}_1^{2N_S}) \rangle \right| > \frac{\xi'}{4},$$

there must be an $h_0 \in \Lambda$ that satisfies that

$$\frac{\tau}{2N_T} \sum_{n=1}^{N_T} \left(|f_0(\mathbf{z}_n'^{(T)}) - h_0(\mathbf{z}_n'^T)| + |f_0(\mathbf{z}_n^{(T)}) - h_0(\mathbf{z}_n^{(T)})| \right)$$

$$+ \frac{1-\tau}{2N_S} \sum_{n=1}^{N_S} \left(|f_0(\mathbf{z}_n'^{(S)}) - h_0(\mathbf{z}_n'^S)| + |f_0(\mathbf{z}_n^{(S)}) - h_0(\mathbf{z}_n^{(S)})| \right) < \frac{\xi'}{8},$$

26

and meanwhile,

$$\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{h}_0(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{h}_0(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{8}.$$

Therefore, for the realization of $\mathbf{Z}$, we arrive at

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{f}(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{4}\right\}$$

$$\leq\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{8}\right\}. \tag{88}$$

Moreover, we denote the event

$$A := \left\{\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{8}\right\}\right\},$$

and let $\mathbf{1}_A$ be the characteristic function of the event A. By Fubini's Theorem, we have

$$\Pr\{A\} = \mathrm{E}\left\{\mathrm{E}_{\overrightarrow{\epsilon}}\{\mathbf{1}_A\}\,\big|\,\mathbf{Z}\right\}$$

$$=\mathrm{E}\left\{\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{8}\right\}\,\big|\,\mathbf{Z}\right\}. \tag{89}$$

Fix a realization of $\mathbf{Z}$ again. According to (57), (84), (85) and Theorem C.3, for any $\tau\in[0,1)$ and given an arbitrary $\xi' > 0$, we have for any $N_S N_T \geq 8(b-a)^2/(\xi')^2$,

$$\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{8}\right\}$$

$$\leq|\Lambda|\max_{h\in\Lambda}\Pr\left\{\left|\frac{\tau}{2N_T}\langle\overrightarrow{\epsilon}_T,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\rangle + \frac{1-\tau}{2N_S}\langle\overrightarrow{\epsilon}_S,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\rangle\right| > \frac{\xi'}{8}\right\}$$

$$=\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\max_{h\in\Lambda}\Pr\left\{\left|\mathrm{E}'_\tau h - \mathrm{E}_\tau h\right| > \frac{\xi'}{4}\right\}$$

$$\leq\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\max_{h\in\Lambda}\Pr\left\{\left|\widetilde{\mathrm{E}}h - \mathrm{E}'_\tau h\right| + \left|\widetilde{\mathrm{E}}h - \mathrm{E}_\tau h\right| > \frac{\xi'}{4}\right\}$$

$$\leq 2\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\max_{h\in\Lambda}\Pr\left\{\left|\widetilde{\mathrm{E}}h - \mathrm{E}_\tau h\right| > \frac{\xi'}{8}\right\}$$

$$\leq 4\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\exp\left\{-\frac{N_S N_T\left(\xi - (1-\tau)D_\mathcal{F}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2 N_T + \tau^2 N_S\right)}\right\}, \tag{90}$$

where $\widetilde{\mathrm{E}}h := \tau\mathrm{E}^{(T)}h + (1-\tau)\mathrm{E}^{(S)}h$.

The combination of (58), (87), (88) and (90) leads to the following result: for any $\tau\in[0,1)$ and given an arbitrary $\xi > (1-\tau)D_\mathcal{F}(S,T)$, we have for any $N_S N_T \geq 8(b-a)^2/(\xi')^2$,

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_\tau f\right| > \xi\right\}$$

$$\leq 8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\exp\left\{-\frac{N_S N_T\left(\xi - (1-\tau)D_\mathcal{F}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2 N_T + \tau^2 N_S\right)}\right\}$$

$$\leq 8\mathcal{N}_1^\tau\left(\mathcal{F},\xi'/8,2(N_S+N_T)\right)\exp\left\{-\frac{N_S N_T\left(\xi - (1-\tau)D_\mathcal{F}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2 N_T + \tau^2 N_S\right)}\right\}. \tag{91}$$

