

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

3 Elif Vural[†] and Hüseyin Karaca[‡]

5 **SM1. Technical lemmas and proofs.**

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

$$8 \quad \mathcal{L}_\alpha(f^s, f^t, h) \triangleq (1 - \alpha)\mathcal{L}^s(f^s, h) + \alpha\mathcal{L}^t(f^t, h).$$

9 Then the expected target loss is bounded as

$$10 \quad \mathcal{L}^t(f^t, h) \leq \mathcal{L}_\alpha(f^s, f^t, h) + (1 - \alpha)RD(f^s, f^t).$$

11 *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

$$12 \quad \mathcal{L}^t(f^t, h) \leq \mathcal{L}^s(f^s, h) + R D(f^s, f^t).$$

13 Using this above, we obtain

$$14 \quad \begin{aligned} \mathcal{L}^t(f^t, h) &\leq \alpha\mathcal{L}^t(f^t, h) + (1 - \alpha)(\mathcal{L}^s(f^s, h) + R D(f^s, f^t)) \\ &= \mathcal{L}_\alpha(f^s, f^t, h) + (1 - \alpha)RD(f^s, f^t). \end{aligned}$$

■

15 **Lemma SM1.2.** Let the conditions in [Assumption 2.3](#) hold. Let

$$16 \quad \hat{\mathcal{L}}_\alpha(f^s, f^t, h) \triangleq (1 - \alpha)\hat{\mathcal{L}}^s(f^s, h) + \alpha\hat{\mathcal{L}}^t(f^t, h)$$

17 denote the empirical weighted loss. Then, we have

$$18 \quad \begin{aligned} P \left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, h \in \mathcal{H}} |\mathcal{L}_\alpha(f^s, f^t, h) - \hat{\mathcal{L}}_\alpha(f^s, f^t, h)| \leq \epsilon \right) \\ \geq 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}}. \end{aligned}$$

*Submitted to the editors February 17, 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>).

19 *Proof.* We characterize the complexity of function spaces via covering numbers [SM9]. We
 20 first derive a bound for the deviation between the expected and empirical target losses. Let
 21 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
 22 $\mathcal{H} \circ \mathcal{F}^t$ with covering number

$$23 \quad \kappa^t = \mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t).$$

24 Take any $g_k^t = h_k \circ f_k^t$, for $k = 1, \dots, \kappa^t$. The random variables $\ell(g_k^t(x_j^t), \mathbf{y}_j^t)$, $j = 1, \dots, M_t$
 25 are independent identically distributed, bounded as $|\ell(g_k^t(x_j^t), \mathbf{y}_j^t)| \leq A_\ell$, and they have mean
 26 $\mathcal{L}^t(f_k^t, h_k)$. From Hoeffding's inequality, we get that for each k , the deviation between the
 27 empirical loss and the expected loss is bounded as

$$28 \quad P\left(|\hat{\mathcal{L}}^t(f_k^t, h_k) - \mathcal{L}^t(f_k^t, h_k)| \geq \frac{\epsilon}{4\alpha}\right) \leq 2e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}}.$$

29 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

$$30 \quad |\hat{\mathcal{L}}^t(f_k^t, h_k) - \mathcal{L}^t(f_k^t, h_k)| \leq \frac{\epsilon}{4\alpha}$$

31 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
 32 that

$$33 \quad \mathfrak{d}^t(g^t, g_k^t) < \frac{\epsilon}{8\alpha L_\ell}.$$

34 This gives

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

36 It is easy to show similarly that

$$37 \quad |\hat{\mathcal{L}}^t(f^t, h) - \hat{\mathcal{L}}^t(f_k^t, h_k)| < \frac{\epsilon}{8\alpha}.$$

38 Then with probability at least

$$39 \quad 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}}$$

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

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

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

$$44 \quad 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}}$$

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

$$47 \quad |\mathcal{L}^s(f^s, h) - \hat{\mathcal{L}}^s(f^s, h)| < \frac{\epsilon}{2(1-\alpha)}.$$

