SUPPLEMENTARY MATERIALS: A Unified Analysis of Generalization and Sample Complexity for Semi-Supervised Domain Adaptation*

Elif Vural† and Hüseyin Karaca‡

SM1. Technical lemmas and proofs.

Lemma SM1.1. *Consider that Assumption 2.1 holds. Let $\mathcal{L}_\alpha(f^s, f^t, h)$ denote the expected weighted loss in the source and target domains given by*

$$\mathcal{L}_\alpha(f^s, f^t, h) \triangleq (1-\alpha)\mathcal{L}^s(f^s, h) + \alpha\mathcal{L}^t(f^t, h).$$

Then the expected target loss is bounded as

$$\mathcal{L}^t(f^t, h) \leq \mathcal{L}_\alpha(f^s, f^t, h) + (1-\alpha)RD(f^s, f^t).$$

Proof. We have $\mathcal{L}^t(f^t, h) = \alpha\mathcal{L}^t(f^t, h) + (1-\alpha)\mathcal{L}^t(f^t, h)$. From Assumption 2.1, we get

$$\mathcal{L}^t(f^t, h) \leq \mathcal{L}^s(f^s, h) + R\ D(f^s, f^t).$$

Using this above, we obtain

$$\begin{aligned}\mathcal{L}^t(f^t, h) &\leq \alpha\mathcal{L}^t(f^t, h) + (1-\alpha)\left(\mathcal{L}^s(f^s, h) + R\ D(f^s, f^t)\right) \\ &= \mathcal{L}_\alpha(f^s, f^t, h) + (1-\alpha)RD(f^s, f^t).\end{aligned}$$

∎

Lemma SM1.2. *Let the conditions in Assumption 2.3 hold. Let*

$$\hat{\mathcal{L}}_\alpha(f^s, f^t, h) \triangleq (1-\alpha)\hat{\mathcal{L}}^s(f^s, h) + \alpha\hat{\mathcal{L}}^t(f^t, h)$$

denote the empirical weighted loss. Then, we have

$$\begin{aligned}&P\left(\sup_{f^s\in\mathcal{F}^s, f^t\in\mathcal{F}^t, h\in\mathcal{H}} |\mathcal{L}_\alpha(f^s, f^t, h) - \hat{\mathcal{L}}_\alpha(f^s, f^t, h)| \leq \epsilon\right) \\ &\geq 1 - 2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t)e^{-\frac{M_t\epsilon^2}{8\alpha^2 A_\ell^2}} - 2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s)e^{-\frac{M_s\epsilon^2}{8(1-\alpha)^2 A_\ell^2}}.\end{aligned}$$

*Submitted to the editors February 18, 2026.

Funding: This work is supported by The Scientific and Technological Research Council of Türkiye (TÜBİTAK) 1515 Frontier R&D Laboratories Support Program for Türk Telekom 6G R&D Lab under project number 5249902 and 2210 National Graduate Scholarship Program.

†Department of Electrical and Electronics Engineering, METU, Ankara, Türkiye (velif@metu.edu.tr, http://blog.metu.edu.tr/velif/).

‡Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Türkiye and Türk Telekom, Ankara, Türkiye (huseyin.karaca@bilkent.edu.tr, hkaraca@turktelekom.com.tr, https://huseyin-karaca.github.io).

Proof. We characterize the complexity of function spaces via covering numbers [SM9]. We first derive a bound for the deviation between the expected and empirical target losses. Let the open balls of radius $\frac{\epsilon}{8\alpha L_\ell}$ around the functions $\{g_k^t\}_{k=1}^{\kappa^t}$ be a cover for the function space $\mathcal{H}\circ\mathcal{F}^t$ with covering number

$$\kappa^t = \mathcal{N}(\mathcal{H}\circ\mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t).$$

Take any $g_k^t = h_k \circ f_k^t$, for $k = 1, \dots, \kappa^t$. The random variables $\ell(g_k^t(x_j^t), \mathbf{y}_j^t)$, $j = 1, \dots, M_t$ are independent identically distributed, bounded as $|\ell(g_k^t(x_j^t), \mathbf{y}_j^t)| \leq A_\ell$, and they have mean $\mathcal{L}^t(f_k^t, h_k)$. From Hoeffding's inequality, we get that for each k, the deviation between the empirical loss and the expected loss is bounded as

$$P\left(|\hat{\mathcal{L}}^t(f_k^t, h_k) - \mathcal{L}^t(f_k^t, h_k)| \geq \frac{\epsilon}{4\alpha}\right) \leq 2e^{-\frac{M_t\epsilon^2}{8\alpha^2 A_\ell^2}}.$$

Then, from union bound, with probability at least $1 - 2\kappa^t e^{-\frac{M_t\epsilon^2}{8\alpha^2 A_\ell^2}}$, the inequality

$$|\hat{\mathcal{L}}^t(f_k^t, h_k) - \mathcal{L}^t(f_k^t, h_k)| \leq \frac{\epsilon}{4\alpha}$$

holds for all $k = 1, \dots, \kappa^t$. Now for any $g^t = h \circ f^t \in \mathcal{H}\circ\mathcal{F}^t$, there exists at least one g_k^t such that

$$\mathfrak{d}^t(g^t, g_k^t) < \frac{\epsilon}{8\alpha L_\ell}.$$

This gives

$$\begin{aligned}
&|\mathcal{L}^t(f^t, h) - \mathcal{L}^t(f_k^t, h_k)| = \left|\int_{\mathcal{Z}^t} \left(\ell(g^t(x^t), \mathbf{y}^t) - \ell(g_k^t(x^t), \mathbf{y}^t)\right)\, d\mu_t\right| \\
&\leq \int_{\mathcal{Z}^t} \left|\ell(g^t(x^t), \mathbf{y}^t) - \ell(g_k^t(x^t), \mathbf{y}^t)\right|\, d\mu_t \leq \int_{\mathcal{Z}^t} L_\ell \|g^t(x^t) - g_k^t(x^t)\|\, d\mu_t \\
&\leq L_\ell \int_{\mathcal{Z}^t} \mathfrak{d}^t(g^t, g_k^t)\, d\mu_t < \frac{\epsilon}{8\alpha}.
\end{aligned}$$

It is easy to show similarly that

$$|\hat{\mathcal{L}}^t(f^t, h) - \hat{\mathcal{L}}^t(f_k^t, h_k)| < \frac{\epsilon}{8\alpha}.$$

Then with probability at least

$$1 - 2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t)e^{-\frac{M_t\epsilon^2}{8\alpha^2 A_\ell^2}}$$

for any $g^t = h \circ f^t \in \mathcal{H}\circ\mathcal{F}^t$ we have

$$\begin{aligned}
&|\mathcal{L}^t(f^t, h) - \hat{\mathcal{L}}^t(f^t, h)| \\
&\leq |\mathcal{L}^t(f^t, h) - \mathcal{L}^t(f_k^t, h_k)| + |\mathcal{L}^t(f_k^t, h_k) - \hat{\mathcal{L}}^t(f_k^t, h_k) + |\hat{\mathcal{L}}^t(f_k^t, h_k) - \hat{\mathcal{L}}^t(f^t, h)| \\
&< \frac{\epsilon}{8\alpha} + \frac{\epsilon}{4\alpha} + \frac{\epsilon}{8\alpha} = \frac{\epsilon}{2\alpha}.
\end{aligned}$$

Replacing α with $1-\alpha$ and applying the same steps for the function space $\mathcal{H}\circ\mathcal{F}^s$, we similarly obtain that with probability at least

$$1-2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s,\frac{\epsilon}{8(1-\alpha)L_\ell},\mathfrak{d}^s)e^{-\frac{M_s\epsilon^2}{8(1-\alpha)^2A_\ell^2}}$$

the difference between the expected and empirical source losses is bounded for any f^s and h as

$$|\mathcal{L}^s(f^s,h)-\hat{\mathcal{L}}^s(f^s,h)|<\frac{\epsilon}{2(1-\alpha)}.$$

Combining these results, we get that with probability at least

$$1-2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t,\frac{\epsilon}{8\alpha L_\ell},\mathfrak{d}^t)e^{-\frac{M_t\epsilon^2}{8\alpha^2A_\ell^2}}-2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s,\frac{\epsilon}{8(1-\alpha)L_\ell},\mathfrak{d}^s)e^{-\frac{M_s\epsilon^2}{8(1-\alpha)^2A_\ell^2}} \tag{SM1.1}$$

the largest difference between the expected and empirical total weighted losses is bounded as

$$\begin{aligned}\sup_{f^s\in\mathcal{F}^s,f^t\in\mathcal{F}^t,h\in\mathcal{H}}&|\mathcal{L}_\alpha(f^s,f^t,h)-\hat{\mathcal{L}}_\alpha(f^s,f^t,h)|\\&\le\alpha\sup|\mathcal{L}^t(f^t,h)-\hat{\mathcal{L}}^t(f^t,h)|+(1-\alpha)\sup|\mathcal{L}^s(f^s,h)-\hat{\mathcal{L}}^s(f^s,h)|\\&\le\epsilon.\end{aligned}$$

Lemma SM1.3. *Let the source and target distributions and the transformations $f^s:\mathcal{X}^s\to\mathcal{X}$ and $f^t:\mathcal{X}^t\to\mathcal{X}$ be such that Assumption 2.5 holds. Also, for given $\epsilon>0$, let the number of source and target samples be such that*

$$N_s>\frac{\sigma_s^2}{\epsilon^2},\quad N_t>\frac{\sigma_t^2}{\epsilon^2}.$$

Then for the source domain we have

$$\begin{aligned}&P\left(\left\|\frac{1}{N_s}\sum_{i=1}^{N_s}f^s(x_i^s)-E[f^s(x^s)]\right\|\ge\epsilon\right)\\&\le\exp\left(-\frac{1}{8}\left(\frac{\sqrt{N_s}\epsilon}{\sigma_s}-1\right)^2\frac{1}{1+\left(\frac{\sqrt{N_s}\epsilon}{\sigma_s}-1\right)\frac{C_s}{2\sqrt{N_s}\sigma_s}}\right)\end{aligned} \tag{SM1.2}$$

and for the target domain we have

$$\begin{aligned}&P\left(\left\|\frac{1}{N_t}\sum_{j=1}^{N_t}f^t(x_j^t)-E[f^t(x^t)]\right\|\ge\epsilon\right)\\&\le\exp\left(-\frac{1}{8}\left(\frac{\sqrt{N_t}\epsilon}{\sigma_t}-1\right)^2\frac{1}{1+\left(\frac{\sqrt{N_t}\epsilon}{\sigma_t}-1\right)\frac{C_t}{2\sqrt{N_t}\sigma_t}}\right).\end{aligned} \tag{SM1.3}$$

Proof. Our proof is based on the following result by Yurinskii [SM43].

Theorem. *[SM43, Theorem 2.1] Let $\zeta_1, \dots, \zeta_N \in \mathcal{B}$ be independent random vectors, where $\mathcal{B}$ is a Banach space. Assume for all $i = 1, \dots, N$*

$$E[\|\zeta_i\|^k] \leq \frac{k!}{2} b_i^2 C^{k-2}, \text{ for } k = 2, 3, \cdots . \tag{SM1.4}$$

If $x > \beta_N / B_N$ where

$$\beta_N \geq E[\|\zeta_1 + \cdots + \zeta_N\|], \quad B_N^2 = b_1^2 + \cdots + b_N^2, \tag{SM1.5}$$

then

$$P\left(\|\zeta_1 + \cdots + \zeta_N\| \geq x B_N\right) \leq \exp\left(-\frac{1}{8}\left(x - \frac{\beta_N}{B_N}\right)^2 \frac{1}{1 + \left(x - \frac{\beta_N}{B_N}\right)\frac{C}{2B_N}}\right).$$

Based on [SM43, Theorem 2.1], we first derive the stated result for the source domain, whose generalization to the target domain is straightforward. First notice that, due to the assumptions (2.8), (2.9), the random vectors $f^s(x_i^s) - E[f^s(x^s)]$ for $i = 1, \dots, N_s$ satisy the condition (SM1.4), for the choices $b_i = \sigma_s$ and $C = C_s$.

Next, we derive a constant β_{N_s} for which the zero-mean random vectors $\zeta_i = f^s(x_i^s) - E[f^s(x^s)]$ for $i = 1, \dots, N_s$ satisfy the condition (SM1.5) for $N = N_s$. From (2.8), we have

$$E[\|\zeta_i\|^2] \leq \sigma_s^2.$$

We consider now

$$\begin{aligned} E\left[\left\|\sum_{i=1}^{N_s} \zeta_i\right\|^2\right] &= E\left[\left\langle \sum_{i=1}^{N_s} \zeta_i, \sum_{j=1}^{N_s} \zeta_j \right\rangle\right] = \sum_{i=1}^{N_s}\sum_{j=1}^{N_s} E[\langle \zeta_i, \zeta_j \rangle] \\ &= \sum_{i=1}^{N_s} E[\langle \zeta_i, \zeta_i \rangle] + \sum_{i=1}^{N_s} \sum_{j \neq i,\, j=1}^{N_s} E[\langle \zeta_i, \zeta_j \rangle] \leq \sigma_s^2 N_s \end{aligned}$$

where the last inequality follows from $E[\|\zeta_i\|^2] \leq \sigma_s^2$, and the fact that we have $E[\langle \zeta_i, \zeta_j \rangle] = 0$ for independent and zero-mean ζ_i and ζ_j for $i \neq j$. From the nonnegativity of the variance, we have $(E[Y])^2 \leq E[Y^2]$ for any random variable Y. Taking

$$Y = \left\|\sum_{i=1}^{N_s} \zeta_i\right\|$$

then yields

$$E\left[\left\|\sum_{i=1}^{N_s} \zeta_i\right\|\right] \leq \left(E\left[\left\|\sum_{i=1}^{N_s} \zeta_i\right\|^2\right]\right)^{1/2} \leq \sigma_s \sqrt{N_s}.$$

Hence defining $\beta_{N_s} = \sigma_s\sqrt{N_s}$, we get

$$E[\|\zeta_1 + \cdots + \zeta_{N_s}\|] \leq \beta_{N_s}. \tag{SM1.6}$$

From the choice $b_i = \sigma_s$, we have $B_{N_s} = \sqrt{N_s}\sigma_s = \beta_{N_s}$. Now for given $\epsilon > 0$, from the assumption $N_s > \sigma_s^2/\epsilon^2$, the following choice for x

$$x = \frac{\sqrt{N_s}\epsilon}{\sigma_s} > 1$$

satisfies the condition $x > \beta_{N_s}/B_{N_s}$ as $\beta_{N_s} = B_{N_s}$. Then from [SM43, Theorem 2.1], we have

$$P\left(\|\zeta_1 + \cdots + \zeta_{N_s}\| \geq N_s\epsilon\right) \leq \exp\left(-\frac{1}{8}\left(\frac{\sqrt{N_s}\epsilon}{\sigma_s} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s}\epsilon}{\sigma_s} - 1\right)\frac{C_s}{2\sqrt{N_s}\sigma_s}}\right).$$

Replacing $\zeta_i = f^s(x_i^s) - E[f^s(x^s)]$ gives the stated result

$$P\left(\left\|\frac{1}{N_s}\sum_{i=1}^{N_s} f^s(x_i^s) - E[f^s(x^s)]\right\| \geq \epsilon\right)$$
$$\leq \exp\left(-\frac{1}{8}\left(\frac{\sqrt{N_s}\epsilon}{\sigma_s} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s}\epsilon}{\sigma_s} - 1\right)\frac{C_s}{2\sqrt{N_s}\sigma_s}}\right).$$

Applying the same analysis for the target domain, it is easy to show similarly that the upper bound for the target domain in (SM1.3) also holds. ∎

Lemma SM1.4. *Let Assumptions 2.5, 2.6 hold. Given $\epsilon > 0$, let the number of source and target samples be such that*

$$N_s > \frac{16\sigma_s^2}{\epsilon^2}, \quad N_t > \frac{16\sigma_t^2}{\epsilon^2}.$$

Let us define the functions

$$a_s(N_s, \epsilon) \triangleq \frac{1}{8}\left(\frac{\sqrt{N_s}\epsilon}{4\sigma_s} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s}\epsilon}{4\sigma_s} - 1\right)\frac{C_s}{2\sqrt{N_s}\sigma_s}}$$
$$a_t(N_t, \epsilon) \triangleq \frac{1}{8}\left(\frac{\sqrt{N_t}\epsilon}{4\sigma_t} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_t}\epsilon}{4\sigma_t} - 1\right)\frac{C_t}{2\sqrt{N_t}\sigma_t}}.$$

Then

$$P\left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t} |D(f^s, f^t) - \hat{D}(f^s, f^t)| < \epsilon\right)$$
$$\geq 1 - \mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^s)\exp(-a_s(N_s, \epsilon)) - \mathcal{N}(\mathcal{F}^t, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^t)\exp(-a_t(N_t, \epsilon)).$$

Proof. We begin with bounding the deviation $|D(f^s, f^t) - \hat{D}(f^s, f^t)|$ between the MMD and its empirical estimate for a fixed pair of transformations. Let f^s and f^t be a given, fixed pair of transformations. We have

$$\begin{aligned} &|D(f^s, f^t) - \hat{D}(f^s, f^t)| \\ &= \left| \|E[f^s(x^s)] - E[f^t(x^t)]\| - \|\frac{1}{N_s}\sum_{i=1}^{N_s} f^s(x_i^s) - \frac{1}{N_t}\sum_{j=1}^{N_t} f^t(x_j^t)\| \right| \\ &\leq \|\frac{1}{N_s}\sum_{i=1}^{N_s} f^s(x_i^s) - E[f^s(x^s)]\| + \|\frac{1}{N_t}\sum_{j=1}^{N_t} f^t(x_j^t) - E[f^t(x^t)]\|. \end{aligned} \tag{SM1.7}$$