According to (91), letting

$$\epsilon := 8\mathcal{N}_1^\tau\left(\mathcal{F},\xi'/8,2(N_S+N_T)\right)\exp\left\{-\frac{N_S N_T\left(\xi - (1-\tau)D_\mathcal{F}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2 N_T + \tau^2 N_S\right)}\right\},$$

27

we have given an arbitrary $\xi > (1-\tau)D_{\mathcal{F}}(S,T)$ and for any $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$, with probability at least $1 - \epsilon$,

$$\sup_{f \in \mathcal{F}} \left| \mathrm{E}_\tau f - \mathrm{E}^{(T)} f \right| \leq (1-\tau)D_{\mathcal{F}}(S,T) + \left(\frac{\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S + N_T)) - \ln(\epsilon/8)}{\frac{N_S N_T}{32(b-a)^2((1-\tau)^2 N_T + \tau^2 N_S)}} \right)^{\frac{1}{2}}.$$

This completes the proof. ∎

D   Comparison with Prior Works

There have been some previous works on the theoretical analysis of domain adaptation with multiple sources (*cf.* [13, 14, 15, 16, 17, 20]) and domain adaptation combining source and target data (*cf.* [13, 18]).

In [14, 15], the function class and the loss function are assumed to satisfy the conditions of "α-triangle inequality" and "uniform convergence bound". Moreover, one has to get some prior information about the disparity between any source domain and the target domain. Under these conditions, some generalization bounds were obtained by using the classical techniques developed under the assumption of same distribution.

Mansour *et al.* [16] proposed another framework to study the problem of domain adaptation with multiple sources. In this framework, one has to know some prior knowledge including the exact distributions of the source domains and the hypothesis function with a small loss on each source domain. Furthermore, the target domain and the hypothesis function on the target domain were deemed as the mixture of the source domains and the mixture of the hypothesis functions on the source domains, respectively. By introducing the Rényi divergence, Mansour *et al.* [17] extended their previous work [16] to a more general setting, where the distribution of the target domain can be arbitrary and one only needs to know an approximation of the exact distribution of each source domain. Ben-David *et al.* [13] also discussed the situation of domain adaptation with the mixture of source domains.

In [13, 18], domain adaptation combining source and target data was originally proposed and meanwhile a theoretical framework was presented to analyze its properties for the classification tasks by introducing the $\mathcal{H}$-divergence. Under the condition of "λ-close", the authors applied the classical techniques developed under the assumption of same distribution to achieve the generalization bounds based on the VC dimension.

Mansour *et al.* [20] introduced the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ to measure the difference between domains and this quantity can be used in both classification and regression tasks. By applying the classical results of statistical learning theory, the authors obtained the generalization bounds based on the Rademacher complexity.

The framework proposed in this paper is suitable for various kinds of tasks including classification and regression, because there is no assumption on the characteristics of domains and the function class except that the function class is composed of bounded functions.

We use the integral probability metric $D_{\mathcal{F}}(S,T)$ to measure the difference between $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$. We show that this quantity actually is a (semi)metric for any non-trivial function class $\mathcal{F}$ and can be bounded by the summation of the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ and the quantity $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$, which measure the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between labeling functions $g_*^{(S)}$ and $g_*^{(T)}$, respectively.

Instead of directly applying the classical techniques, based on the integral probability metric, the generalization bounds for the two types of domain adaptation are derived by using the specific Hoeffding-type deviation inequality and symmetrization inequality for the corresponding kind of domain adaptation, respectively. By the resultant generalization bounds, we can provide a rigorous theoretical analysis of the asymptotic convergence and the rate of convergence of the learning process for either kind of domain adaptation.

Based on the derived generalization bounds, we provide a rigorous theoretical analysis of the asymptotic convergence and the rate of convergence of the learning process for either kind of domain adap-

28

tation. We also consider the choices of $\mathbf{w}$ and τ that affect the rate of convergence of the learning processes for the two types of domain adaptation, respectively. The numerical experiments support our results as well.

29