48 Combining these results, we get that with probability at least

$$49 \quad (\text{SM1.1}) \quad 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}}$$

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

$$51 \quad \begin{aligned} \sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, h \in \mathcal{H}} |\mathcal{L}_\alpha(f^s, f^t, h) - \hat{\mathcal{L}}_\alpha(f^s, f^t, h)| \\ \leq \alpha \sup |\mathcal{L}^t(f^t, h) - \hat{\mathcal{L}}^t(f^t, h)| + (1 - \alpha) \sup |\mathcal{L}^s(f^s, h) - \hat{\mathcal{L}}^s(f^s, h)| \\ \leq \epsilon. \end{aligned} \quad \blacksquare$$

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

$$55 \quad N_s > \frac{\sigma_s^2}{\epsilon^2}, \quad N_t > \frac{\sigma_t^2}{\epsilon^2}.$$

56 Then for the source domain we have

$$57 \quad (\text{SM1.2}) \quad \begin{aligned} P \left(\left\| \frac{1}{N_s} \sum_{i=1}^{N_s} f^s(x_i^s) - E[f^s(x^s)] \right\| \geq \epsilon \right) \\ \leq \exp \left(-\frac{1}{8} \left(\frac{\sqrt{N_s} \epsilon}{\sigma_s} - 1 \right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s} \epsilon}{\sigma_s} - 1 \right) \frac{C_s}{2\sqrt{N_s} \sigma_s}} \right) \end{aligned}$$

58 and for the target domain we have

$$59 \quad (\text{SM1.3}) \quad \begin{aligned} P \left(\left\| \frac{1}{N_t} \sum_{j=1}^{N_t} f^t(x_j^t) - E[f^t(x^t)] \right\| \geq \epsilon \right) \\ \leq \exp \left(-\frac{1}{8} \left(\frac{\sqrt{N_t} \epsilon}{\sigma_t} - 1 \right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_t} \epsilon}{\sigma_t} - 1 \right) \frac{C_t}{2\sqrt{N_t} \sigma_t}} \right). \end{aligned}$$

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

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

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

64 If $x > \beta_N / B_N$ where

65 (SM1.5)
$$\beta_N \geq E[\|\zeta_1 + \dots + \zeta_N\|], \quad B_N^2 = b_1^2 + \dots + b_N^2,$$

66 then

67
$$P(\|\zeta_1 + \dots + \zeta_N\| \geq xB_N) \leq \exp\left(-\frac{1}{8}\left(x - \frac{\beta_N}{B_N}\right)^2 \frac{1}{1 + \left(x - \frac{\beta_N}{B_N}\right) \frac{C}{2B_N}}\right).$$

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

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

74
$$E[\|\zeta_i\|^2] \leq \sigma_s^2.$$

75 We consider now

76
$$\begin{aligned} E\left[\left\|\sum_{i=1}^{N_s} \zeta_i\right\|^2\right] &= E\left[\left\langle\sum_{i=1}^{N_s} \zeta_i, \sum_{j=1}^{N_s} \zeta_j\right\rangle\right] = \sum_{i=1}^{N_s} \sum_{j=1}^{N_s} E[\langle \zeta_i, \zeta_j \rangle] \\ &= \sum_{i=1}^{N_s} E[\langle \zeta_i, \zeta_i \rangle] + \sum_{i=1}^{N_s} \sum_{j \neq i, j=1}^{N_s} E[\langle \zeta_i, \zeta_j \rangle] \leq \sigma_s^2 N_s \end{aligned}$$

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

80
$$Y = \left\|\sum_{i=1}^{N_s} \zeta_i\right\|$$

81 then yields

82
$$E\left[\left\|\sum_{i=1}^{N_s} \zeta_i\right\|\right] \leq \left(E\left[\left\|\sum_{i=1}^{N_s} \zeta_i\right\|^2\right]\right)^{1/2} \leq \sigma_s \sqrt{N_s}.$$