Replacing ϵ by $\epsilon/4$ in Lemma SM1.3, we observe that with probability at least

$$1 - \exp(-a_s(N_s, \epsilon)) - \exp(-a_t(N_t, \epsilon))$$

we have

$$\|\frac{1}{N_s}\sum_{i=1}^{N_s} f^s(x_i^s) - E[f^s(x^s)]\| \leq \frac{\epsilon}{4}, \quad \|\frac{1}{N_t}\sum_{j=1}^{N_t} f^t(x_j^t) - E[f^t(x^t)]\| \leq \frac{\epsilon}{4}$$

which yields from (SM1.7)

$$|D(f^s, f^t) - \hat{D}(f^s, f^t)| \leq \frac{\epsilon}{2}.$$

In order to extend the above bound to the whole space of transformations, we consider covers of the function classes $\mathcal{F}^s$ and $\mathcal{F}^t$, consisting of open balls of radius $\epsilon/8$ respectively around the functions $\{f_k^s\}_{k=1}^{\kappa^s}$ and $\{f_l^t\}_{l=1}^{\kappa^t}$, where κ^s and κ^t are the covering numbers

$$\kappa^s = \mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^s), \quad \kappa^t = \mathcal{N}(\mathcal{F}^t, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^t).$$

From the union bound, it follows that with probability at least

$$1 - \kappa^s \exp(-a_s(N_s, \epsilon)) - \kappa^t \exp(-a_t(N_t, \epsilon))$$

for all $k = 1, \ldots, \kappa^s$ and $l = 1, \ldots, \kappa^t$,

$$|D(f_k^s, f_l^t) - \hat{D}(f_k^s, f_l^t)| \leq \frac{\epsilon}{2}. \tag{SM1.8}$$

Now, let us consider an arbitrary pair of transformations $f^s \in \mathcal{F}^s$ and $f^t \in \mathcal{F}^t$. As the balls around $\{f_k^s\}_{k=1}^{\kappa^s}$ and $\{f_l^t\}_{l=1}^{\kappa^t}$ form $\epsilon/8$-covers of the function classes, there exists a source transformation f_k^s and a target transformation f_l^t such that

$$\mathfrak{d}_{\mathcal{X}}^s(f^s, f_k^s) < \frac{\epsilon}{8}, \quad \mathfrak{d}_{\mathcal{X}}^t(f^t, f_l^t) < \frac{\epsilon}{8}.$$

We can then bound the difference between the MMD and its sample mean for f^s and f^t as follows.

$$|D(f^s,f^t)-\hat{D}(f^s,f^t)| \le |D(f^s,f^t)-D(f^s_k,f^t_l)| + |D(f^s_k,f^t_l)-\hat{D}(f^s_k,f^t_l)| + |\hat{D}(f^s_k,f^t_l)-\hat{D}(f^s,f^t)| \tag{SM1.9}$$

Next, we bound each one of the terms on the right hand side of the above inequality. The first term can be upper bounded as

$$\begin{aligned} |D(f^s,f^t)-D(f^s_k,f^t_l)| &= \left| \|E[f^s(x^s)]-E[f^t(x^t)]\| - \|E[f^s_k(x^s)]-E[f^t_l(x^t)]\| \right| \\ &\le \|E[f^s(x^s)]-E[f^s_k(x^s)]\| + \|E[f^t(x^t)]-E[f^t_l(x^t)]\| \\ &= \|E[f^s(x^s)-f^s_k(x^s)]\| + \|E[f^t(x^t)-f^t_l(x^t)]\| \\ &\le E[\|f^s(x^s)-f^s_k(x^s)\|] + E[\|f^t(x^t)-f^t_l(x^t)\|] \end{aligned} \tag{SM1.10}$$

where the last inequality follows from Jensen's inequality, observing the fact that a norm over a Hilbert space is a convex function. From the definition of the metrics $\mathfrak{d}^s_{\mathcal{X}}$ and $\mathfrak{d}^t_{\mathcal{X}}$, we have

$$\begin{aligned} \|f^s(x^s)-f^s_k(x^s)\| &\le \mathfrak{d}^s_{\mathcal{X}}(f^s,f^s_k) \\ \|f^t(x^t)-f^t_l(x^t)\| &\le \mathfrak{d}^t_{\mathcal{X}}(f^t,f^t_l) \end{aligned}$$

for all $x^s \in \mathcal{X}^s$ and $x^t \in \mathcal{X}^t$. Using this in (SM1.10), we get

$$|D(f^s,f^t)-D(f^s_k,f^t_l)| \le \mathfrak{d}^s_{\mathcal{X}}(f^s,f^s_k) + \mathfrak{d}^t_{\mathcal{X}}(f^t,f^t_l) < \frac{\epsilon}{8}+\frac{\epsilon}{8} = \frac{\epsilon}{4}.$$

With a similar analysis by replacing the expectations with the sample means, it is easy to show that the third term in the inequality (SM1.9) can also be upper bounded as

$$|\hat{D}(f^s_k,f^t_l)-\hat{D}(f^s,f^t)| < \frac{\epsilon}{4}.$$

Now, remembering also the probabilistic upper bound (SM1.8) that holds for the second term in (SM1.9) for all k and l, we get that with probability at least

$$1-\kappa^s \exp(-a_s(N_s,\epsilon)) - \kappa^t \exp(-a_t(N_t,\epsilon))$$

we have for all $f^s \in \mathcal{F}^s$ and $f^t \in \mathcal{F}^t$,

$$|D(f^s,f^t)-\hat{D}(f^s,f^t)| < \frac{\epsilon}{4}+\frac{\epsilon}{2}+\frac{\epsilon}{4} = \epsilon.$$

Hence, we get the stated result

$$\begin{aligned} &P\left(\sup_{f^s\in\mathcal{F}^s, f^t\in\mathcal{F}^t} |D(f^s,f^t)-\hat{D}(f^s,f^t)| < \epsilon \right) \\ &\ge 1-\kappa^s \exp(-a_s(N_s,\epsilon)) - \kappa^t \exp(-a_t(N_t,\epsilon)) \\ &= 1-\mathcal{N}(\mathcal{F}^s,\frac{\epsilon}{8},\mathfrak{d}^s_{\mathcal{X}}) \exp(-a_s(N_s,\epsilon)) - \mathcal{N}(\mathcal{F}^t,\frac{\epsilon}{8},\mathfrak{d}^t_{\mathcal{X}}) \exp(-a_t(N_t,\epsilon)). \end{aligned}$$

Lemma SM1.5. *Let the condition in Assumption 3.2 hold. Then the mappings $f^{sl}: \mathcal{X}^s \to \mathcal{X}^l$ and $f^{tl}: \mathcal{X}^t \to \mathcal{X}^l$ for $l = 1, \ldots, L-1$, and the mappings $f^s: \mathcal{X}^s \to \mathcal{X}$ and $f^t: \mathcal{X}^t \to \mathcal{X}$ are measurable. Moreover, assuming that $E[\sqrt{k^l(\boldsymbol{\xi}^{sl}, \boldsymbol{\xi}^{sl})}] < \infty$ and $E[\sqrt{k^l(\boldsymbol{\xi}^{tl}, \boldsymbol{\xi}^{tl})}] < \infty$, the functions $E[f^{sl}(x^s)]: \mathbb{R}^{d_l} \to \mathbb{R}$ and $E[f^{tl}(x^t)]: \mathbb{R}^{d_l} \to \mathbb{R}$ defined as*

$$E[f^{sl}(x^s)](\cdot) \triangleq E[f^{sl}(x^s)(\cdot)]$$
$$E[f^{tl}(x^t)](\cdot) \triangleq E[f^{tl}(x^t)(\cdot)]$$

through the Borel probability measures μ_s and μ_t in the source and target domains are in the RKHSs $\mathcal{X}^l$. Consequently, the functions

$$E[f^s(x^s)] \triangleq (E[f^{s1}(x^s)], \ldots, E[f^{s(L-1)}(x^s)])$$
$$E[f^t(x^t)] \triangleq (E[f^{t1}(x^t)], \ldots, E[f^{t(L-1)}(x^t)])$$

are in the Hilbert space $\mathcal{X}$.

Proof. We prove the statements only for the source domain, as the proofs for the target domain are the same. Let $\boldsymbol{\xi}^{sl}(x^s) \in \mathbb{R}^{d_l}$ denote the feature in layer l for the source input $x^s \in \mathbb{R}^{d_0}$, where we regard $\boldsymbol{\xi}^{sl}(\cdot): \mathbb{R}^{d_0} \to \mathbb{R}^{d_l}$ as a function. In the relation

$$\boldsymbol{\xi}^{sl}(x^s) = \eta^l(\mathbf{W}^{sl}\boldsymbol{\xi}^{s(l-1)}(x^s) + \mathbf{b}^{sl})$$

the expression $\mathbf{W}^{sl}\boldsymbol{\xi}^{s(l-1)}(x^s) + \mathbf{b}^{sl}$ is a continuous mapping of $\boldsymbol{\xi}^{s(l-1)}(x^s)$, and the function η^l is continuous. Hence, based on a simple induction argument it follows that $\boldsymbol{\xi}^{sl}(\cdot): \mathcal{X}^s = \mathbb{R}^{d_0} \to \mathbb{R}^{d_l}$ is a continuous, thus measurable function (a Borel map).

We next show that the mappings $f^{sl}: \mathcal{X}^s \to \mathcal{X}^l$ are measurable. Let $\mathcal{B}(\cdot)$ denote the Borel σ-algebra of a metric space. We recall from (3.4) that $f^{sl}(x^s) = \phi^l(\boldsymbol{\xi}^{sl}(x^s)) \in \mathcal{X}^l$. Consider a sequence $\{\boldsymbol{\xi}_n^{sl}\} \subset \mathbb{R}^{d_l}$ with $\lim_{n\to\infty} \boldsymbol{\xi}_n^{sl} = \boldsymbol{\xi}_*^{sl}$ for some $\boldsymbol{\xi}_*^{sl} \in \mathbb{R}^{d_l}$. As the kernel $k^l(\cdot,\cdot)$ is assumed to be a continuous function, we have

$$\lim_{n\to\infty} \|\phi^l(\boldsymbol{\xi}_n^{sl}) - \phi^l(\boldsymbol{\xi}_*^{sl})\|_{\mathcal{X}^l}^2 = \lim_{n\to\infty} \left(k^l(\boldsymbol{\xi}_n^{sl}, \boldsymbol{\xi}_n^{sl}) - 2k^l(\boldsymbol{\xi}_n^{sl}, \boldsymbol{\xi}_*^{sl}) + k^l(\boldsymbol{\xi}_*^{sl}, \boldsymbol{\xi}_*^{sl})\right) = 0$$

where $\|\cdot\|_{\mathcal{X}^l}$ denotes the norm in the RKHS $\mathcal{X}^l$. It thus follows that

$$\lim_{n\to\infty} \phi^l(\boldsymbol{\xi}_n^{sl}) = \phi^l(\boldsymbol{\xi}_*^{sl})$$

and hence $\phi^l: \mathbb{R}^{d_l} \to \mathcal{X}^l$ is a continuous function. ϕ^l is thus measurable with respect to the Borel σ-algebra $\mathcal{B}(\mathcal{X}^l)$ of the RKHS $\mathcal{X}^l$. Since $\boldsymbol{\xi}^{sl}(\cdot): \mathcal{X}^s \to \mathbb{R}^{d_l}$ is a measurable mapping as well, we conclude that the mapping $f^{sl} = \phi^l(\boldsymbol{\xi}^{sl}(\cdot)): \mathcal{X}^s \to \mathcal{X}^l$ is measurable with respect to $\mathcal{B}(\mathcal{X}^l)$, for $l = 1, \ldots, L-1$.

We next show that the mappings $f^s \in \mathcal{F}^s$ are measurable. Since the kernel $k^l(\cdot,\cdot)$ is assumed to be continuous, the RKHS $\mathcal{X}^l$ is separable for all l [SM34]. The separability of the RKHSs ensures that

$$\mathcal{B}(\mathcal{X}) = \bigotimes_{l=1}^{L-1} \mathcal{B}(\mathcal{X}^l)$$

where the right hand side denotes the σ-algebra generated by all finite products of Borel sets in $\mathcal{B}(\mathcal{X}^l)$'s [SM7]. Hence, denoting the set product of some collection of Borel sets $B^1 \in \mathcal{B}(\mathcal{X}^1), \cdots, B^{L-1} \in \mathcal{B}(\mathcal{X}^{L-1})$ as

$$B^1 \times B^2 \times \cdots \times B^{L-1} = \{(f^1, f^2, \ldots, f^{L-1}) : f^l \in B^l,\, l = 1, \ldots, L-1\},$$

the σ-algebra generated by

$$B = \{B^1 \times \cdots \times B^{L-1} : B^1 \in \mathcal{B}(\mathcal{X}^1), \cdots, B^{L-1} \in \mathcal{B}(\mathcal{X}^{L-1})\}$$

is equal to the Borel σ-algebra $\mathcal{B}(\mathcal{X})$. Then, in order to show that $f^s : \mathcal{X}^s \to \mathcal{X}$ is measurable, it is sufficient to show that the inverse image $(f^s)^{-1}(B)$ of the set B is contained in $\mathcal{B}(\mathcal{X}^s)$. For any element $B^1 \times \cdots \times B^{L-1}$ in B, we have

$$\begin{aligned}(f^s)^{-1}(B^1 \times \cdots \times B^{L-1}) &= \{x^s \in \mathcal{X}^s : f^s(x^s) \in B^1 \times \cdots \times B^{L-1}\} \\ &= \{x^s \in \mathcal{X}^s : f^{s1}(x^s) \in B^1,\ \cdots\ , f^{s(L-1)}(x^s) \in B^{L-1}\} \\ &= \bigcap_{l=1}^{L-1} (f^{sl})^{-1}(B^l).\end{aligned}$$

Since each f^{sl} is measurable, $(f^{sl})^{-1}(B^l) \in \mathcal{B}(\mathcal{X}^s)$. Hence, $(f^s)^{-1}(B^1 \times \cdots \times B^{L-1}) \in \mathcal{B}(\mathcal{X}^s)$ and we conclude that $f^s : \mathcal{X}^s \to \mathcal{X}$ is a measurable mapping.

In order to prove the second part of the lemma, let us fix $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$, and for fixed $\boldsymbol{\xi}$ consider the function $f^{sl}(\cdot)(\boldsymbol{\xi}) : \mathcal{X}^s = \mathbb{R}^{d_0} \to \mathbb{R}$ given by

$$f^{sl}(\cdot)(\boldsymbol{\xi}) = k^l(\boldsymbol{\xi}^{sl}(\cdot), \boldsymbol{\xi}).$$

From the continuity of the kernel k^l and the measurability of the function $\boldsymbol{\xi}^{sl}(\cdot)$, it is easy to conclude that the function $f^{sl}(\cdot)(\boldsymbol{\xi})$ is measurable for any fixed $\boldsymbol{\xi}$. Hence, based on the Borel probability measure μ_s in the source domain, the expectation $E_{x^s}[f^{sl}(x^s)(\boldsymbol{\xi})]$ for fixed $\boldsymbol{\xi}$ is well defined, as well as the function $E_{x^s}[f^{sl}(x^s)] : \mathbb{R}^{d_l} \to \mathbb{R}$ given by

$$E_{x^s}[f^{sl}(x^s)](\boldsymbol{\xi}) \triangleq E_{x^s}[f^{sl}(x^s)(\boldsymbol{\xi})].$$

Next, we would like to show that $E_{x^s}[f^{sl}(x^s)] \in \mathcal{X}^l$. Consider the linear functional $T_{\mu_s} : \mathcal{X}^l \to \mathbb{R}$ on the RKHS $\mathcal{X}^l$ defined by

$$T_{\mu_s}(\psi) \triangleq E_{x^s}[\psi(\boldsymbol{\xi}^{sl})]$$

for $\psi \in \mathcal{X}^l$. Following the steps as in the proof of [SM20, Lemma 3], the linear functional T_{μ_s} is observed to be bounded since

$$\begin{aligned}|T_{\mu_s}(\psi)| = \left|E_{x^s}[\psi(\boldsymbol{\xi}^{sl})]\right| &\le E_{x^s}\left[|\psi(\boldsymbol{\xi}^{sl})|\right] = E_{x^s}\left[\left|\langle k^l(\boldsymbol{\xi}^{sl}, \cdot), \psi(\cdot)\rangle_{\mathcal{X}^l}\right|\right] \\ &\le E_{x^s}\left[\|k^l(\boldsymbol{\xi}^{sl}, \cdot)\|_{\mathcal{X}^l}\|\psi\|_{\mathcal{X}^l}\right] = E_{x^s}\left[\sqrt{k^l(\boldsymbol{\xi}^{sl}, \boldsymbol{\xi}^{sl})}\right]\|\psi\|_{\mathcal{X}^l}.\end{aligned}$$

Hence, by the Riesz Representation Theorem [SM2, Theorem 12.5],[SM20, Lemma 3], there exists an element $\psi^{sl} \in \mathcal{X}^l$ in the RKHS $\mathcal{X}^l$ (called the mean embedding), such that

$$T_{\mu_s}(\psi) = \langle \psi, \psi^{sl} \rangle_{\mathcal{X}^l}$$

for all $\psi \in \mathcal{X}^l$. In particular, setting $\psi = \phi^l(\boldsymbol{\xi})$ for an arbitrary $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$, we have

$$T_{\mu_s}(\phi^l(\boldsymbol{\xi})) = \langle \phi^l(\boldsymbol{\xi}), \psi^{sl} \rangle_{\mathcal{X}^l} = \psi^{sl}(\boldsymbol{\xi}). \tag{SM1.11}$$

But it also holds that

$$\begin{aligned} T_{\mu_s}(\phi^l(\boldsymbol{\xi})) &= E_{x^s}[\phi^l(\boldsymbol{\xi})(\boldsymbol{\xi}^{sl})] = E_{x^s}[k^l(\boldsymbol{\xi}, \boldsymbol{\xi}^{sl})] = E_{x^s}[k^l(\boldsymbol{\xi}^{sl}, \boldsymbol{\xi})] \\ &= E_{x^s}[\phi^l(\boldsymbol{\xi}^{sl})(\boldsymbol{\xi})] = E_{x^s}[f^{sl}(x^s)(\boldsymbol{\xi})] = E_{x^s}[f^{sl}(x^s)](\boldsymbol{\xi}). \end{aligned} \tag{SM1.12}$$

From the equality of the expressions in (SM1.11) and (SM1.12), we observe that

$$E_{x^s}[f^{sl}(x^s)] = \psi^{sl} \in \mathcal{X}^l.$$

It then simply follows from the construction of $\mathcal{X}$ that

$$E_{x^s}[f^s(x^s)] \triangleq (E_{x^s}[f^{s1}(x^s)], \dots, E_{x^s}[f^{s(L-1)}(x^s)])$$

■

is in the Hilbert space $\mathcal{X}$.