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

84 (SM1.6)
$$E[\|\zeta_1 + \dots + \zeta_{N_s}\|] \leq \beta_{N_s}.$$

85 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
86 assumption $N_s > \sigma_s^2/\epsilon^2$, the following choice for x

87
$$x = \frac{\sqrt{N_s} \epsilon}{\sigma_s} > 1$$

88 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

89
$$P(\|\zeta_1 + \dots + \zeta_{N_s}\| \geq N_s \epsilon) \leq \exp\left(-\frac{1}{8} \left(\frac{\sqrt{N_s} \epsilon}{\sigma_s} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s} \epsilon}{\sigma_s} - 1\right) \frac{C_s}{2\sqrt{N_s} \sigma_s}}\right).$$

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

91
$$\begin{aligned} P\left(\left\|\frac{1}{N_s} \sum_{i=1}^{N_s} f^s(x_i^s) - E[f^s(x^s)]\right\| \geq \epsilon\right) \\ \leq \exp\left(-\frac{1}{8} \left(\frac{\sqrt{N_s} \epsilon}{\sigma_s} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s} \epsilon}{\sigma_s} - 1\right) \frac{C_s}{2\sqrt{N_s} \sigma_s}}\right). \end{aligned}$$

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

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

96
$$N_s > \frac{16\sigma_s^2}{\epsilon^2}, \quad N_t > \frac{16\sigma_t^2}{\epsilon^2}.$$

97 Let us define the functions

98
$$\begin{aligned} a_s(N_s, \epsilon) &\triangleq \frac{1}{8} \left(\frac{\sqrt{N_s} \epsilon}{4\sigma_s} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_s} \epsilon}{4\sigma_s} - 1\right) \frac{C_s}{2\sqrt{N_s} \sigma_s}} \\ a_t(N_t, \epsilon) &\triangleq \frac{1}{8} \left(\frac{\sqrt{N_t} \epsilon}{4\sigma_t} - 1\right)^2 \frac{1}{1 + \left(\frac{\sqrt{N_t} \epsilon}{4\sigma_t} - 1\right) \frac{C_t}{2\sqrt{N_t} \sigma_t}}. \end{aligned}$$

99 Then

100
$$\begin{aligned} P\left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t} |D(f^s, f^t) - \hat{D}(f^s, f^t)| < \epsilon\right) \\ \geq 1 - \mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^s) \exp(-a_s(N_s, \epsilon)) - \mathcal{N}(\mathcal{F}^t, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^t) \exp(-a_t(N_t, \epsilon)). \end{aligned}$$

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

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

105 Replacing ϵ by $\epsilon/4$ in [Lemma SM1.3](#), we observe that with probability at least

$$106 \quad 1 - \exp(-a_s(N_s, \epsilon)) - \exp(-a_t(N_t, \epsilon))$$

107 we have

$$108 \quad \left\| \frac{1}{N_s} \sum_{i=1}^{N_s} f^s(x_i^s) - E[f^s(x^s)] \right\| \leq \frac{\epsilon}{4}, \quad \left\| \frac{1}{N_t} \sum_{j=1}^{N_t} f^t(x_j^t) - E[f^t(x^t)] \right\| \leq \frac{\epsilon}{4}$$

109 which yields from [\(SM1.7\)](#)

$$110 \quad |D(f^s, f^t) - \hat{D}(f^s, f^t)| \leq \frac{\epsilon}{2}.$$

111 In order to extend the above bound to the whole space of transformations, we consider
 112 covers of the function classes \mathcal{F}^s and \mathcal{F}^t , consisting of open balls of radius $\epsilon/8$ respectively
 113 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

$$114 \quad \kappa^s = \mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^s), \quad \kappa^t = \mathcal{N}(\mathcal{F}^t, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^t).$$

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

$$116 \quad 1 - \kappa^s \exp(-a_s(N_s, \epsilon)) - \kappa^t \exp(-a_t(N_t, \epsilon))$$

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

$$118 \quad (SM1.8) \quad |D(f_k^s, f_l^t) - \hat{D}(f_k^s, f_l^t)| \leq \frac{\epsilon}{2}.$$

119 Now, let us consider an arbitrary pair of transformations $f^s \in \mathcal{F}^s$ and $f^t \in \mathcal{F}^t$. As the
 120 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
 121 transformation f_k^s and a target transformation f_l^t such that

$$122 \quad \mathfrak{d}_{\mathcal{X}}^s(f^s, f_k^s) < \frac{\epsilon}{8}, \quad \mathfrak{d}_{\mathcal{X}}^t(f^t, f_l^t) < \frac{\epsilon}{8}.$$

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

$$125 \quad (\text{SM1.9}) \quad |D(f^s, f^t) - \hat{D}(f^s, f^t)| \leq |D(f^s, f^t) - D(f_k^s, f_l^t)| + |D(f_k^s, f_l^t) - \hat{D}(f_k^s, f_l^t)| \\ + |\hat{D}(f_k^s, f_l^t) - \hat{D}(f^s, f^t)|$$

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

$$128 \quad (\text{SM1.10}) \quad |D(f^s, f^t) - D(f_k^s, f_l^t)| = \|E[f^s(x^s)] - E[f^t(x^t)]\| - \|E[f_k^s(x^s)] - E[f_l^t(x^t)]\| \\ \leq \|E[f^s(x^s)] - E[f_k^s(x^s)]\| + \|E[f^t(x^t)] - E[f_l^t(x^t)]\| \\ = \|E[f^s(x^s) - f_k^s(x^s)]\| + \|E[f^t(x^t) - f_l^t(x^t)]\| \\ \leq E[\|f^s(x^s) - f_k^s(x^s)\|] + E[\|f^t(x^t) - f_l^t(x^t)\|]$$

129 where the last inequality follows from Jensen's inequality, observing the fact that a norm over
 130 a Hilbert space is a convex function. From the definition of the metrics $\mathfrak{d}_{\mathcal{X}}^s$ and $\mathfrak{d}_{\mathcal{X}}^t$, we have

$$131 \quad \|f^s(x^s) - f_k^s(x^s)\| \leq \mathfrak{d}_{\mathcal{X}}^s(f^s, f_k^s) \\ \|f^t(x^t) - f_l^t(x^t)\| \leq \mathfrak{d}_{\mathcal{X}}^t(f^t, f_l^t)$$

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

$$133 \quad |D(f^s, f^t) - D(f_k^s, f_l^t)| \leq \mathfrak{d}_{\mathcal{X}}^s(f^s, f_k^s) + \mathfrak{d}_{\mathcal{X}}^t(f^t, f_l^t) < \frac{\epsilon}{8} + \frac{\epsilon}{8} = \frac{\epsilon}{4}.$$

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

$$136 \quad |\hat{D}(f_k^s, f_l^t) - \hat{D}(f^s, f^t)| < \frac{\epsilon}{4}.$$

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

$$139 \quad 1 - \kappa^s \exp(-a_s(N_s, \epsilon)) - \kappa^t \exp(-a_t(N_t, \epsilon))$$

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

$$141 \quad |D(f^s, f^t) - \hat{D}(f^s, f^t)| < \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.$$

142 Hence, we get the stated result

$$143 \quad P \left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t} |D(f^s, f^t) - \hat{D}(f^s, f^t)| < \epsilon \right) \\ \geq 1 - \kappa^s \exp(-a_s(N_s, \epsilon)) - \kappa^t \exp(-a_t(N_t, \epsilon)) \\ = 1 - \mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^s) \exp(-a_s(N_s, \epsilon)) - \mathcal{N}(\mathcal{F}^t, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^t) \exp(-a_t(N_t, \epsilon)).$$
■