Lemma SM1.6. *Let Assumptions 3.1-3.3 hold. Then, the transformation function classes $\mathcal{F}^s, \mathcal{F}^t$ in (3.6) and the composite function classes $\mathcal{G}^s, \mathcal{G}^t$ in (3.8) are compact metric spaces, respectively under the metrics $\mathfrak{d}^s_{\mathcal{X}}, \mathfrak{d}^t_{\mathcal{X}}$ in (2.10), and the metrics $\mathfrak{d}^s, \mathfrak{d}^t$ in (2.4).*

Proof. We prove the statements only for $\mathcal{F}^s$ and $\mathcal{G}^s$ as the proofs for the target domain are similar. We first show that $\mathcal{F}^s$ is compact with respect to the metric $\mathfrak{d}^s_{\mathcal{X}}$. Let

$$\boldsymbol{\Phi}^s = \{\boldsymbol{\Theta}^s = (\boldsymbol{\Theta}^{s1}, \dots, \boldsymbol{\Theta}^{sL}) : \ |\boldsymbol{\Theta}^{sl}_{ij}| \leq A_\Theta, \forall i, j, l\}$$

denote the parameter space over which the source network parameters are defined. Regarding $\boldsymbol{\Phi}^s$ as the Cartesian product of the corresponding matrix spaces at layers $l = 1, \dots, L$, it follows from the bound $|\boldsymbol{\Theta}^{sl}_{ij}| \leq A_\Theta$ on the network parameters that the finite dimensional set $\boldsymbol{\Phi}^s$ is closed and bounded, hence compact.

We next define a mapping $\mathcal{M}_{\mathcal{F}^s} : \boldsymbol{\Phi}^s \to \mathcal{F}^s$ such that

$$\mathcal{M}_{\mathcal{F}^s}(\boldsymbol{\Theta}^s) = f^s_{\boldsymbol{\Theta}^s} = \big(f^{s1}_{\boldsymbol{\Theta}^s}, \dots, f^{s(L-1)}_{\boldsymbol{\Theta}^s}\big) \tag{SM1.13}$$

where the notation $f^s_{\boldsymbol{\Theta}^s}(x^s)$ stands for the function $f^s(x^s)$ defined in (3.6) by explicitly referring to its dependence on the network parameters $\boldsymbol{\Theta}^s$. In the following, we show that the mapping $\mathcal{M}_{\mathcal{F}^s}$ is continuous. Let us consider a sequence $\{\boldsymbol{\Theta}^s_n\} \subset \boldsymbol{\Phi}^s$ converging to an element $\boldsymbol{\Theta}^s_* \in \boldsymbol{\Phi}^s$. Since the relation (3.1) between the features of adjacent layers is given by a linear mapping

followed by a continuous activation function η^l, the mapping $\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s)$ is a continuous function of $\boldsymbol{\Theta}^s$, i.e.

$$\lim_{n\to\infty} \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s) = \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s). \tag{SM1.14}$$

In fact, due to the assumptions on the boundedness (3.2) of the source samples, the boundedness (3.3) of the network parameters, and the Lipschitz continuity (3.13) of the activation functions η^l, it is easy to show that the convergence in (SM1.14) is uniform on $\mathcal{X}^s$. Hence, for any given $\epsilon > 0$, one can find some n_0 such that for $n \geq n_0$, we have

$$\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s) - \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s)\| < \epsilon$$

for all $x^s \in \mathcal{X}^s$, for $l = 1, \ldots, L-1$. Then we have

$$\begin{aligned}
&\|f^{sl}_{\boldsymbol{\Theta}^s_n}(x^s) - f^{sl}_{\boldsymbol{\Theta}^s_*}(x^s)\|^2_{\mathcal{X}^l} = \|\phi^l(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s)) - \phi^l(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s))\|^2_{\mathcal{X}^l} \\
&= k^l\big(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s)\big) - 2k^l\big(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s)\big) + k^l\big(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s)\big) \\
&\leq 2L_K\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_n}(x^s) - \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_*}(x^s)\| < 2L_K\epsilon
\end{aligned}$$

for all $x^s \in \mathcal{X}^s$ due to the Lipschitz continuity of the kernels k^l. This gives

$$\|f^s_{\boldsymbol{\Theta}^s_n}(x^s) - f^s_{\boldsymbol{\Theta}^s_*}(x^s)\|^2_{\mathcal{X}} = \sum_{l=1}^{L-1} \|f^{sl}_{\boldsymbol{\Theta}^s_n}(x^s) - f^{sl}_{\boldsymbol{\Theta}^s_*}(x^s)\|^2_{\mathcal{X}^l} < 2(L-1)L_K\epsilon.$$

We have thus obtained

$$\|f^s_{\boldsymbol{\Theta}^s_n}(x^s) - f^s_{\boldsymbol{\Theta}^s_*}(x^s)\|_{\mathcal{X}} < \sqrt{2(L-1)L_K}\sqrt{\epsilon}$$

for all $n \geq n_0$ and for all $x^s \in \mathcal{X}^s$, which shows that $f^s_{\boldsymbol{\Theta}^s_n}(x^s)$ converges to $f^s_{\boldsymbol{\Theta}^s_*}(x^s)$ uniformly on $\mathcal{X}^s$. Then we have

$$\begin{aligned}
\lim_{n\to\infty} \mathfrak{d}^s_{\mathcal{X}}(f^s_{\boldsymbol{\Theta}^s_n}, f^s_{\boldsymbol{\Theta}^s_*}) &= \lim_{n\to\infty} \sup_{x^s\in\mathcal{X}^s} \|f^s_{\boldsymbol{\Theta}^s_n}(x^s) - f^s_{\boldsymbol{\Theta}^s_*}(x^s)\|_{\mathcal{X}} \\
&= \sup_{x^s\in\mathcal{X}^s} \lim_{n\to\infty} \|f^s_{\boldsymbol{\Theta}^s_n}(x^s) - f^s_{\boldsymbol{\Theta}^s_*}(x^s)\|_{\mathcal{X}} = 0
\end{aligned}$$

where the second equality follows from the uniform convergence of $f^s_{\boldsymbol{\Theta}^s_n}(x^s)$. We have thus shown that the mapping $\mathcal{M}_{\mathcal{F}^s} : \boldsymbol{\Phi}^s \to \mathcal{F}^s$ defined in (SM1.13) is continuous. Since the set $\boldsymbol{\Phi}^s$ is compact, we conclude that the function space $\mathcal{F}^s$ is a compact metric space.

Next, in order to show the compactness of $\mathcal{G}^s$, we proceed in a similar fashion. Let us define a mapping $\mathcal{M}_{\mathcal{G}^s} : \boldsymbol{\Phi}^s \to \mathcal{G}^s$ with $\mathcal{M}_{\mathcal{G}^s}(\boldsymbol{\Theta}^s) = g^s_{\boldsymbol{\Theta}^s}$, where the notation $g^s_{\boldsymbol{\Theta}^s}(x^s) = \boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s}(x^s)$ refers to the network output function defined in (3.7) by clarifying its dependence on the network parameters. Similarly to (SM1.14), it is easy to observe that $\boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s}(x^s)$ is a continuous function of $\boldsymbol{\Theta}^s$ and for any sequence $\{\boldsymbol{\Theta}^s_n\}$ converging to an element $\boldsymbol{\Theta}^s_* \in \boldsymbol{\Phi}^s$

$$\lim_{n\to\infty} g^s_{\boldsymbol{\Theta}^s_n}(x^s) = \lim_{n\to\infty} \boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s_n}(x^s) = \boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s_*}(x^s) = g^s_{\boldsymbol{\Theta}^s_*}(x^s)$$

uniformly. Hence,

$$\lim_{n\to\infty} \mathfrak{d}^s(g^s_{\boldsymbol{\Theta}^s_n}, g^s_{\boldsymbol{\Theta}^s_*}) = \lim_{n\to\infty} \sup_{x^s\in\mathcal{X}^s} \|g^s_{\boldsymbol{\Theta}^s_n}(x^s) - g^s_{\boldsymbol{\Theta}^s_*}(x^s)\| = \sup_{x^s\in\mathcal{X}^s} \lim_{n\to\infty} \|g^s_{\boldsymbol{\Theta}^s_n}(x^s) - g^s_{\boldsymbol{\Theta}^s_*}(x^s)\| = 0.$$

Hence, the mapping $\mathcal{M}_{\mathcal{G}^s} : \boldsymbol{\Phi}^s \to \mathcal{G}^s$ is continuous. Then, from the compactness of $\boldsymbol{\Phi}^s$, it follows that the function space $\mathcal{G}^s$ is compact as well. ∎

Lemma SM1.7. *Let Assumptions 3.1, 3.3, 3.4 hold. Then, the covering numbers of the function classes $\mathcal{F}^s$ and $\mathcal{F}^t$ are upper bounded as*

$$\mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}^s_{\mathcal{X}}) \le \prod_{l=1}^{L-1} \left(\frac{4A_\Theta L_K Q}{\epsilon^2} + 1\right)^{d_l(d_{l-1}+1)}$$
$$\mathcal{N}(\mathcal{F}^t, \epsilon, \mathfrak{d}^t_{\mathcal{X}}) \le \prod_{l=1}^{L-1} \left(\frac{4A_\Theta L_K Q}{\epsilon^2} + 1\right)^{d_l(d_{l-1}+1)}$$

where the dimension-dependent constant Q is defined as

$$Q \triangleq \sum_{l=1}^{L-1} Q_l$$

with

$$Q_l \triangleq (L_\eta R_{l-1}\sqrt{d_l d_{l-1}} + L_\eta\sqrt{d_l}) + \sum_{i=1}^{l-1}(L_\eta R_{i-1}\sqrt{d_i d_{i-1}} + L_\eta\sqrt{d_i}) \prod_{k=i+1}^{l} L_\eta A_\Theta \sqrt{d_k d_{k-1}} \tag{SM1.15}$$

for $l = 2, \ldots, L$ and $Q_1 \triangleq L_\eta\sqrt{d_1 d_0}\, R_0 + L_\eta\sqrt{d_1}$. Here

$$R_l \triangleq (A_\eta A_\Theta)^l (A_x\sqrt{d_0} + 1)\sqrt{d_1}\prod_{k=1}^{l-1}\sqrt{d_{k+1}d_k} + \sum_{i=2}^{l-1}(A_\eta A_\Theta)^{l+1-i}\sqrt{d_i}\prod_{k=i}^{l-1}\sqrt{d_{k+1}d_k} + A_\eta A_\Theta\sqrt{d_l}$$

under condition (3.15) and $R_l \triangleq C_\eta\sqrt{d_l}$ under condition (3.14) for $l = 2, \ldots, L-1$, where $R_0 \triangleq A_x$ and $R_1 \triangleq A_\eta A_\Theta\sqrt{d_1 d_0}A_x + A_\eta A_\Theta\sqrt{d_1}$.

Proof. We obtain the bound only for the source domain, as the derivation for the target domain is identical. Our proof is based on constructing an ϵ-cover for the compact metric space $\mathcal{F}^s$. For two mappings $f^s_1, f^s_2 \in \mathcal{F}^s$ defined respectively by the parameter vectors $\boldsymbol{\Theta}^s_1, \boldsymbol{\Theta}^s_2$

we have

$$
\begin{aligned}
(\mathfrak{d}^s_{\mathcal{X}}(f^s_1, f^s_2))^2 &= \sup_{x^s \in \mathcal{X}^s} \|f^s_1(x^s) - f^s_2(x^s)\|^2_{\mathcal{X}} \\
&= \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} \|\phi^l(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s)) - \phi^l(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s))\|^2_{\mathcal{X}^l} \\
&= \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s)\right) - 2k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s)\right) \\
&\qquad + k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s)\right) \qquad \text{(SM1.16)} \\
&\le \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} \left|k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s)\right) - k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s)\right)\right| \\
&\qquad + \left|k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s)\right) - k^l\left(\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s), \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s)\right)\right| \\
&\le \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} 2L_K \|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_1}(x^s) - \boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s_2}(x^s)\|
\end{aligned}
$$

where the last inequality is due to the Lipschitz continuity of the kernels k^l. We next construct a cover for the set of parameter vectors $\boldsymbol{\Theta}^s$, which will define a cover for $\mathcal{F}^s$ using the relation in (SM1.16). From (3.3) the network parameter vectors of layer l are in the compact set

$$
\boldsymbol{\Theta}^l = \{\boldsymbol{\Theta}^l = [\mathbf{W}^l \ \mathbf{b}^l] \in \mathbb{R}^{d_l \times (d_{l-1}+1)} : |\mathbf{W}^l_{ij}| \le A_\Theta, |\mathbf{b}^l_i| \le A_\Theta, \forall i, j, l\}. \qquad \text{(SM1.17)}
$$

Then there exists a cover of $\boldsymbol{\Theta}^l$ consisting of open balls around a set $\mathfrak{G}^l = \{\boldsymbol{\Theta}^l_m\}_{m=1}^{\kappa^l}$ of regularly sampled grid points, with a distance of δ between adjacent grid centers in each dimension. The maximal overall distance between two adjacent grid centers is then $\delta\sqrt{d_l(d_{l-1}+1)}$. Hence, the distance between any parameter vector $\boldsymbol{\Theta}^l \in \boldsymbol{\Theta}^l$ and the nearest grid center $\boldsymbol{\Theta}^l_m$ is at most

$$
\frac{\delta\sqrt{d_l(d_{l-1}+1)}}{2}
$$

with the number of balls in the cover being

$$
\kappa^l = \left(\frac{2A_\Theta}{\delta} + 1\right)^{d_l(d_{l-1}+1)}.
$$

From the Cartesian product of the grid centers at layers $l = 1, \ldots, L-1$, we then obtain a product grid

$$
\mathfrak{G} = \mathfrak{G}^1 \times \cdots \times \mathfrak{G}^{L-1} = \{\boldsymbol{\Theta}_k\}_{k=1}^{\kappa^1 \ldots \kappa^{L-1}} \qquad \text{(SM1.18)}
$$

which defines a cover for the overall parameter space

$$
\boldsymbol{\Phi} = \{\boldsymbol{\Theta} = (\boldsymbol{\Theta}^1, \ldots, \boldsymbol{\Theta}^{L-1}) : \ |\boldsymbol{\Theta}^l_{ij}| \le A_\Theta, \forall i, j, l\}
$$

consisting of

$$\kappa_{\mathfrak{G}} = \prod_{l=1}^{L-1} \kappa^l = \prod_{l=1}^{L-1} \left(\frac{2A_\Theta}{\delta} + 1 \right)^{d_l(d_{l-1}+1)}$$

balls. Then for any $f^s \in \mathcal{F}^s$ with parameters $\boldsymbol{\Theta}^s$, there exists some $f^s_k \in \mathcal{F}^s$ with parameters $\boldsymbol{\Theta}_k = (\boldsymbol{\Theta}^1_k, \boldsymbol{\Theta}^2_k, \dots, \boldsymbol{\Theta}^{L-1}_k) \in \mathfrak{G}$ in the product grid such that

$$\|\boldsymbol{\Theta}^{sl} - \boldsymbol{\Theta}^l_k\| < \delta\sqrt{d_l(d_{l-1}+1)}. \tag{SM1.19}$$

For any $x^s \in \mathcal{X}^s$, the distance between the l-th layer features of these parameters can be bounded as