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

$$\begin{aligned} 148 \quad E[f^{sl}(x^s)](\cdot) &\triangleq E[f^{sl}(x^s)(\cdot)] \\ E[f^{tl}(x^t)](\cdot) &\triangleq E[f^{tl}(x^t)(\cdot)] \end{aligned}$$

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

$$\begin{aligned} 151 \quad E[f^s(x^s)] &\triangleq (E[f^{s1}(x^s)], \dots, E[f^{s(L-1)}(x^s)]) \\ E[f^t(x^t)] &\triangleq (E[f^{t1}(x^t)], \dots, E[f^{t(L-1)}(x^t)]) \end{aligned}$$

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

153 **Proof.** We prove the statements only for the source domain, as the proofs for the target
 154 domain are the same. Let $\xi^{sl}(x^s) \in \mathbb{R}^{d_l}$ denote the feature in layer l for the source input
 155 $x^s \in \mathbb{R}^{d_0}$, where we regard $\xi^{sl}(\cdot) : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_l}$ as a function. In the relation

$$156 \quad \xi^{sl}(x^s) = \eta^l(\mathbf{W}^{sl}\xi^{s(l-1)}(x^s) + \mathbf{b}^{sl})$$

157 the expression $\mathbf{W}^{sl}\xi^{s(l-1)}(x^s) + \mathbf{b}^{sl}$ is a continuous mapping of $\xi^{s(l-1)}(x^s)$, and the function
 158 η^l is continuous. Hence, based on a simple induction argument it follows that $\xi^{sl}(\cdot) : \mathcal{X}^s =$
 159 $\mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_l}$ is a continuous, thus measurable function (a Borel map).

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

$$164 \quad \lim_{n \rightarrow \infty} \|\phi^l(\xi_n^{sl}) - \phi^l(\xi_*^{sl})\|_{\mathcal{X}^l}^2 = \lim_{n \rightarrow \infty} \left(k^l(\xi_n^{sl}, \xi_n^{sl}) - 2k^l(\xi_n^{sl}, \xi_*^{sl}) + k^l(\xi_*^{sl}, \xi_*^{sl}) \right) = 0$$

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

$$166 \quad \lim_{n \rightarrow \infty} \phi^l(\xi_n^{sl}) = \phi^l(\xi_*^{sl})$$

167 and hence $\phi^l : \mathbb{R}^{d_l} \rightarrow \mathcal{X}^l$ is a continuous function. ϕ^l is thus measurable with respect to the
 168 Borel σ -algebra $\mathcal{B}(\mathcal{X}^l)$ of the RKHS \mathcal{X}^l . Since $\xi^{sl}(\cdot) : \mathcal{X}^s \rightarrow \mathbb{R}^{d_l}$ is a measurable mapping as
 169 well, we conclude that the mapping $f^{sl} = \phi^l(\xi^{sl}(\cdot)) : \mathcal{X}^s \rightarrow \mathcal{X}^l$ is measurable with respect to
 170 $\mathcal{B}(\mathcal{X}^l)$, for $l = 1, \dots, L - 1$.

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

$$174 \quad \mathcal{B}(\mathcal{X}) = \bigotimes_{l=1}^{L-1} \mathcal{B}(\mathcal{X}^l)$$

175 where the right hand side denotes the σ -algebra generated by all finite products of Borel
 176 sets in $\mathcal{B}(\mathcal{X}^l)$'s [SM7]. Hence, denoting the set product of some collection of Borel sets
 177 $B^1 \in \mathcal{B}(\mathcal{X}^1), \dots, B^{L-1} \in \mathcal{B}(\mathcal{X}^{L-1})$ as

$$178 \quad B^1 \times B^2 \times \dots \times B^{L-1} = \{(f^1, f^2, \dots, f^{L-1}) : f^l \in B^l, l = 1, \dots, L-1\},$$

179 the σ -algebra generated by

$$180 \quad B = \{B^1 \times \dots \times B^{L-1} : B^1 \in \mathcal{B}(\mathcal{X}^1), \dots, B^{L-1} \in \mathcal{B}(\mathcal{X}^{L-1})\}$$

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

$$\begin{aligned} 184 \quad (f^s)^{-1}(B^1 \times \dots \times B^{L-1}) &= \{x^s \in \mathcal{X}^s : f^s(x^s) \in B^1 \times \dots \times B^{L-1}\} \\ &= \{x^s \in \mathcal{X}^s : f^{s1}(x^s) \in B^1, \dots, f^{s(L-1)}(x^s) \in B^{L-1}\} \\ &= \bigcap_{l=1}^{L-1} (f^{sl})^{-1}(B^l). \end{aligned}$$

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

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

$$189 \quad f^{sl}(\cdot)(\xi) = k^l(\xi^{sl}(\cdot), \xi).$$

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

$$194 \quad E_{x^s}[f^{sl}(x^s)](\xi) \triangleq E_{x^s}[f^{sl}(x^s)(\xi)].$$

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

$$197 \quad T_{\mu_s}(\psi) \triangleq E_{x^s}[\psi(\xi^{sl})]$$

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

$$\begin{aligned} 200 \quad |T_{\mu_s}(\psi)| &= \left| E_{x^s}[\psi(\xi^{sl})] \right| \leq E_{x^s} \left[|\psi(\xi^{sl})| \right] = E_{x^s} \left[\left| \langle k^l(\xi^{sl}, \cdot), \psi(\cdot) \rangle_{\mathcal{X}^l} \right| \right] \\ &\leq E_{x^s} \left[\|k^l(\xi^{sl}, \cdot)\|_{\mathcal{X}^l} \|\psi\|_{\mathcal{X}^l} \right] = E_{x^s} \left[\sqrt{k^l(\xi^{sl}, \xi^{sl})} \right] \|\psi\|_{\mathcal{X}^l}. \end{aligned}$$

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

203
$$T_{\mu_s}(\psi) = \langle \psi, \psi^{sl} \rangle_{\mathcal{X}^l}$$

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

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

206 But it also holds that

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

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

209
$$E_{x^s}[f^{sl}(x^s)] = \psi^{sl} \in \mathcal{X}^l.$$

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

211
$$E_{x^s}[f^s(x^s)] \triangleq (E_{x^s}[f^{s1}(x^s)], \dots, E_{x^s}[f^{s(L-1)}(x^s)])$$
 ■

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

213 **Lemma SM1.6.** *Let Assumptions 3.1-3.3 hold. Then, the transformation function classes
 214 $\mathcal{F}^s, \mathcal{F}^t$ in (3.6) and the composite function classes $\mathcal{G}^s, \mathcal{G}^t$ in (3.8) are compact metric spaces,
 215 respectively under the metrics $\mathfrak{d}_{\mathcal{X}}^s, \mathfrak{d}_{\mathcal{X}}^t$ in (2.10), and the metrics $\mathfrak{d}^s, \mathfrak{d}^t$ in (2.4).*

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

218
$$\Phi^s = \{\Theta^s = (\Theta^{s1}, \dots, \Theta^{sL}) : |\Theta_{ij}^{sl}| \leq A_\Theta, \forall i, j, l\}$$

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

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

224 (SM1.13)
$$\mathcal{M}_{\mathcal{F}^s}(\Theta^s) = f_{\Theta^s}^s = (f_{\Theta^s}^{s1}, \dots, f_{\Theta^s}^{s(L-1)})$$

225 where the notation $f_{\Theta^s}^s(x^s)$ stands for the function $f^s(x^s)$ defined in (3.6) by explicitly referring
 226 to its dependence on the network parameters Θ^s . In the following, we show that the mapping
 227 $\mathcal{M}_{\mathcal{F}^s}$ is continuous. Let us consider a sequence $\{\Theta_n^s\} \subset \Phi^s$ converging to an element $\Theta_*^s \in \Phi^s$.
 228 Since the relation (3.1) between the features of adjacent layers is given by a linear mapping

229 followed by a continuous activation function η^l , the mapping $\xi_{\Theta^s}^{sl}(x^s)$ is a continuous function
230 of Θ^s , i.e.