(SM1.20)

$$\begin{aligned}
\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| &= \left\| \eta^l \left(\mathbf{W}^{sl} \boldsymbol{\xi}^{s(l-1)}_{\boldsymbol{\Theta}^s}(x^s) + \mathbf{b}^{sl} \right) - \eta^l \left(\mathbf{W}^l_k \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{b}^l_k \right) \right\| \\
&\le L_\eta \left\| \mathbf{W}^{sl} \boldsymbol{\xi}^{s(l-1)}_{\boldsymbol{\Theta}^s}(x^s) + \mathbf{b}^{sl} - \mathbf{W}^l_k \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) - \mathbf{b}^l_k \right\| \\
&= L_\eta \left\| \mathbf{W}^{sl} \boldsymbol{\xi}^{s(l-1)}_{\boldsymbol{\Theta}^s}(x^s) - \mathbf{W}^{sl} \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{W}^{sl} \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) - \mathbf{W}^l_k \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{b}^{sl} - \mathbf{b}^l_k \right\| \\
&\le L_\eta \|\mathbf{W}^{sl}\| \, \|\boldsymbol{\xi}^{s(l-1)}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\| + L_\eta \|\mathbf{W}^{sl} - \mathbf{W}^l_k\| \, \|\boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\| + L_\eta \|\mathbf{b}^{sl} - \mathbf{b}^l_k\|
\end{aligned}$$

where $\mathbf{W}^l_k$, $\mathbf{b}^l_k$, and $\boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}$ denote the l-th layer network parameters and features generated by the parameter vector $\boldsymbol{\Theta}_k$; and $\|\cdot\|$ and $\|\cdot\|_F$ respectively denote the operator norm and the Frobenius norm of a matrix. From (SM1.17) and (SM1.19), we have

$$\begin{aligned}
\|\mathbf{W}^{sl}\| \le \|\mathbf{W}^{sl}\|_F &\le A_\Theta \sqrt{d_l d_{l-1}} \\
\|\mathbf{W}^{sl} - \mathbf{W}^l_k\| \le \|\mathbf{W}^{sl} - \mathbf{W}^l_k\|_F &< \delta \sqrt{d_l d_{l-1}} \\
\|\mathbf{b}^{sl} - \mathbf{b}^l_k\| &< \delta \sqrt{d_l}.
\end{aligned}$$

These bounds together with the inequality in (SM1.20) yield

$$\begin{aligned}
\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| < L_\eta A_\Theta \sqrt{d_l d_{l-1}} \, &\|\boldsymbol{\xi}^{s(l-1)}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\| \\
&+ L_\eta \delta \sqrt{d_l d_{l-1}} \, \|\boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\| + L_\eta \delta \sqrt{d_l}.
\end{aligned} \tag{SM1.21}$$

In order to study (SM1.21), we first obtain an upper bound on the term $\|\boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\|$. Notice that for the condition (3.14), we simply have

$$\begin{aligned}
\|\boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| = \|\eta^l \left(\mathbf{W}^l \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{b}^l \right)\| &= \left(\sum_{i=1}^{d_l} \left(\eta^l_i (\mathbf{W}^l \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{b}^l) \right)^2 \right)^{1/2} \\
&\le C_\eta \sqrt{d_l}.
\end{aligned} \tag{SM1.22}$$

Next, for the condition (3.15) we have

$$\begin{aligned}
\|\boldsymbol{\xi}^0_{\boldsymbol{\Theta}_k}(x^s)\| &= \|x^s\| \le A_x \\
\|\boldsymbol{\xi}^1_{\boldsymbol{\Theta}_k}(x^s)\| &= \|\eta^1 \left(\mathbf{W}^1 \boldsymbol{\xi}^0_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{b}^1 \right)\| \le A_\eta \, \|\mathbf{W}^1 \boldsymbol{\xi}^0_{\boldsymbol{\Theta}_k}(x^s) + \mathbf{b}^1\| \\
&\le A_\eta \left(\|\mathbf{W}^1\| \|\boldsymbol{\xi}^0_{\boldsymbol{\Theta}_k}(x^s)\| + \|\mathbf{b}^1\| \right) \le A_\eta A_\Theta \sqrt{d_1 d_0} A_x + A_\eta A_\Theta \sqrt{d_1}
\end{aligned}$$

for layers $l = 0$ and $l = 1$. For $l \geq 2$, one can similarly establish a recursive relation between the parameter vectors of layers l and $l-1$, which yields

$$\begin{aligned}\|\boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| &\leq A_\eta \left(\|\mathbf{W}^l\|\|\boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\| + \|\mathbf{b}^l\|\right)\\ &\leq A_\eta A_\Theta \sqrt{d_l d_{l-1}}\|\boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\| + A_\eta A_\Theta \sqrt{d_l}\\ &\leq (A_\eta A_\Theta)^l (A_x\sqrt{d_0}+1)\sqrt{d_1}\prod_{k=1}^{l-1}\sqrt{d_{k+1}d_k}\\ &+ \sum_{i=2}^{l-1}(A_\eta A_\Theta)^{l+1-i}\sqrt{d_i}\prod_{k=1}^{l-1}\sqrt{d_{k+1}d_k} + A_\eta A_\Theta\sqrt{d_l}.\end{aligned}$$

Hence, combining this with (SM1.22), we get

$$\|\boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| \leq R_l \tag{SM1.23}$$

for $l = 2, \ldots, L-1$, where R_l is the constant defined in Lemma SM1.7. Using this in (SM1.21), we obtain

$$\begin{aligned}\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| &< L_\eta A_\Theta \sqrt{d_l d_{l-1}}\,\|\boldsymbol{\xi}^{s(l-1)}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{l-1}_{\boldsymbol{\Theta}_k}(x^s)\|\\ &+ L_\eta\delta\sqrt{d_l d_{l-1}}\,R_{l-1} + L_\eta\delta\sqrt{d_l}.\end{aligned} \tag{SM1.24}$$

For layer $l = 1$, we have

$$\begin{aligned}\|\boldsymbol{\xi}^{s1}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^1_{\boldsymbol{\Theta}_k}(x^s)\| &< L_\eta A_\Theta\sqrt{d_1 d_0}\,\|\boldsymbol{\xi}^{s0}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^0_{\boldsymbol{\Theta}_k}(x^s)\|\\ &+ L_\eta\delta\sqrt{d_1 d_0}\,R_0 + L_\eta\delta\sqrt{d_1}\\ &= L_\eta\delta\sqrt{d_1 d_0}\,R_0 + L_\eta\delta\sqrt{d_1}\end{aligned}$$

since $\boldsymbol{\xi}^{s0}_{\boldsymbol{\Theta}^s}(x^s) = \boldsymbol{\xi}^0_{\boldsymbol{\Theta}_k}(x^s) = x^s$. This relation together with the recursive inequality in (SM1.24) yields

$$\begin{aligned}\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| &< \delta\Bigg((L_\eta R_{l-1}\sqrt{d_l d_{l-1}} + L_\eta\sqrt{d_l})\\ &+ \sum_{i=1}^{l-1}(L_\eta R_{i-1}\sqrt{d_i d_{i-1}} + L_\eta\sqrt{d_i})\prod_{k=i+1}^{l} L_\eta A_\Theta\sqrt{d_k d_{k-1}}\Bigg)\\ &= Q_l\delta\end{aligned} \tag{SM1.25}$$

for $l = 1, \ldots, L-1$. Hence, we have shown that for any $f^s \in \mathcal{F}^s$ with parameters $\boldsymbol{\Theta}^s$, there exists some $f^s_k \in \mathcal{F}^s$ with parameters $\boldsymbol{\Theta}_k \in \mathfrak{G}$ in the product grid such that

$$\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| < Q_l\delta$$

for any $x^s \in \mathcal{X}^s$. We can now use this in (SM1.16) to bound the distance $\mathfrak{d}^s_{\mathcal{X}}(f^s, f^s_k)$ as

$$(\mathfrak{d}^s_{\mathcal{X}}(f^s, f^s_k))^2 \leq \sup_{x^s\in\mathcal{X}^s}\sum_{l=1}^{L-1} 2L_K\|\boldsymbol{\xi}^{sl}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^l_{\boldsymbol{\Theta}_k}(x^s)\| < 2L_K\delta\sum_{l=1}^{L-1}Q_l = 2L_K\delta Q. \tag{SM1.26}$$

Therefore, the set $\{f_k^s\}_{k=1}^{\kappa_{\mathfrak{G}}} \subset \mathcal{F}^s$ provides a cover for $\mathcal{F}^s$ with covering radius $\sqrt{2L_K\delta Q}$. In order to obtain a covering radius of $\epsilon = \sqrt{2L_K\delta Q}$, we set

$$\delta = \frac{\epsilon^2}{2L_K Q}$$

which provides a grid consisting of

$$\prod_{l=1}^{L-1} \kappa^l = \prod_{l=1}^{L-1} \left(\frac{4A_\Theta L_K Q}{\epsilon^2} + 1 \right)^{d_l(d_{l-1}+1)}$$

balls that covers $\mathcal{F}^s$. Hence, we obtain the upper bound

$$\mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}_{\mathcal{X}}^s) \leq \prod_{l=1}^{L-1} \left(\frac{4A_\Theta L_K Q}{\epsilon^2} + 1 \right)^{d_l(d_{l-1}+1)}$$

for the covering number stated in the lemma. ∎

Lemma SM1.8. *Let Assumptions 3.1, 3.3, 3.4 hold. Then, the covering numbers of the function classes $\mathcal{H} \circ \mathcal{F}^s$ and $\mathcal{H} \circ \mathcal{F}^t$ are upper bounded as*

$$\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s) \leq \prod_{l=1}^{L} \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)}$$

$$\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \epsilon, \mathfrak{d}^t) \leq \prod_{l=1}^{L} \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)}.$$

Proof. We prove the statement of the lemma only for the source function space $\mathcal{H} \circ \mathcal{F}^s$, as the derivations for the target domain are identical. In order to bound the covering number for $\mathcal{H} \circ \mathcal{F}^s$, we proceed as in the proof of Lemma SM1.7 and extend the grid construction in (SM1.18) to include layer L as well. This defines a grid

$$\mathfrak{G}_{\mathcal{H}\circ\mathcal{F}} = \mathfrak{G}^1 \times \cdots \times \mathfrak{G}^L = \{\mathbf{\Theta}_k\}_{k=1}^{\kappa^1 \ldots \kappa^L} \tag{SM1.27}$$

providing a cover for the parameter space

$$\mathbf{\Phi}_{\mathcal{H}\circ\mathcal{F}} = \{\mathbf{\Theta} = (\mathbf{\Theta}^1, \ldots, \mathbf{\Theta}^L) : \ |\mathbf{\Theta}_{ij}^l| \leq A_\Theta, \forall i, j, l\}$$

consisting of

$$\prod_{l=1}^{L} \kappa^l = \prod_{l=1}^{L} \left(\frac{2A_\Theta}{\delta} + 1 \right)^{d_l(d_{l-1}+1)} \tag{SM1.28}$$

balls. Then for any $g^s \in \mathcal{H} \circ \mathcal{F}^s$ with network parameters $\mathbf{\Theta}^s$, there exists some $g_k^s \in \mathcal{H} \circ \mathcal{F}^s$ with network parameters $\mathbf{\Theta}_k = (\mathbf{\Theta}_k^1, \mathbf{\Theta}_k^2, \ldots, \mathbf{\Theta}_k^L) \in \mathfrak{G}_{\mathcal{H}\circ\mathcal{F}}$ in the grid such that

$$\|\mathbf{\Theta}^{sl} - \mathbf{\Theta}_k^l\| < \delta\sqrt{d_l(d_{l-1}+1)}$$

for $l = 1, \ldots, L$. Proceeding in a similar fashion to the derivations in (SM1.20) and (SM1.21), we obtain

$$\begin{aligned}\|\boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{L}_{\boldsymbol{\Theta}_k}(x^s)\| &\leq L_\eta \|\mathbf{W}^{sL}\| \, \|\boldsymbol{\xi}^{s(L-1)}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{L-1}_{\boldsymbol{\Theta}_k}(x^s)\| \\ &\quad + L_\eta \|\mathbf{W}^{sL} - \mathbf{W}^{L}_k\| \, \|\boldsymbol{\xi}^{L-1}_{\boldsymbol{\Theta}_k}(x^s)\| + L_\eta \|\mathbf{b}^{sL} - \mathbf{b}^{L}_k\| \\ &< L_\eta A_\Theta \sqrt{d_L d_{L-1}} \, \|\boldsymbol{\xi}^{s(L-1)}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{L-1}_{\boldsymbol{\Theta}_k}(x^s)\| \\ &\quad + L_\eta \delta \sqrt{d_L d_{L-1}} \, \|\boldsymbol{\xi}^{L-1}_{\boldsymbol{\Theta}_k}(x^s)\| + L_\eta \delta \sqrt{d_L}\end{aligned} \tag{SM1.29}$$

for any $x^s \in \mathcal{X}^s$. Combining this inequality with the bounds in (SM1.23) and (SM1.25) gives

$$\begin{aligned}\|\boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{L}_{\boldsymbol{\Theta}_k}(x^s)\| &< L_\eta A_\Theta \sqrt{d_L d_{L-1}} \, Q_{L-1}\delta \\ &\quad + L_\eta \delta \sqrt{d_L d_{L-1}} \, R_{L-1} + L_\eta \delta \sqrt{d_L} \\ &= Q_L \delta.\end{aligned}$$

Recalling the definition of the distance $\mathfrak{d}^s$ in (2.4), we then have

$$\mathfrak{d}^s(g^s, g^s_k) = \sup_{x^s \in \mathcal{X}^s} \|g^s(x^s) - g^s_k(x^s)\| = \sup_{x^s \in \mathcal{X}^s} \|\boldsymbol{\xi}^{sL}_{\boldsymbol{\Theta}^s}(x^s) - \boldsymbol{\xi}^{L}_{\boldsymbol{\Theta}_k}(x^s)\| < Q_L \delta.$$

Hence, the grid $\mathfrak{G}_{\mathcal{H}\circ\mathcal{F}}$ in (SM1.27) provides a cover for $\mathcal{H} \circ \mathcal{F}^s$ with covering radius $Q_L\delta$. For a covering radius of ϵ, we set $\epsilon = Q_L\delta$, which results in a cover with

$$\prod_{l=1}^{L} \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)} \tag{SM1.30}$$

balls due to (SM1.28). We thus get the covering number upper bound

$$\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s) \leq \prod_{l=1}^{L} \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)}$$

stated in the lemma. ∎

Lemma SM1.9. *Let Assumption 3.7 hold. Assume also that the composite function classes $\mathcal{V}^s$ and $\mathcal{V}^t$ are compact with respect to the metrics*

$$\begin{aligned}\mathfrak{d}^s_{\mathcal{V}}(v^s_1, v^s_2) &\triangleq \sup_{x^s \in \mathcal{X}^s} |v^s_1(x^s) - v^s_2(x^s)| \\ \mathfrak{d}^t_{\mathcal{V}}(v^t_1, v^t_2) &\triangleq \sup_{x^t \in \mathcal{X}^t} |v^t_1(x^t) - v^t_2(x^t)|\end{aligned}$$

where $v^s_1, v^s_2 \in \mathcal{V}^s$ and $v^t_1, v^t_2 \in \mathcal{V}^t$. Then,

$$\begin{aligned}&P\left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, \Delta \in \mathcal{D}} |D_\Delta(f^s, f^t) - \hat{D}_\Delta(f^s, f^t)| \leq \epsilon \right) \\ &\geq 1 - 2\mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}^s_{\mathcal{V}}) \exp\left(-\frac{N_s \epsilon^2}{72 C^2_{\mathcal{D}}} \right) - 2\mathcal{N}(\mathcal{V}^t, \frac{\epsilon}{6}, \mathfrak{d}^t_{\mathcal{V}}) \exp\left(-\frac{N_t \epsilon^2}{72 C^2_{\mathcal{D}}} \right).\end{aligned}$$

Proof. Due to the assumption of compactness of the function classes $\mathcal{V}^s$ and $\mathcal{V}^t$, there exists an ϵ-cover of each function space. Let us denote the cover numbers of $\mathcal{V}^s$ and $\mathcal{V}^t$ as

$$\kappa^s = \mathcal{N}(\mathcal{V}^s, \epsilon, \mathfrak{d}^s_{\mathcal{V}}), \qquad \kappa^t = \mathcal{N}(\mathcal{V}^t, \epsilon, \mathfrak{d}^t_{\mathcal{V}})$$

respectively, and the corresponding sets of ball centers as $\{v^s_k\}_{k=1}^{\kappa^s}$ and $\{v^t_l\}_{l=1}^{\kappa^t}$. Then, for any $v^s \in \mathcal{V}^s$ and any $v^t \in \mathcal{V}^t$ there exist some $v^s_k \in \mathcal{V}^s$ and $v^t_l \in \mathcal{V}^t$ such that

$$\begin{aligned} \mathfrak{d}^s_{\mathcal{V}}(v^s, v^s_k) &= \sup_{x^s \in \mathcal{X}^s} |v^s(x^s) - v^s_k(x^s)| < \epsilon \\ \mathfrak{d}^t_{\mathcal{V}}(v^t, v^t_l) &= \sup_{x^t \in \mathcal{X}^t} |v^t(x^t) - v^t_l(x^t)| < \epsilon. \end{aligned} \tag{SM1.31}$$

Let us denote

$$\begin{aligned} D(v^s_k, v^t_l) &\triangleq \left|E[v^s_k(x^s)] - E[v^t_l(x^t)]\right| \\ \hat{D}(v^s_k, v^t_l) &\triangleq \left|\frac{1}{N_s}\sum_{i=1}^{N_s} v^s_k(x^s_i) - \frac{1}{N_t}\sum_{j=1}^{N_t} v^t_l(x^t_j)\right|. \end{aligned}$$

Take any $f^s \in \mathcal{F}^s$, $f^t \in \mathcal{F}^t$ and $\Delta \in \mathcal{D}$. We have

$$\begin{aligned} &|D_\Delta(f^s, f^t) - \hat{D}_\Delta(f^s, f^t)| \\ &= |D_\Delta(f^s, f^t) - D(v^s_k, v^t_l) + D(v^s_k, v^t_l) - \hat{D}(v^s_k, v^t_l) + \hat{D}(v^s_k, v^t_l) - \hat{D}_\Delta(f^s, f^t)| \\ &\leq |D_\Delta(f^s, f^t) - D(v^s_k, v^t_l)| + |D(v^s_k, v^t_l) - \hat{D}(v^s_k, v^t_l)| + |\hat{D}(v^s_k, v^t_l) - \hat{D}_\Delta(f^s, f^t)|. \end{aligned} \tag{SM1.32}$$