231 (SM1.14)
$$\lim_{n \rightarrow \infty} \xi_{\Theta_n^s}^{sl}(x^s) = \xi_{\Theta_*^s}^{sl}(x^s).$$

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

236
$$\|\xi_{\Theta_n^s}^{sl}(x^s) - \xi_{\Theta_*^s}^{sl}(x^s)\| < \epsilon$$

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

238
$$\begin{aligned} \|f_{\Theta_n^s}^{sl}(x^s) - f_{\Theta_*^s}^{sl}(x^s)\|_{\mathcal{X}^l}^2 &= \|\phi^l(\xi_{\Theta_n^s}^{sl}(x^s)) - \phi^l(\xi_{\Theta_*^s}^{sl}(x^s))\|_{\mathcal{X}^l}^2 \\ &= k^l(\xi_{\Theta_n^s}^{sl}(x^s), \xi_{\Theta_n^s}^{sl}(x^s)) - 2k^l(\xi_{\Theta_n^s}^{sl}(x^s), \xi_{\Theta_*^s}^{sl}(x^s)) + k^l(\xi_{\Theta_*^s}^{sl}(x^s), \xi_{\Theta_*^s}^{sl}(x^s)) \\ &\leq 2L_K \|\xi_{\Theta_n^s}^{sl}(x^s) - \xi_{\Theta_*^s}^{sl}(x^s)\| < 2L_K \epsilon \end{aligned}$$

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

240
$$\|f_{\Theta_n^s}^s(x^s) - f_{\Theta_*^s}^s(x^s)\|_{\mathcal{X}}^2 = \sum_{l=1}^{L-1} \|f_{\Theta_n^s}^{sl}(x^s) - f_{\Theta_*^s}^{sl}(x^s)\|_{\mathcal{X}^l}^2 < 2(L-1)L_K \epsilon.$$

241 We have thus obtained

242
$$\|f_{\Theta_n^s}^s(x^s) - f_{\Theta_*^s}^s(x^s)\|_{\mathcal{X}} < \sqrt{2(L-1)L_K} \sqrt{\epsilon}$$

243 for all $n \geq n_0$ and for all $x^s \in \mathcal{X}^s$, which shows that $f_{\Theta_n^s}^s(x^s)$ converges to $f_{\Theta_*^s}^s(x^s)$ uniformly
244 on \mathcal{X}^s . Then we have

245
$$\begin{aligned} \lim_{n \rightarrow \infty} \mathfrak{d}_{\mathcal{X}}^s(f_{\Theta_n^s}^s, f_{\Theta_*^s}^s) &= \lim_{n \rightarrow \infty} \sup_{x^s \in \mathcal{X}^s} \|f_{\Theta_n^s}^s(x^s) - f_{\Theta_*^s}^s(x^s)\|_{\mathcal{X}} \\ &= \sup_{x^s \in \mathcal{X}^s} \lim_{n \rightarrow \infty} \|f_{\Theta_n^s}^s(x^s) - f_{\Theta_*^s}^s(x^s)\|_{\mathcal{X}} = 0 \end{aligned}$$

246 where the second equality follows from the uniform convergence of $f_{\Theta_n^s}^s(x^s)$. We have thus
247 shown that the mapping $\mathcal{M}_{\mathcal{F}^s} : \Phi^s \rightarrow \mathcal{F}^s$ defined in (SM1.13) is continuous. Since the set Φ^s
248 is compact, we conclude that the function space \mathcal{F}^s is a compact metric space.