We proceed by bounding each one of the three terms at the right hand side of the inequality in (SM1.32). The first term can be upper bounded as

$$\begin{aligned} |D_\Delta(f^s, f^t) - D(v^s_k, v^t_l)| &= \left||E[v^s(x^s)] - E[v^t(x^t)]| - |E[v^s_k(x^s)] - E[v^t_l(x^t)]|\right| \\ &\leq \left|E[v^s(x^s)] - E[v^t(x^t)] - E[v^s_k(x^s)] + E[v^t_l(x^t)]\right| \\ &\leq |E[v^s(x^s)] - E[v^s_k(x^s)]| + |E[v^t(x^t)] - E[v^t_l(x^t)]| < 2\epsilon \end{aligned} \tag{SM1.33}$$

where the last inequality follows from (SM1.31). For the third term in (SM1.32), one can similarly show that

$$|\hat{D}(v^s_k, v^t_l) - \hat{D}_\Delta(f^s, f^t)| < 2\epsilon. \tag{SM1.34}$$

We lastly study the second term in (SM1.32). We have

$$\begin{aligned} &|D(v^s_k, v^t_l) - \hat{D}(v^s_k, v^t_l)| \\ &= \left|\left|E[v^s_k(x^s)] - E[v^t_l(x^t)]\right| - \left|\frac{1}{N_s}\sum_{i=1}^{N_s} v^s_k(x^s_i) - \frac{1}{N_t}\sum_{j=1}^{N_t} v^t_l(x^t_j)\right|\right| \\ &\leq \left|E[v^s_k(x^s)] - E[v^t_l(x^t)] - \frac{1}{N_s}\sum_{i=1}^{N_s} v^s_k(x^s_i) + \frac{1}{N_t}\sum_{j=1}^{N_t} v^t_l(x^t_j)\right| \\ &\leq \left|\frac{1}{N_s}\sum_{i=1}^{N_s} v^s_k(x^s_i) - E[v^s_k(x^s)]\right| + \left|\frac{1}{N_t}\sum_{j=1}^{N_t} v^t_l(x^t_j) - E[v^t_l(x^t)]\right|. \end{aligned} \tag{SM1.35}$$

As the domain discriminator is bounded due to Assumption 3.7, from Hoeffding's inequality we have

$$P\left(\left|\frac{1}{N_s}\sum_{i=1}^{N_s} v_k^s(x_i^s) - E[v_k^s(x^s)]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{N_s\epsilon^2}{2C_{\mathcal{D}}^2}\right)$$

for a fixed $v_k^s \in \mathcal{V}^s$, and a similar inequality can be obtained for a fixed $v_l^t \in \mathcal{V}^t$. Applying the union bound over all ball centers $\{v_k^s\}_{k=1}^{\kappa^s}$ and $\{v_l^t\}_{l=1}^{\kappa^t}$, we get that with probability at least

$$1 - 2\kappa^s \exp\left(-\frac{N_s\epsilon^2}{2C_{\mathcal{D}}^2}\right) - 2\kappa^t \exp\left(-\frac{N_t\epsilon^2}{2C_{\mathcal{D}}^2}\right)$$

we have

$$\left|\frac{1}{N_s}\sum_{i=1}^{N_s} v_k^s(x_i^s) - E[v_k^s(x^s)]\right| < \epsilon \quad \text{and} \quad \left|\frac{1}{N_t}\sum_{j=1}^{N_t} v_l^t(x_j^t) - E[v_l^t(x^t)]\right| < \epsilon$$

for all ball centers, which implies from (SM1.35)

$$|D(v_k^s, v_l^t) - \hat{D}(v_k^s, v_l^t)| < 2\epsilon.$$

Combining this result with the bounds in (SM1.32)-(SM1.34), we get

$$\begin{aligned} &P\left(\sup_{f^s\in\mathcal{F}^s, f^t\in\mathcal{F}^t, \Delta\in\mathcal{D}} |D_\Delta(f^s, f^t) - \hat{D}_\Delta(f^s, f^t)| \leq 6\epsilon\right) \\ &\geq 1 - 2\kappa^s \exp\left(-\frac{N_s\epsilon^2}{2C_{\mathcal{D}}^2}\right) - 2\kappa^t \exp\left(-\frac{N_t\epsilon^2}{2C_{\mathcal{D}}^2}\right). \end{aligned}$$

Replacing ϵ with $\epsilon/6$, we get the statement of the lemma. ∎

SM2. Discussion of the results in relation with previous literature. We now discuss our findings in relation with previous literature. To the best of our knowledge, our study is the first to propose an in-depth characterization of the sample complexity of domain-adaptive neural networks. A substantial body of work has focused on the effect of domain discrepancy on generalization performance, while another line of research has examined the sample complexity of neural networks, however, in a single-domain setting. We briefly overview these results below, along with a few relevant studies on the performance of domain alignment methods. For clarity and consistency, we restate the findings of prior work using our own notation. The presence of the parameter δ in the bounds signifies that the result holds with probability at least $1-\delta$.

SM2.1. Effect of domain discrepancy on generalization performance. One of the earliest analyses examining the effect of the deviation between the source and target distributions is the study by Ben-David et al. [SM6]. The gap between the expected target loss and the empirical source loss is shown to be bounded by

$$O\left(\sqrt{\frac{\dim_{VC}(\mathcal{H})}{M_s}+\log(\delta^{-1})}\right)+d_{\mathcal{H}}(D_S,D_T)+\lambda$$

ignoring the logarithmic factors, where $\dim_{VC}(\mathcal{H})$ denotes the VC-dimension of the hypothesis space $\mathcal{H}$, M_s is the number of of labeled source samples, and λ is a measure of the proximity of the true label function to the hypothesis class $\mathcal{H}$. Here $d_{\mathcal{H}}(D_S,D_T)$ is the $\mathcal{A}$-distance [SM6] between the source and target distributions D_S and D_T, given by

$$d_{\mathcal{H}}(D_S,D_T)=2\sup_{A\in\mathcal{A}}|P_{D_S}(A)-P_{D_T}(A)|$$

where $\mathcal{A}$ is the set of domain subsets with characteristic functions in $\mathcal{H}$, and $P_{(\cdot)}$ denotes probability with respect to a distribution.

In a succeeding study [SM5], this result has been extended to algorithms minimizing a convex combination of source and target losses, where the hypothesis that minimizes the empirical weighted loss is shown to generalize to the target domain within an error of

$$O\Bigg(\sqrt{\frac{\alpha^2}{\gamma}+\frac{(1-\alpha)^2}{1-\gamma}}\sqrt{\frac{\dim_{VC}(\mathcal{H})+\log(\delta^{-1})}{M}} + (1-\alpha)\left(\sqrt{\frac{\dim_{VC}(\mathcal{H})\log(\delta^{-1})}{N}}+\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T)+\lambda\right)\Bigg).$$

Here the distribution distance $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_S,D_T)$ denotes the empirical divergence between the source and the target distributions over the symmetric difference hypothesis space $\mathcal{H}\Delta\mathcal{H}$, which corresponds to the set of disagreements [SM5]. $N=N_s=N_t$ denotes the number of all samples in the two domains, and M is the total number of labeled samples, with $M_s=(1-\gamma)M$ source samples and $M_t=\gamma M$ target samples. This result has some implications paralel to our study, in that the optimal weight α of the target loss should decrease with the scarcity of

target labels, i.e., as γ decreases. A high domain discrepancy $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ also drives the weighted loss towards the target loss, by decreasing the weight $1-\alpha$ of the source loss.

Similar findings have been presented in the study of Mansour et al. in terms of the Rademacher complexities of the hypothesis space [SM26]. However, in [SM26] the deviation between the source and the target domains has been characterized in terms of the discrepancy $\mathrm{disc}_\ell(D_S, D_T)$, which quantifies how the loss-induced disagreement between any pair of hypotheses may differ across D_S and D_T.

Following these pioneering works, many other domain divergence measures have been proposed in succeeding studies [SM32]. Deng et al. have explored a robust variant of the discrepancy in [SM26] based on the adversarial Rademacher complexity definition [SM11], which has been shown to vary with the number of samples M and the network width d at rate $O(\sqrt{d/M})$ for two-layer ReLU neural networks. Zhang et al. have proposed an alternative characterization of distribution distance based on the margin disparity discrepancy, leading to generalization bounds in terms of the Rademacher complexities and the covering numbers of hypothesis spaces [SM45]. Zellinger et al. have presented performance bounds depending on the VC-dimension of the function classes by formulating the domain discrepancy in terms of the difference between the moments of the source and target distributions [SM44]. Other recent efforts along this line include studies involving margin-aware risks with links to optimal transport distances [SM12], information-theoretic bounds based on mutual information [SM38, SM42], hypothesis-specific divergence measures [SM39], and risk definitions based on stochastic predictors [SM33].

Remark SM2.1. We note that all these aforementioned works assume that a common classifier is learnt in the original source and target domains; i.e., their setting is essentially different from ours as they do not at all consider learning a transformation or a mapping that aligns the two domains. The main distinction among these works lies in the specific distribution discrepancy each one proposes to characterize the misalignment between the domains, with the purpose of deriving tighter error bounds. Meanwhile, the reported labeled and unlabeled sample complexities, or otherwise the errors, follow the classical dependence on the VC-dimensions or the Rademacher complexities of the hypothesis classes in consideration, consistent with well-established results in learning theory. From the perspective of domain alignment algorithms, one may want to regard the domain discrepancies in these bounds as the distance obtained after mapping the two domains to a shared domain, an interpretation that arguably extends to transformation learning. While this view holds to some extent, many of the discrepancy measures used in these works (including their empirical approximations) are defined in a theoretical manner, and are difficult to estimate in practice. Although efficient computational techniques may exist for some of these discrepancy measures, they often lack accompanying learning guarantees. In contrast, our main results in Theorems 2.7-3.11 offer a practical means of assessing the generalization capability of domain alignment algorithms, as they are based on the empirical distribution distance computed directly on the aligned training data.

SM2.2. Performance bounds for domain alignment algorithms. To the best of our knowledge, a very limited number of theoretical analyses have investigated the performance of learning domain-aligning transformations or representations. A multi-task domain adaptation

method is proposed in [SM46], which learns the similarity between source and target samples through a linear transformation $\mathbf{G}$. Assuming the incoherence of the projections corresponding to different tasks, the estimation error of the transformation $\mathbf{G}$ is shown to be bounded by $O(d_T\sqrt{\log(d_S)/n})$, where d_S and d_T denote the dimensions of the source and target Euclidean domains, and n is the number of tasks. While this bound is subsequently leveraged in [SM46] to design suitable classifiers based on the incoherence principle, the scope of their analysis is limited to linear transformations.

A performance analysis of conditional distribution matching is presented in [SM37], showing that the generalization gap in the target domain is bounded by

$$O\left(1+\frac{1}{\sqrt{M_t}}+\sqrt{\frac{\log(\delta^{-1})}{M_s+M_t}}\right)$$

when the source domain is mapped to the target domain through a location and scale transform.

Fang et al. have considered semi-supervised domain alignment algorithms as in our work [SM13]. However, their analysis is significantly different from ours since it does not explore the sample complexity of learning domain transformations, but instead treats the sample complexity as a known problem parameter. Their study aims to demonstrate that the need for labeled target data can be alleviated under certain assumptions by relying on the source and unlabeled target data.

Transferring representations from a source task to a target task is a problem different from but connected to domain adaptation. Wang et al. have provided an extensive analysis of transfer learning and multitask learning through domain-invariant feature representations by minimizing a combined empirical loss under regularization [SM40]. The performance gap between the source and target losses is shown to vary at rate

$$O\left(\mathrm{dist}_{\mathcal{Y}}(f^s,f^t)+\sqrt{\frac{\log(\delta^{-1})}{M_s+M_t}}\right).$$

Here $\mathrm{dist}_{\mathcal{Y}}(f^s,f^t)$ denotes the $\mathcal{Y}$-discrepancy [SM29] between the two domains once transformed to a shared domain, which is, however, not easy to estimate in practice.

Galanti et al. have modeled the transfer learning problem in a setting where a target task and multiple source tasks are drawn from the same distribution of distributions, and considered that a neural network architecture is partially transferred to the target task [SM15]. Their analysis implies that for accurate transfer, the number of source tasks and the number of samples per source task must scale with the number of edges, respectively, in the transferred component and the target-specific component of the network. In a recent work, Jiao et al. have considered a model that distinguishes between shared and domain-specific features in multi-domain deep transfer learning and shown that transferability between tasks improves the convergence rates in the target task [SM22]. McNamara and Balcan have investigated representation learning on a source task and fine-tuning on a target task [SM28]. The accuracy on the source task is shown to carry over to the target task within a performance gap of

$O(\sqrt{\dim_{VC}(\mathcal{H}\circ\mathcal{F})/M_s}+\sqrt{\dim_{VC}(\mathcal{H})/M_t})$, where $\mathcal{F}$ is the space of feature representations and $\mathcal{H}$ is the space of classifiers. The significance of this result lies in the fact that the number M_t of labeled target samples should scale with the dimension of only the classifier $\mathcal{H}$, rather than the more complex composite hypothesis space $\mathcal{H}\circ\mathcal{F}$. A paralel finding is presented in [SM35] for the problem of transfer learning in a multi-task setting, demonstrating that the number of labeled samples for a new task needs to scale only with the complexity of its own task-specific map, assuming the abundance of the training data for the previous tasks.

Remark SM2.2. Although our domain adaptation setting differs essentially from that considered in these transfer learning studies, they are comparable in their shared focus on handling the scarcity of labeled target samples. Whereas these works tie sample complexity to the richness of the target function class, which can be still large for deep neural networks, our analysis indicates that in a domain adaptation scenario the limitedness of target labels can be tolerated through strategically choosing the weight parameter as $\alpha = O(\sqrt{M_t})$, independently of the complexity of the target function class.

SM2.3. Sample complexity of neural networks in a single domain. Sample complexity of neural networks is a well-explored topic in statistical learning theory, a comprehensive overview of which can be found in [SM1], [SM4]. Although this classical line of research pertains to learning algorithms in a single domain and does not extend to domain adaptation scenarios, we find it instructive to briefly review these results and compare them to our bounds on domain adaptive neural networks.

The sample complexity of a feed-forward network consisting of W weights, L layers and s output units, with fixed piecewise-polynomial activation functions is reported as [SM1, Theorem 21.5]

$$O\left(\frac{s(WL\log(W)+WL^2)\log(\epsilon^{-1})+\log(\delta^{-1})}{\epsilon^2}\right) \tag{SM2.1}$$

in order to attain an error of ϵ. Denoting the network width as d, the number of weights W in an L-layer network is obtained as $W = d^2L$. Then, the sample complexity $M = O(d^2L^3)$ in (SM2.1) points to a quadratic dependence on d and a cubic dependence on L. This polynomial dependence is in line with our results in Theorems 3.6 and 3.11, where the sample complexity of labeled source data has been obtained as $M_s = O(d^2L^2)$. The dependence on L is quadratic, hence slightly tighter in our bounds.

A more recent trend in the exploration of sample complexity of neural networks is the characterization of the complexity in a dimension-independent way under particular assumptions. Neyshabur et al. have shown that the sample complexity depends exponentially on the network depth; nevertheless, its dependence on the network width can be removed under group norm regularization of network weights [SM31]. In succeeding studies, the exponential dependence on the network size has been reduced to polynomial [SM41], quadratic [SM30], linear [SM19] and logarithmic [SM3] factors. Harvey et al. have shown that the VC-dimension of neural networks with ReLU activation functions is $O(WL\log(W))$, resulting in comparable bounds to our work [SM21]. In some more recent works, it has been shown that the dependence on network width can be removed for one-layer networks [SM36] and reduced to logarithmic factors for two-layer networks [SM10] under bounded Frobenius norm and spectral

norm constraints. We note that these results essentially rely on the condition that the norms of the weight matrices be upper bounded in a dimension-independent manner, and would translate to rather pessimistic sample complexities under the removal of this assumption.

Remark SM2.3. While the above studies have contributed to a comprehensive understanding of neural network classifiers, they all focus on the single-domain scenario, assuming identical distributions for training and test data. To the best of our knowledge, our work is the first to provide a detailed analysis of the sample complexity of domain-adaptive neural networks. We note that our analysis does not impose any special constraints on the weight matrices, such as norm regularization. Under the incorporation of norm constraints, we would expect to arrive at tighter bounds consistently with the approaches in single-domain settings, which is left as a potential future direction of our study.

SM3. Detailed experimental results. This section provides comprehensive details of the experimental validation presented in Section 5 of the main article, including complete setup descriptions, implementation details, and extended discussions of the results.

SM3.1. General domain alignment methods: Detailed setup and results.

SM3.1.1. Synthetic data experiments. We validate our theoretical findings on a synthetic data set with two classes. The source and target data sets are generated by applying two different geometric transformations to 400 samples drawn from the standard normal distribution in $\mathbb{R}^2$. We simulate a learning algorithm that learns geometric transformations to map the source and target samples to a common domain and then trains a classifier in the shared domain. Here we emulate a setting where the transformations f^s and f^t are treated as if learnt from data, however, with some error. In practice, f^s and f^t are formed by perturbing the ground truth geometric transformations with some transformation estimation error τ. We test a range of estimation error levels τ in the experiments. The classifier trained after mapping the samples to the common domain is chosen as a regularized ridge regression algorithm solving

$$\min_{\mathbf{w}\in\mathbb{R}^2} \quad \frac{1-\alpha}{M_s}\sum_{i=1}^{M_s}(\mathbf{w}^T f^s(x_i^s)-\mathbf{y}_i^s)^2+\frac{\alpha}{M_t}\sum_{j=1}^{M_t}(\mathbf{w}^T f^t(x_j^t)-\mathbf{y}_j^t)^2+\lambda\|\mathbf{w}\|^2.$$

The target misclassification rate is evaluated over 1000 test samples drawn from the target distribution and classified through the learnt hypothesis $\mathbf{w}$ and target transformation f^t.

The variation of the target misclassification rate with the number M_t of labeled target samples is shown for different values of the weight α for the target loss. In order to interpret these results, it is helpful to recall our theoretical analysis: Theorem 2.1 states that the expected target loss $\mathcal{L}^t(f^t,h)$ deviates from its reference value based on the empirical weighted loss $\hat{\mathcal{L}}_\alpha(f^s,f^t,h)$ and the distance $D(f^s,f^t)$ by an amount of ϵ. In order to achieve this with high and fixed probability, the term $M_t\epsilon^2$ in the probability expression must be constant (ignoring logarithmic factors and assuming that the generic covering numbers grow at a typical geometric rate). This implies that the expected target loss should decrease at rate $\epsilon = O(\sqrt{1/M_t})$ as M_t increases. We observe that the decay in the target error with M_t is consistent with Theorem 2.1. The fitted theoretical rates of decay $O(\sqrt{1/M_t})$ closely match the experimental data. We can also observe that large M_t values favor larger α values, while α must be chosen smaller at small M_t values. This aligns with the conclusion that the parameter α must be chosen as $\alpha = O(\sqrt{M_t})$ in order to control the probability term as M_t decreases.

We also study the variation of the target misclassification rate with the estimation error τ of the geometric transformations. The parameter τ here is taken as the norm of the error matrix that is added to the ground truth transformation matrix. Hence, τ can be regarded as a parameter proportional to the distribution distance $D(f^s,f^t)$. The misclassification rate tends to increase with τ at an approximately linear rate, as confirmed by the theoretical linear rate of increase fitted to the experimental data. These results are coherent with the prediction of Theorem 2.1 that the expected target loss should increase proportionally to the distribution distance $D(f^s,f^t)$.

SM3.1.2. MIT-CBCL face recognition experiments. We experiment on the MIT-CBCL image data set [SM27]. The data set consists of a total of 3240 synthetic face images belonging to 10 subjects. The images of each subject are rendered under 36 different illumination conditions and 9 poses, with Pose 1 corresponding to the frontal view and Pose 9 corresponding to a nearly profile view. We consider the images rendered under Pose 1 as the source domain, and repeat experiments by taking images from Poses 2, 5 and 9 as the target domain in each trial. First, using all labeled and unlabeled images, we compute a mapping between the source and target domains by the method proposed in [SM14], which finds a transformation that aligns the PCA bases of the source and target domains. We then train an SVM classifier using all labeled samples from the two domains. The unlabeled target samples are finally classified with the learnt transformation and classifier.

The misclassification rates of unlabeled target samples are evaluated with respect to the number of labeled target and source samples. We observe that the misclassification rates are reduced effectively with the increase in the number of labeled samples. As predicted by our theory, the target loss asymptotically reduces to an error component resulting from the empirical loss and the distribution distance, at rates $O(\sqrt{1/M_t})$ and $O(\sqrt{1/M_s})$ with increasing M_t and M_s. The experimental results seem consistent with this expectation. The theoretical curves fitted to the experimental data with the expected rates of decrease are indicated in the plots for visual comparison.

SM3.2. Domain-adaptive neural networks: Detailed setup and results. We experimentally verify our results in Theorems 3.6 and 3.7 regarding the sample complexity of domain-adaptive neural networks. For both MMD-based and adversarial architectures, we characterize the sample complexity with respect to the depth L and the width d of the network, and investigate the optimal value of the weight α of the target loss.

In our experiments, the MNIST handwritten digit data set [SM24] is used as the source data set, which consists of 60000 images. The target data set is taken as MNIST-M [SM16], which contains 59000 handwritten digit images with colored backgrounds. We train the neural networks with labeled and unlabeled training samples from the source and target domains, and then evaluate the target accuracy of the learnt models, defined as the correct classification rate of test samples from the target domain. In all experiments, algorithm hyperparameters and fixed variables are chosen to keep the neural network in the overfitting regime, enabling the characterization of the sample complexity of the models under consideration. Operating in the overfitting regime means that the network's target accuracy drops below a predefined threshold as network complexity increases for a fixed number of training samples. Increasing the amount of labeled or unlabeled data delays the onset of this accuracy drop, which allows us to determine the required sample size for a given level of network complexity without using excessive computational resources.

SM3.2.1. Experimental methodology. In the experiments, the primary goal is to characterize how target accuracy behaves as network complexity increases, measured either by the number of layers L or the width parameter d. In the overfitting regime, for fixed sample sizes M_s or N_s, target accuracy decreases as network complexity grows. Our experiments indicate that this decrease can be approximated as a linear decline. Therefore, in addition to the data points obtained experimentally, we estimate values for untested complexity levels by applying

linear extrapolation, improving computational efficiency. The target accuracy vs. network complexity curves along with the corresponding linear regression fits are presented in the left panels of the relevant figures.

The right panels analyze how the number of labeled or unlabeled samples required to guarantee a fixed target accuracy changes with network size. Specific target accuracy levels are selected, and the network complexities at which these accuracy levels are maintained are identified by drawing a horizontal line at the chosen target accuracy and finding its intersections with the linearly extrapolated curves. Once the maximum network complexity at which the network maintains the predefined target accuracy is identified, the variation of these complexity values with respect to M_s or N_s is plotted. A quadratic curve in the form of ax^2+bx+c with the constraint $\frac{-b}{2a} \geq 0$ is fitted to the experimental results for visual evaluation.

For the optimal α experiments, the relationship between target accuracy and α is first characterized, and the optimal value α_{opt} is identified for each M_t value (while keeping M_s constant). Our experimental results suggest that the variation of target accuracy with α can be approximated as quadratic, which is meaningful since α is expected to have an optimal value that maximizes target accuracy. The α value corresponding to the peak of the fitted parabola is denoted as α_{opt}. Subsequently, to validate the theoretical relationship $\alpha_{opt} = O(\sqrt{M_t})$, the data are plotted with α_{opt} on the vertical axis and M_t on the horizontal axis, and a curve of the form $a + b\sqrt{x}$ is fitted. This procedure is repeated for different M_s values to verify consistency.

SM3.2.2. MMD-based domain adaptation networks. In our analysis of MMD-based domain adaptation networks, we consider the architecture proposed in the pioneering study [SM25] as our benchmark. We build on our previous experimental study [SM23] and employ a neural network structure similar to the baseline model in [SM25], beginning with convolutional layers and followed by several fully connected MMD layers. The MMD layer parameters are coupled between the source and target domains. The dimensions (widths) of all MMD layers are set as equal. Batch normalization is applied after each layer in order to stabilize the performance. We use the PyTorch implementation of the network available in [SM8] and adapt it for the minimization of the objective function

$$\frac{1-\alpha}{M_s}\sum_{i=1}^{M_s}\ell(h\circ f(x_i^s),\mathbf{y}_i^s)+\frac{\alpha}{M_t}\sum_{i=1}^{M_t}\ell(h\circ f(x_j^t),\mathbf{y}_j^t)+\beta\sum_{l=1}^{L-1}(\hat{D}^l)^2(f^l,f^l)$$

where $\ell(\cdot,\cdot)$ is set as the cross-entropy loss function and the source and target feature transformations are coupled as $f^s = f^t = f$ and $f^{sl} = f^{tl} = f^l$.

Implementation details.. One significant modification with respect to the original implementation is the handling of the convolutional layers in the feature extractor. In the original article [SM25], these layers were fixed during training, i.e., their parameters were not updated. In our implementation, the convolutional layers are trained iteratively during each training process. The parameter L represents the number of fully connected MMD layers in the network, and all of these layers can be easily configured since they are fully connected (linear) layers. The parameter d represents the factor by which the width of the MMD layers in the original implementation [SM8] is multiplied. When new fully connected layers are added

to the network (i.e., when L increases), batch normalization layers are inserted between the added layers to stabilize training and improve performance.

Tuning of β.. The parameter β determines the weight of the MMD term in the objective function. Since the MMD loss is computed separately for each layer and then summed, the overall MMD loss increases as L increases. To prevent this term from dominating and suppressing the classification information from the labeled data, the parameter β is chosen to be inversely proportional to the number of layers L.

Training details.. During training, we use the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and momentum of 0.9. The batch size is set to 512. The number of training epochs is selected to increase proportionally with the size of the network in order to ensure convergence. The complete experimental configurations, including fixed, independent, and group (control) variable values for each experiment, are summarized in Table SM1.

Experiment	**Fixed Variables**	**Independent Variables**	**Group (Control) Variables**
L vs. M_s	M_t: 115, N_s: 60000, N_t: 59000, d: 80, α: 0.2, β: $1.75/L$	L: {4, 5, 6, 7, 8}	M_s: {117, 234, 351, 819}
L vs. N_s	M_t: 750, M_s: 2750, d: 80, α: 0.2, β: $1.75/(1 + L/4)$	L: {2, 3, 4, 5, 6}	N_s, N_t: {30000, 36000, 42000, 54000}
d vs. M_s	M_t: 115, N_s: 60000, N_t: 59000, L: 1, α: 0.25, β: 1	d: {40, 80, 120, 160, 320, 640, 1280}	M_s: {117, 234, 351, 819}
Optimum α	M_s: {234, 819, 1755}, N_s: 6000, N_t: 5900, L: 4, d: 80, β: 0.875	α: {0.2, 0.4, 0.6, 0.8}	M_t: {115, 345, 460, 575}

Table SM1

Experimental configurations for MMD-based domain adaptation experiments.

Sample complexity with respect to depth and width.. We study the sample complexity of labeled source samples M_s and all source samples N_s with respect to the number L of MMD layers in the network. The experiments show the decrease in the target accuracy as the number L of MMD layers increases when the network is in the overfitting regime, for different M_s and N_s values. We aim to characterize the sample complexity of M_s and N_s with respect to L in this experiment. Therefore, we determine several desired target accuracy levels and identify the smallest M_s and N_s values that ensure this target accuracy as L grows (in cases where obtaining the exact value of L exceeded our computational resources, we resorted to linear extrapolation to approximately infer the corresponding L value). We recall from Theorem 3.6 that the sample complexities of M_s and N_s are expected to grow at quadratic rates $M_s = O(L^2)$ and $N_s = O(L^2)$ as the network depth L increases. The experimental findings confirm this prediction, as the increase in the required sample size for attaining a reference target accuracy level indeed follows a quadratic increase with L. The curves are obtained by fitting quadratic polynomials to the experimental data for visual evaluation.

A similar experiment is conducted for sample complexity with respect to the network width. The parameter d represents the factor by which the network width in the original implementation [SM8] is multiplied in our experiment. Hence, d is directly proportional to the shared width parameter of the MMD layers. The results are also consistent with the theoretical findings in Theorem 3.6, which states that the sample complexity must increase at a quadratic rate $M_s = O(d^2)$ as the network width increases.

Optimal weight parameter α.. We recall from Theorem 3.6 that, in order to maximize the target accuracy, the weight parameter α of the target classification loss must scale as $\alpha = O(\sqrt{M_t})$ as the number M_t of labeled target samples varies. We experimentally validate this result. We examine the variation of the target accuracy with the weight parameter α, which follows a non-monotonic variation with α as expected. We approximately identify the optimal value α_{opt} of the weight parameter for each value of M_t by applying polynomial fitting to the plots. In order to visually observe the prediction of Theorem 3.6, we also fit a curve of $O(\sqrt{M_t})$ to each data sequence. The experimental data seems consistent with the fitted curves, which supports the statement of Theorem 3.6 that the optimal weight parameter must scale at rate $\alpha_{opt} = O(\sqrt{M_t})$.

SM3.2.3. Adversarial domain adaptation networks. In order to experimentally evaluate our findings in Section 3.3.2, we adopt the model proposed in [SM17], which is a well-known representative of adversarial domain adaptation architectures. We use the PyTorch implementation of this model available in [SM18], by adapting it to the semi-supervised setting studied in our analysis. We train the adversarial network to minimize the objective function

$$\frac{1-\alpha}{M_s}\sum_{i=1}^{M_s}\ell(h\circ f(x_i^s),\mathbf{y}_i^s)+\frac{\alpha}{M_t}\sum_{i=1}^{M_t}\ell(h\circ f(x_j^t),\mathbf{y}_j^t)$$
$$-\frac{\beta}{N_s+N_t}\left(\sum_{i=1}^{N_s}\ell_{\mathcal{D}}(\Delta\circ f(x_i^s),l_i^s)+\sum_{j=1}^{N_t}\ell_{\mathcal{D}}(\Delta\circ f(x_j^t),l_j^t)\right)$$

where the label loss $\ell(\cdot,\cdot)$ and the domain discrimator loss $\ell_{\mathcal{D}}(\cdot,\cdot)$ are selected as the negative log likelihood function, and the source and target feature extractor networks are coupled as $f^s = f^t = f$.

Implementation details.. Two major modifications were applied to the original implementation [SM18]: (i) transitioning from a fully unsupervised structure to the semi-supervised framework studied in our analysis, and (ii) adding the capability to systematically increase the network capacity. The first modification is reflected in the summation limits of the objective function, which now distinguish between labeled samples (used for classification loss) and all samples (used for domain discrimination loss). The second modification is implemented through the parameters L and d as described below.

The feature extractor network contains only convolutional layers, while the label predictor and domain discriminator networks consist of fully connected layers in the implementation in [SM18]. The original architecture in [SM17] consists of a feature extractor with two convolutional layers, a label predictor with three fully connected layers, and a domain discriminator with two fully connected layers. In our implementation, the model is updated to take L as an

input parameter and dynamically stack the corresponding number of convolutional and fully connected layers. Specifically, when analyzing the sample complexity of labeled data (M_s), we set the number of layers in the feature extractor and label predictor networks as equal, which is represented by the parameter L. Likewise, when studying the sample complexity of all data (N_s), the number of layers in the feature extractor and domain discriminator networks are equated and denoted as L. The convolutional layers in the feature extractor, being shared by both networks, are modified in all experiments. Batch normalization and ReLU layers are included after each convolutional or fully connected layer, following standard practice.

We use a similar strategy to adjust the network width, where we scale the number of convolutional channels and the fully connected layer width in the original paper [SM17] with the same factor d. In the original architecture, the convolutional layers have 64 channels and the fully connected layers have 100 neurons. Hence, the number of convolutional channels is scaled proportionally to the width of the label predictor and the domain discriminator networks, respectively, when studying the sample complexities of M_s and N_s. For instance, in a M_s vs. d experiment with $d = 2$ and $L = 3$, the resulting architecture consists of a feature extractor with 4 convolutional layers (each with 128 channels), a label predictor with 4 fully connected layers (each with 200 neurons), and a domain discriminator with 2 fully connected layers (each with 100 neurons).

Training details.. Training is conducted using the Adam optimizer with a learning rate of 0.001. The batch size is set to 128, and the network is trained for 100 epochs. The complete experimental configurations for each experiment are summarized in Table SM2.

Experiment	**Fixed Variables**	**Independent Variables**	**Group (Control) Variables**
L vs. M_s	M_t: 20, N_s, N_t: 6000, d: 1	L: {4, 5, 6, 7, 8, 9, 10}	M_s: {60, 120, 180, 240}
L vs. N_s	M_t: 20, M_s: 240, d: 1	L: {3, 5, 7, 9}	N_s, N_t: {750, 1500, 3000, 12000}
d vs. M_s	M_t: 30, N_s, N_t: 6000, L: 2	d: {1, 4, 8, 16, 24, 32}	M_s: {120, 180, 240, 300, 360}
d vs. N_s	M_t: 30, M_s: 300, L: 2	d: {1, 4, 8, 16, 24, 32}	N_s, N_t: {750, 1500, 3000, 6000}
Optimum α	M_s: {240, 360, 480}, N_s, N_t: 6000, L: 3, d: 1	α: varied in $[0, 1]$	M_t: {60, 120, 210, 240}

Table SM2

Experimental configurations for adversarial domain adaptation experiments.

Sample complexity with respect to depth and width.. The sample complexities of the number of source samples with the network depth L and width d are presented in the figures. Similarly to the MMD experiments, we show the variation of the target accuracy with L or d at different M_s and N_s values, and then investigate the smallest M_s and N_s values ensuring a reference target accuracy level as L or d increases. The results of these experiments align with the theoretical bounds in Theorem 3.7, confirming the quadratic growth in the sample complexities

$M_s, N_s = O(L^2)$ and $M_s, N_s = O(d^2)$ as the network depth L and width d increase.

Optimal weight parameter α.. We lastly study the choice of the parameter α weighting the target classification loss in the objective function for the adversarial setting. The results confirm the theoretical prediction that the optimal value of the weight parameter should scale at rate $\alpha_{opt} = O(\sqrt{M_t})$ as the number of labeled samples varies.

Overall, our experimental findings are in line with the theoretical bounds presented in Theorems 3.6 and 3.7, supporting our sample complexity and optimal weight choice analyses for both MMD-based and adversarial domain adaptation networks.

SM4. Derivation of Lipschitz constants for common nonlinear activation functions. Here we derive Lipschitz constants for some widely used nonlinear activation functions. Let $\eta : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ represent an activation function in layer l giving the output $\zeta = \eta(\boldsymbol{\xi})$ for the input $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$.

SM4.1. ReLU activation. We begin with the rectified linear unit (ReLU) function $\eta_R : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ given by

$$\boldsymbol{\zeta}(k) = \max\{0, \boldsymbol{\xi}(k)\} \tag{SM4.1}$$

where $\boldsymbol{\zeta} = \eta_R(\boldsymbol{\xi})$, and the notation $(\cdot)(k)$ denotes the k-th entry of a vector. For two vectors $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^{d_l}$, we have

$$\begin{aligned} \|\eta_R(\boldsymbol{\xi}_1) - \eta_R(\boldsymbol{\xi}_2)\|^2 &= \sum_{k=1}^{d_l} (\max\{0, \boldsymbol{\xi}_1(k)\} - \max\{0, \boldsymbol{\xi}_2(k)\})^2 \\ &\leq \sum_{k=1}^{d_l} (\boldsymbol{\xi}_1(k) - \boldsymbol{\xi}_2(k))^2 = \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|^2 \end{aligned} \tag{SM4.2}$$

where $\max\{\cdot, \cdot\}$ denotes the maximum of two scalar values. We thus get

$$\|\eta_R(\boldsymbol{\xi}_1) - \eta_R(\boldsymbol{\xi}_2)\| \leq \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|$$

which gives the Lipschitz constant of the ReLU function as $L_R = 1$.

SM4.2. Softplus activation. Next, we consider the softplus function $\eta_{SP} : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ given by

$$\boldsymbol{\zeta}(k) = \log\left(1 + e^{\boldsymbol{\xi}(k)}\right) \tag{SM4.3}$$

where $\boldsymbol{\zeta} = \eta_{SP}(\boldsymbol{\xi})$. The derivative of the components of the softplus function can be upper bounded as

$$\left|\frac{d}{dt}\log(1 + e^t)\right| = \left|\frac{e^t}{1 + e^t}\right| < 1 \tag{SM4.4}$$

for all $t \in \mathbb{R}$. Then for $\boldsymbol{\zeta}_1 = \eta_{SP}(\boldsymbol{\xi}_1)$ and $\boldsymbol{\zeta}_2 = \eta_{SP}(\boldsymbol{\xi}_2)$ with $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^{d_l}$, from the mean value theorem we get

$$|\boldsymbol{\zeta}_1(k) - \boldsymbol{\zeta}_2(k)| \leq |\boldsymbol{\xi}_1(k) - \boldsymbol{\xi}_2(k)| \tag{SM4.5}$$

which implies

$$\|\eta_{SP}(\boldsymbol{\xi}_1) - \eta_{SP}(\boldsymbol{\xi}_2)\| \leq \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|. \tag{SM4.6}$$

Hence, we obtain the Lipschitz constant of the softplus function as $L_{SP} = 1$.

SM4.3. Softmax activation. Lastly, we consider the softmax function $\eta_{SM} : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ given by

$$\eta_{SM}(\boldsymbol{\xi}) = [\eta^1_{SM}(\boldsymbol{\xi})\ \eta^2_{SM}(\boldsymbol{\xi})\ \cdots\ \eta^{d_l}_{SM}(\boldsymbol{\xi})]^T$$

where $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$ and each k-th component $\eta^k_{SM}(\boldsymbol{\xi}) : \mathbb{R}^{d_l} \to \mathbb{R}$ of the softmax activation is defined as

$$\eta^k_{SM}(\boldsymbol{\xi}) = \frac{e^{\boldsymbol{\xi}(k)}}{\sum_{n=1}^{d_l} e^{\boldsymbol{\xi}(n)}}. \tag{SM4.7}$$

Since the functions $\eta^k_{SM}(\boldsymbol{\xi})$ are differentiable for all k, for any two $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^{d_l}$, it follows from the multivariable mean value theorem that there exists some $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$ lying in the line segment between $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ such that

$$\eta^k_{SM}(\boldsymbol{\xi}_1) - \eta^k_{SM}(\boldsymbol{\xi}_2) = (\nabla \eta^k_{SM}(\boldsymbol{\xi}))^T(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)$$

where $\nabla \eta^k_{SM}(\boldsymbol{\xi}) \in \mathbb{R}^{d_l}$ denotes the gradient of η^k_{SM} at $\boldsymbol{\xi}$. The following inequality is then obtained

$$|\eta^k_{SM}(\boldsymbol{\xi}_1) - \eta^k_{SM}(\boldsymbol{\xi}_2)| \leq \sup_{\boldsymbol{\xi} \in \mathbb{R}^{d_l}} \|\nabla \eta^k_{SM}(\boldsymbol{\xi})\|\, \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|. \tag{SM4.8}$$

In the sequel, in order to find a Lipschitz constant for the softmax function, we derive a bound on the norm $\|\nabla \eta^k_{SM}(\boldsymbol{\xi})\|$ of its gradient.

For the case $k \neq n$, the derivative of $\eta^k_{SM}(\boldsymbol{\xi})$ with respect to the n-th entry $\boldsymbol{\xi}(n)$ of $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$ is obtained as

$$\frac{\partial \eta^k_{SM}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(n)} = \frac{\partial}{\partial \boldsymbol{\xi}(n)} \left(\frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right) = -\frac{e^{\boldsymbol{\xi}(k)} e^{\boldsymbol{\xi}(n)}}{\left(\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)} \right)^2}.$$

Since all $e^{\boldsymbol{\xi}(1)}, \ldots, e^{\boldsymbol{\xi}(d_l)}$ are positive, it is easy to show that $(e^{\boldsymbol{\xi}(1)} + \ldots, + e^{\boldsymbol{\xi}(d_l)})^2 \geq 4 e^{\boldsymbol{\xi}(k)} e^{\boldsymbol{\xi}(n)}$. Using this in the above expression, we get the bound

$$\left| \frac{\partial \eta^k_{SM}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(n)} \right| \leq \frac{1}{4}. \tag{SM4.9}$$

Next, for the case $k = n$, we have

$$\frac{\partial \eta^k_{SM}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(k)} = \frac{\partial}{\partial \boldsymbol{\xi}(k)} \left(\frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right) = \left(\frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right) \left(1 - \frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right).$$

Letting $\alpha = e^{\boldsymbol{\xi}(k)} / \sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}$ in the above expression and observing that the maximum value of the function $\alpha(1-\alpha)$ in the interval $\alpha \in [0, 1]$ is $1/4$, we get

$$\left| \frac{\partial \eta^k_{SM}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(k)} \right| \leq \frac{1}{4}. \tag{SM4.10}$$

Combining the results (SM4.9) and (SM4.10), the gradient of $\eta_{SM}^k(\boldsymbol{\xi})$ can be bounded as

$$\|\nabla \eta_{SM}^k(\boldsymbol{\xi})\| \leq \frac{\sqrt{d_l}}{4}$$

for any $\boldsymbol{\xi} \in \mathbb{R}^{d_l}$. Using this in (SM4.8) gives

$$|\eta_{SM}^k(\boldsymbol{\xi}_1) - \eta_{SM}^k(\boldsymbol{\xi}_2)| \leq \frac{\sqrt{d_l}}{4}\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|$$

for any $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in \mathbb{R}^{d_l}$, which implies

$$\|\eta_{SM}(\boldsymbol{\xi}_1) - \eta_{SM}(\boldsymbol{\xi}_2)\| \leq \frac{d_l}{4}\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|.$$

Defining

$$d_{\max} = \max_{l=1,\ldots,L} d_l$$

we thus get the Lipschitz constant of the softmax function as $L_{SM} = d_{\max}/4$.

SM5. Proof of Corollary 3.5.

Proof. In order to analyze the dependence of $\mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}^s_{\mathcal{X}})$ on d and L, we first study how the term R_l in Lemma SM1.7 grows with the dimension d and the number of layers L. For condition (3.14), we have

$$R_l = C_\eta \sqrt{d_l} = O(d^{1/2}).$$

For condition (3.15), representing the relevant constant terms as c for simplicity, we have

$$R_l = O\big((cd)^l\big).$$

We next study the term Q_l in (SM1.15). For condition (3.14), we obtain

$$Q_l = O(c^{l-1}\, d^{l+\frac{1}{2}})$$

which results in

$$Q = O(c^{L-2} d^{L-\frac{1}{2}}). \tag{SM5.1}$$

Meanwhile, condition (3.15) yields

$$Q_l = O\big((l-1)\, c^{l-1}\, d^l\big)$$

resulting in

$$Q = O\big((L-2)\, c^{L-2}\, d^{L-1}\big). \tag{SM5.2}$$

For simplicity, we may combine the results in (SM5.1) and (SM5.2) through a slightly more pessimistic but brief common upper bound as

$$Q = O\big(L\, c^{L-2} d^L\big)$$

which is valid for both of the conditions in (3.14) and (3.15). Then, from the expressions of the covering numbers $\mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}^s_{\mathcal{X}})$ and $\mathcal{N}(\mathcal{F}^t, \epsilon, \mathfrak{d}^t_{\mathcal{X}})$ in Lemma SM1.7, we conclude

$$\mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}^s_{\mathcal{X}}) = O\left(\left(\frac{cQ}{\epsilon^2}\right)^{d^2 L}\right) = O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2}\right)$$

where we have taken the liberty to replace the ϵ^2 term in the denominator with ϵ for simplicity, as they will lead to equivalent bounds. Similarly,

$$\mathcal{N}(\mathcal{F}^t, \epsilon, \mathfrak{d}^t_{\mathcal{X}}) = O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2}\right).$$

We next analyze the covering number $\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s)$ for the hypothesis space $\mathcal{H} \circ \mathcal{F}^s$. For condition (3.14), we have

$$Q_L = O(c^{L-1}\, d^{L+\frac{1}{2}})$$

which gives from Lemma SM1.8

$$\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s,\epsilon,\mathfrak{d}^s)=O\left(\left(\frac{cQ_L}{\epsilon}\right)^{d^2L}\right)=O\left(\frac{(cd)^{d^2L^2}}{\epsilon^{d^2L}}\right) \tag{SM5.3}$$

if the $d^2L/2$ term added to the d^2L^2 term in the exponent is ignored for simplicity. Next, for condition (3.15) we obtain

$$Q_L=O\big((L-1)\,c^{L-1}\,d^L\big)$$

resulting in

$$\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s,\epsilon,\mathfrak{d}^s)=O\left(\left(\frac{cQ_L}{\epsilon}\right)^{d^2L}\right)=O\left(\left(\frac{L}{\epsilon}\right)^{d^2L}(cd)^{d^2L^2}\right). \tag{SM5.4}$$

Combining the bounds in (SM5.3) and (SM5.4), we arrive at the common upper bound

$$\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s,\epsilon,\mathfrak{d}^s)=O\left(\left(\frac{L}{\epsilon}\right)^{d^2L}(cd)^{d^2L^2}\right)$$

which covers both conditions. Identical derivations for the target domain yield

$$\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t,\epsilon,\mathfrak{d}^t)=O\left(\left(\frac{L}{\epsilon}\right)^{d^2L}(cd)^{d^2L^2}\right). \qquad \blacksquare$$

SM6. Proof of Theorem 3.6.

Proof. We first notice that, owing to Lemma SM1.5, we can analyze MMD-based domain adaptation networks within the setting of Theorem 2.7. The compactness of the function spaces $\mathcal{F}^s$, $\mathcal{F}^t$, $\mathcal{H}\circ\mathcal{F}^s$, and $\mathcal{H}\circ\mathcal{F}^t$ follow from Assumptions 3.1-3.3 due to Lemma SM1.6. Assumptions 2.3 and 2.6 are thereby satisfied; hence, the statement of Theorem 2.7 applies to the current setting in consideration.

We recall from Theorem 2.7 that the expected target loss in (3.16) is attained with probability at least

$$\begin{aligned}1-2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t,\frac{\epsilon}{8\alpha L_\ell},\mathfrak{d}^t)e^{-\frac{M_t\epsilon^2}{8\alpha^2A_\ell^2}}-2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s,\frac{\epsilon}{8(1-\alpha)L_\ell},\mathfrak{d}^s)e^{-\frac{M_s\epsilon^2}{8(1-\alpha)^2A_\ell^2}}\\-\mathcal{N}(\mathcal{F}^s,\frac{\epsilon}{8},\mathfrak{d}^s_{\mathcal{X}})\exp(-a_s(N_s,\epsilon))-\mathcal{N}(\mathcal{F}^t,\frac{\epsilon}{8},\mathfrak{d}^t_{\mathcal{X}})\exp(-a_t(N_t,\epsilon)).\end{aligned} \tag{SM6.1}$$

Our proof is then based on identifying the rate at which the number of samples should grow with L and d so that each one of the terms subtracted from 1 in the expression (SM6.1) remains fixed. This will in return guarantee that the generalization gap of $O(\epsilon)$ in (3.16) be attained with high probability.

We begin with the term $\mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}^s_{\mathcal{X}}) \exp(-a_s(N_s, \epsilon))$. Recalling the definition of $a_s(N_s, \epsilon)$ from Lemma SM1.4, we have

$$a_s(N_s, \epsilon) = \theta(N_s \epsilon^2)$$

where we use the notation $\theta(\cdot)$ to refer to asymptotic tight bounds. Combining this with Corollary 3.5, we obtain

$$\begin{aligned}\mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}^s_{\mathcal{X}}) \exp(-a_s(N_s, \epsilon)) &= O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2} \exp(-N_s \epsilon^2)\right) \\ &= O\left(\exp\left(d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(cd) - N_s \epsilon^2\right)\right).\end{aligned}$$

We conclude that the total number N_s of source samples required to ensure a lower bound on the probability expression (SM6.1) scales as

$$N_s = O\left(\frac{d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}{\epsilon^2}\right),$$

yielding the sample complexity stated in the theorem. An identical derivation based on bounding the term $\mathcal{N}(\mathcal{F}^t, \frac{\epsilon}{8}, \mathfrak{d}^t_{\mathcal{X}}) \exp(-a_t(N_t, \epsilon))$ shows that N_t has the same sample complexity.

Next, we examine the terms involving the number of labeled samples. Proceeding similarly, we get

$$\begin{aligned}\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} &= O\left(\left(\frac{L\alpha}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2} \exp\left(-\frac{M_t \epsilon^2}{\alpha^2}\right)\right) \\ &= O\left(\exp\left(d^2 L \log\left(\frac{L\alpha}{\epsilon}\right) + d^2 L^2 \log(cd) - \frac{M_t \epsilon^2}{\alpha^2}\right)\right).\end{aligned}$$

Recalling that $0 \leq \alpha \leq 1$, we conclude that upper bounding the choice of the weight parameter α by the rate

$$\alpha = O\left(\left(\frac{M_t \epsilon^2}{d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}\right)^{1/2}\right)$$

ensures that the probability term $\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}}$ remain bounded.

Finally, for the number of labeled samples in the source domain, we have

$$\begin{aligned}&\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha) L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}} \\ &\quad = O\left(\left(\frac{L(1-\alpha)}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2} \exp\left(-\frac{M_s \epsilon^2}{(1-\alpha)^2}\right)\right) \\ &\quad = O\left(\exp\left(d^2 L \log\left(\frac{L(1-\alpha)}{\epsilon}\right) + d^2 L^2 \log(cd) - \frac{M_s \epsilon^2}{(1-\alpha)^2}\right)\right).\end{aligned}$$

Recalling again the bound $0 \leq 1-\alpha \leq 1$, we observe that the sample complexity

$$M_s = O\left(\frac{d^2 L \, \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}{\epsilon^2}\right)$$

ensures a lower bound on the probability expression (SM6.1), which concludes the proof of the theorem. ∎

SM7. Derivation of the bound and the Lipschitz constant for the cross-entropy loss.

We first discuss the magnitude bound A_ℓ for the widely used cross-entropy loss function. Let $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y} \subset \mathbb{R}^m$ be two nonnegative label vectors in the label set $\mathcal{Y} = [0,1] \times \cdots \times [0,1] \subset \mathbb{R}^m$. In its naïve form, the cross-entropy loss between $\mathbf{y}_1$ and $\mathbf{y}_2$ is given by

$$\ell(\mathbf{y}_1, \mathbf{y}_2) = -\sum_{k=1}^{m} \log(\mathbf{y}_1(k))\, \mathbf{y}_2(k) \tag{SM7.1}$$

where $\mathbf{y}(k)$ denotes the k-th entry of the vector $\mathbf{y}$. While the original form (SM7.1) of the cross-entropy loss is not bounded, often the following modification is made in order to avoid numerical issues in practical implementations

$$\ell(\mathbf{y}_1, \mathbf{y}_2) = -\sum_{k=1}^{m} \log(\mathbf{y}_1(k) + \delta)\, \mathbf{y}_2(k)$$

where $0 < \delta < 1$ is a positive constant. We then have

$$|\ell(\mathbf{y}_1, \mathbf{y}_2)| \leq \sum_{k=1}^{m} |-\log(\mathbf{y}_1(k)+\delta)\mathbf{y}_2(k)| \leq m \max\{|\log(\delta)|,\ \log(1+\delta)\}.$$

Assuming that δ is very small, we get the following bound on the loss magnitude

$$|\ell(\mathbf{y}_1, \mathbf{y}_2)| \leq A_\ell \triangleq m\, |\log(\delta)|.$$

We next derive the Lipschitz constant L_ℓ of the cross-entropy loss function. For any $\mathbf{y}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ we have

$$\begin{aligned} |\ell(\mathbf{y}_1, \mathbf{y}) - \ell(\mathbf{y}_2, \mathbf{y})| &= \left| -\sum_{k=1}^{m} \log(\mathbf{y}_1(k)+\delta)\mathbf{y}(k) + \sum_{k=1}^{m} \log(\mathbf{y}_2(k)+\delta)\mathbf{y}(k) \right| \\ &\leq \sum_{k=1}^{m} |\log(\mathbf{y}_2(k)+\delta) - \log(\mathbf{y}_1(k)+\delta)|\,. \end{aligned} \tag{SM7.2}$$

For any $t \geq \delta$, we have

$$\left|\frac{d}{dt}\log(t)\right| = \left|\frac{1}{t}\right| \leq \frac{1}{\delta}$$

which gives

$$\left|\frac{\log(\mathbf{y}_2(k)+\delta) - \log(\mathbf{y}_1(k)+\delta)}{\mathbf{y}_2(k) - \mathbf{y}_1(k)}\right| \leq \frac{1}{\delta}$$

due to the mean value theorem. Using this in (SM7.2), we get

$$|\ell(\mathbf{y}_1, \mathbf{y}) - \ell(\mathbf{y}_2, \mathbf{y})| \leq \sum_{k=1}^{m} \delta^{-1} |\mathbf{y}_2(k) - \mathbf{y}_1(k)| \leq \delta^{-1} \sqrt{m}\, \|\mathbf{y}_2 - \mathbf{y}_1\|$$

which shows that the cross-entropy loss is Lipschitz continuous with respect to the first argument with constant

$$L_\ell \triangleq \delta^{-1} \sqrt{m}.$$

SM8. Proof of Theorem 3.11.

Proof. We begin by bounding the expected target loss as

$$\mathcal{L}^t(f^t, h) \leq \mathcal{L}^s(f^s, h) + R_A\, D_\Delta(f^s, f^t)$$

using Assumption 3.10. It follows that

$$\begin{aligned} \mathcal{L}^t(f^t, h) &= \alpha \mathcal{L}^t(f^t, h) + (1-\alpha)\mathcal{L}^t(f^t, h) \\ &\leq \alpha \mathcal{L}^t(f^t, h) + (1-\alpha)\left(\mathcal{L}^s(f^s, h) + R_A\, D_\Delta(f^s, f^t)\right) \\ &= \mathcal{L}_\alpha(f^s, f^t, h) + (1-\alpha) R_A\, D_\Delta(f^s, f^t). \end{aligned} \tag{SM8.1}$$

We next aim to upper bound the expected loss $\mathcal{L}_\alpha(f^s, f^t, h)$ and the expected distribution distance $D_\Delta(f^s, f^t)$ in terms of their empirical counterparts. It follows from Assumptions 3.1 and 3.8 that the source hypothesis space $\mathcal{G}^s = \mathcal{H} \circ \mathcal{F}^s$, the target hypothesis space $\mathcal{G}^t = \mathcal{H} \circ \mathcal{F}^t$, the source domain discriminator space $\mathcal{V}^s = \mathcal{D} \circ \mathcal{F}^s$ and the target domain discriminator space $\mathcal{V}^t = \mathcal{D} \circ \mathcal{F}^t$ are compact with respect to the metrics $\mathfrak{d}^s, \mathfrak{d}^t, \mathfrak{d}^s_{\mathcal{V}}, \mathfrak{d}^t_{\mathcal{V}}$ respectively, which can be shown by following similar steps as in the proof of Lemma SM1.6 in Appendix 6.

Due to the compactness of $\mathcal{G}^s, \mathcal{G}^t$ and the assumptions on the classification loss function ℓ, we have

$$\begin{aligned} &P\left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, h \in \mathcal{H}} |\mathcal{L}_\alpha(f^s, f^t, h) - \hat{\mathcal{L}}_\alpha(f^s, f^t, h)| \leq \epsilon\right) \\ &\geq 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha) L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}} \end{aligned} \tag{SM8.2}$$

from Lemma SM1.2. Similarly, the compactness of $\mathcal{V}^s, \mathcal{V}^t$ together with Assumption 3.7 implies that

$$\begin{aligned} &P\left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, \Delta \in \mathcal{D}} |D_\Delta(f^s, f^t) - \hat{D}_\Delta(f^s, f^t)| \leq \epsilon\right) \\ &\geq 1 - 2\mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}^s_{\mathcal{V}}) \exp\left(-\frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2}\right) - 2\mathcal{N}(\mathcal{V}^t, \frac{\epsilon}{6}, \mathfrak{d}^t_{\mathcal{V}}) \exp\left(-\frac{N_t \epsilon^2}{72 C_{\mathcal{D}}^2}\right) \end{aligned} \tag{SM8.3}$$

due to Lemma SM1.9.

Combining the results in (SM8.1), (SM8.2), and (SM8.3), we get that with probability at least

$$1 - 2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s)e^{-\frac{M_s\epsilon^2}{8(1-\alpha)^2A_\ell^2}} - 2\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t)e^{-\frac{M_t\epsilon^2}{8\alpha^2A_\ell^2}} \\ - 2\mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}^s_{\mathcal{V}})\exp\left(-\frac{N_s\epsilon^2}{72C_{\mathcal{D}}^2}\right) - 2\mathcal{N}(\mathcal{V}^t, \frac{\epsilon}{6}, \mathfrak{d}^t_{\mathcal{V}})\exp\left(-\frac{N_t\epsilon^2}{72C_{\mathcal{D}}^2}\right) \tag{SM8.4}$$

the expected target loss is bounded as

$$\mathcal{L}^t(f^t, h) \leq \hat{\mathcal{L}}_\alpha(f^s, f^t, h) + (1-\alpha)R_A\,\hat{D}_\Delta(f^s, f^t) + (1-\alpha)R_A\epsilon + \epsilon.$$

In the sequel, we examine each one of the terms in the probability expression in (SM8.4). As for the covering numbers of $\mathcal{H}\circ\mathcal{F}^s$ and $\mathcal{H}\circ\mathcal{F}^t$, Assumptions 3.1, 3.4, and 3.8 ensure that the result in Lemma 3.4 applies to this setting as well, which implies that the rate of growth of $\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s, \epsilon, \mathfrak{d}^s)$ and $\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t, \epsilon, \mathfrak{d}^t)$ with L and d is upper bounded by

$$O\left(\left(\frac{L}{\epsilon}\right)^{d^2L}(cd)^{d^2L^2}\right)$$

due to Corollary 3.5. Then, following the very same steps as in the proof of Theorem 3.6, we get that upper bounding the weight parameter α by

$$\alpha = O\left(\left(\frac{M_t\epsilon^2}{d^2L\log\left(\frac{L}{\epsilon}\right) + d^2L^2\log(d)}\right)^{1/2}\right),$$

together with scaling M_s at rate

$$M_s = O\left(\frac{d^2L\,\log\left(\frac{L}{\epsilon}\right) + d^2L^2\log(d)}{\epsilon^2}\right)$$

ensures an upper bound on the terms

$$\mathcal{N}(\mathcal{H}\circ\mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s)e^{-\frac{M_s\epsilon^2}{8(1-\alpha)^2A_\ell^2}}$$

and

$$\mathcal{N}(\mathcal{H}\circ\mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t)e^{-\frac{M_t\epsilon^2}{8\alpha^2A_\ell^2}}$$

in the probability expression in (SM8.4).

Then, in order to analyze the covering numbers of $\mathcal{V}^s$ and $\mathcal{V}^t$, we proceed with the following reasoning: Noting the paralel between the structures of the domain discriminator and the feature extractor network parameters considered in Assumptions 3.8, 3.4 and 3.9, we observe

that the function space $\mathcal{V}^s = \mathcal{D} \circ \mathcal{F}^s$ has an identical construction to the function space $\mathcal{G}^s = \mathcal{H} \circ \mathcal{F}^s$, if the metric

$$\mathfrak{d}^s(g_1^s, g_2^s) = \sup_{x^s \in \mathcal{X}^s} \|g_1^s(x^s) - g_2^s(x^s)\|$$

based on the Euclidean distance in $\mathbb{R}^m$ is replaced by its counterpart

$$\mathfrak{d}_{\mathcal{V}}^s(v_1^s, v_2^s) = \sup_{x^s \in \mathcal{X}^s} |v_1^s(x^s) - v_2^s(x^s)|$$

which uses the Euclidean distance in $\mathbb{R}$ instead. Hence, the latter is a special case of the former that can be obtained by setting $m = 1$. Consequently, the analysis of the covering number $\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s)$ in Corollary 3.5 immediately applies to $\mathcal{N}(\mathcal{D} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}_{\mathcal{V}}^s)$ as well, only by replacing the number of layers L with the total number of layers $L + K - 1$ in the cascade network formed by the combination of the feature extractor and the domain discriminator networks. We thus get

$$\mathcal{N}(\mathcal{V}^s, \epsilon, \mathfrak{d}_{\mathcal{V}}^s) = O\left(\left(\frac{L+K}{\epsilon}\right)^{d^2(L+K)} (cd)^{d^2(L+K)^2}\right)$$

which yields

$$\begin{aligned} &\mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}_{\mathcal{V}}^s) \exp\left(-\frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2}\right) \\ &= O\left(\left(\frac{L+K}{\epsilon}\right)^{d^2(L+K)} (cd)^{d^2(L+K)^2} \exp\left(-\frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2}\right)\right) \\ &= O\left(\exp\left(d^2(L+K)\log\left(\frac{L+K}{\epsilon}\right) + d^2(L+K)^2 \log(cd) - \frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2}\right)\right). \end{aligned} \tag{SM8.5}$$

We thus conclude that the sample complexity

$$N_s = O\left(\frac{d^2(L+K)\log\left(\frac{L+K}{\epsilon}\right) + d^2(L+K)^2 \log(d)}{\epsilon^2}\right)$$

ensures an upper bound on the term (SM8.5). The same arguments also hold for the target domain, resulting in the sample complexity

$$N_t = O\left(\frac{d^2(L+K)\log\left(\frac{L+K}{\epsilon}\right) + d^2(L+K)^2 \log(d)}{\epsilon^2}\right)$$

for the number of target samples, which concludes the proof of the theorem. ∎

REFERENCES

[1] P. L. Anthony, M. Bartlett, *Neural Network Learning - Theoretical Foundations*, Cambridge University Press, Cambridge, UK, 2002.

[2] G. Bachman and L. Narici, *Functional Analysis*, Academic Press, New York and London, 1966.

[3] P. L. Bartlett, D. J. Foster, and M. Telgarsky, *Spectrally-normalized margin bounds for neural networks*, in Advances in Neural Information Processing Systems 30, 2017, pp. 6240–6249.

[4] P. L. Bartlett, A. Montanari, and A. Rakhlin, *Deep learning: a statistical viewpoint*, Acta Numerica, 30 (2021), pp. 87–201.

[5] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, *A theory of learning from different domains*, Machine Learning, 79 (2010), pp. 151–175.

[6] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, *Analysis of representations for domain adaptation*, in Proc. Advances in Neural Information Processing Systems 19, 2006, pp. 137–144.

[7] V. I. Bogachev, *Measure Theory*, Springer, Berlin Heidelberg, 2007.

[8] C. Cai, *Deep adaptation networks (DAN) in PyTorch*, 2020. [Online]. Available: https://github.com/CuthbertCai/pytorch_DAN. Accessed: 2024-11-13.

[9] F. Cucker and S. Smale, *On the Mathematical Foundations of Learning*, Bulletin of the American Mathematical Society, 39 (2002), pp. 1–49.

[10] A. Daniely and E. Granot, *On the sample complexity of two-layer networks: Lipschitz vs. element-wise Lipschitz activation*, in International Conference on Algorithmic Learning Theory, vol. 237, 2024, pp. 505–517.

[11] Y. Deng et al., *On the hardness of robustness transfer: A perspective from Rademacher complexity over symmetric difference hypothesis space*, arXiv preprint: http://arxiv.org/abs/2302.12351, (2023).

[12] S. Dhouib, I. Redko, and C. Lartizien, *Margin-aware adversarial domain adaptation with optimal transport*, in Proc. Int. Conf. Machine Learning,, vol. 119, 2020, pp. 2514–2524.

[13] Z. Fang, J. Lu, F. Liu, and G. Zhang, *Semi-supervised heterogeneous domain adaptation: Theory and algorithms*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 1087–1105.

[14] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, *Unsupervised visual domain adaptation using subspace alignment*, in IEEE International Conference on Computer Vision, 2013, pp. 2960–2967.

[15] T. Galanti, L. Wolf, and T. Hazan, *A theoretical framework for deep transfer learning*, Information and Inference: A Journal of the IMA, 5 (2016), pp. 159–209.

[16] Y. Ganin and V. Lempitsky, *Unsupervised domain adaptation by backpropagation*, in Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 1180–1189.

[17] Y. Ganin et al., *Domain-adversarial training of neural networks*, J. Mach. Learn. Res., 17 (2016), pp. 59:1–59:35.

[18] GitHub repository, *Dann_py3*, 2023. [Online]. Available: https://github.com/fungtion/DANN_py3.git.

[19] N. Golowich, A. Rakhlin, and O. Shamir, *Size-independent sample complexity of neural networks*, in Conference On Learning Theory, vol. 75, 2018, pp. 297–299.

[20] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, *A kernel two-sample test*, J. Mach. Learn. Res., 13 (2012), pp. 723–773.

[21] N. Harvey, C. Liaw, and A. Mehrabian, *Nearly-tight VC-dimension bounds for piecewise linear neural networks*, in Proc. Conf. Learning Theory, vol. 65, 2017, pp. 1064–1068.

[22] Y. Jiao, H. Lin, Y. Luo, and J. Z. Yang, *Deep transfer learning: Model framework and error analysis*, arXiv preprint: http://arxiv.org/abs/2410.09383, (2024).

[23] H. Karaca et al., *An experimental study of the sample complexity of domain adaptation*, in IEEE Signal Processing and Communications Applications Conference, 2023, pp. 1–4.

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[25] M. Long, Y. Cao, J. Wang, and M. I. Jordan, *Learning transferable features with deep adaptation networks*, in Proc 32nd International Conference on Machine Learning, vol. 37, pp. 97–105.

[26] Y. Mansour, M. Mohri, and A. Rostamizadeh, *Domain adaptation: Learning bounds and algorithms*, in The 22nd Conference on Learning Theory, 2009.

[27] Massachusetts Institute of Technology, *MIT-CBCL face recognition database*. Available:

http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html.

[28] D. McNamara and M. Balcan, *Risk bounds for transferring representations with and without fine-tuning*, in Proc. Int. Conf. Machine Learning,, vol. 70, 2017, pp. 2373–2381.

[29] M. Mohri and A. M. Medina, *New analysis and algorithm for learning with drifting distributions*, in Int. Conf. Algorithmic Learning Theory, vol. 7568, 2012, pp. 124–138.

[30] B. Neyshabur, S. Bhojanapalli, and N. Srebro, *A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks*, in Int. Conf. Learning Representations, 2018.

[31] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, in Prof. 28th Conference on Learning Theory, vol. 40, 2015, pp. 1376–1401.

[32] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, *A survey on domain adaptation theory*, arXiv preprint: http://arxiv.org/abs/2004.11829, (2020).

[33] A. Sicilia, K. Atwell, M. Alikhani, and S. J. Hwang, *PAC-Bayesian domain adaptation bounds for multiclass learners*, in Proc. Conf. Uncertainty in Artificial Intelligence, vol. 180, 2022, pp. 1824–1834.

[34] M. Subedi and J. Cortez, *Reproducing Kernel Hilbert Spaces - Part III*. https://www.math.uh.edu/~dlabate/LectureNote_06.pdf. Accessed: 2022-03-22.

[35] N. Tripuraneni, M. I. Jordan, and C. Jin, *On the theory of transfer learning: The importance of task diversity*, in Advances in Neural Information Processing Systems, 2020.

[36] G. Vardi, O. Shamir, and N. Srebro, *The sample complexity of one-hidden-layer neural networks*, in Advances in Neural Information Processing Systems 35, 2022.

[37] X. Wang and J. Schneider, *Generalization bounds for transfer learning under model shift*, in Proc. Conf. Uncertainty in Artificial Intelligence, 2015, pp. 922–931.

[38] Z. Wang and Y. Mao, *Information-theoretic analysis of unsupervised domain adaptation*, in Int. Conf. Learning Representations, 2023.

[39] Z. Wang and Y. Mao, *On f-divergence principled domain adaptation: An improved framework*, in Advances in Neural Information Processing Systems, 2024.

[40] B. Wang et al., *Gap minimization for knowledge sharing and transfer*, J. Mach. Learn. Res., 24 (2023), pp. 33:1–33:57.

[41] C. Wei and T. Ma, *Data-dependent sample complexity of deep neural networks via Lipschitz augmentation*, in Advances in Neural Information Processing Systems 32, 2019, pp. 9722–9733.

[42] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, *On the generalization for transfer learning: An information-theoretic analysis*, IEEE Trans. Inf. Theory, 70 (2024), pp. 7089–7124.

[43] V. V. Yurinskii, *Exponential inequalities for sums of random vectors*, Journal of Multivariate Analysis, 6 (1976), pp. 473–499.

[44] W. Zellinger, B. A. Moser, and S. Saminger-Platz, *On generalization in moment-based domain adaptation*, Ann. Math. Artif. Intell., 89 (2021), pp. 333–369.

[45] Y. Zhang, T. Liu, M. Long, and M. I. Jordan, *Bridging theory and algorithm for domain adaptation*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 7404–7413.

[46] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, *Multi-class heterogeneous domain adaptation*, Journal of Machine Learning Research, 20 (2019), pp. 1–31.