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

254
$$\lim_{n \rightarrow \infty} g_{\Theta_n^s}^s(x^s) = \lim_{n \rightarrow \infty} \xi_{\Theta_n^s}^{sL}(x^s) = \xi_{\Theta_*^s}^{sL}(x^s) = g_{\Theta_*^s}^s(x^s)$$

255 uniformly. Hence,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathfrak{d}^s(g_{\Theta_n^s}^s, g_{\Theta_*^s}^s) &= \lim_{n \rightarrow \infty} \sup_{x^s \in \mathcal{X}^s} \|g_{\Theta_n^s}^s(x^s) - g_{\Theta_*^s}^s(x^s)\| \\ 256 \quad &= \sup_{x^s \in \mathcal{X}^s} \lim_{n \rightarrow \infty} \|g_{\Theta_n^s}^s(x^s) - g_{\Theta_*^s}^s(x^s)\| = 0. \end{aligned}$$

257 Hence, the mapping $\mathcal{M}_{\mathcal{G}^s} : \Phi^s \rightarrow \mathcal{G}^s$ is continuous. Then, from the compactness of Φ^s , it
258 follows that the function space \mathcal{G}^s is compact as well. ■

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

$$\begin{aligned} 261 \quad \mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}_{\mathcal{X}}^s) &\leq \prod_{l=1}^{L-1} \left(\frac{4A_{\Theta}L_KQ}{\epsilon^2} + 1 \right)^{d_l(d_{l-1}+1)} \\ \mathcal{N}(\mathcal{F}^t, \epsilon, \mathfrak{d}_{\mathcal{X}}^t) &\leq \prod_{l=1}^{L-1} \left(\frac{4A_{\Theta}L_KQ}{\epsilon^2} + 1 \right)^{d_l(d_{l-1}+1)} \end{aligned}$$

262 where the dimension-dependent constant Q is defined as

$$263 \quad Q \triangleq \sum_{l=1}^{L-1} Q_l$$

264 with

$$\begin{aligned} 265 \quad (SM1.15) \quad Q_l &\triangleq (L_{\eta}R_{l-1}\sqrt{d_l d_{l-1}} + L_{\eta}\sqrt{d_l}) \\ &+ \sum_{i=1}^{l-1} (L_{\eta}R_{i-1}\sqrt{d_i d_{i-1}} + L_{\eta}\sqrt{d_i}) \prod_{k=i+1}^l L_{\eta}A_{\Theta}\sqrt{d_k d_{k-1}} \end{aligned}$$

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

$$\begin{aligned} 267 \quad R_l &\triangleq (A_{\eta}A_{\Theta})^l (A_x\sqrt{d_0} + 1) \sqrt{d_1} \prod_{k=1}^{l-1} \sqrt{d_{k+1}d_k} \\ &+ \sum_{i=2}^{l-1} (A_{\eta}A_{\Theta})^{l+1-i} \sqrt{d_i} \prod_{k=i}^{l-1} \sqrt{d_{k+1}d_k} + A_{\eta}A_{\Theta}\sqrt{d_l} \end{aligned}$$

268 under condition (3.15) and $R_l \triangleq C_{\eta}\sqrt{d_l}$ under condition (3.14) for $l = 2, \dots, L-1$, where
269 $R_0 \triangleq A_x$ and $R_1 \triangleq A_{\eta}A_{\Theta}\sqrt{d_1 d_0} A_x + A_{\eta}A_{\Theta}\sqrt{d_1}$.

270 **Proof.** We obtain the bound only for the source domain, as the derivation for the target
271 domain is identical. Our proof is based on constructing an ϵ -cover for the compact metric
272 space \mathcal{F}^s . For two mappings $f_1^s, f_2^s \in \mathcal{F}^s$ defined respectively by the parameter vectors Θ_1^s, Θ_2^s

273 we have

$$\begin{aligned}
 (\delta_{\mathcal{X}}^s(f_1^s, f_2^s))^2 &= \sup_{x^s \in \mathcal{X}^s} \|f_1^s(x^s) - f_2^s(x^s)\|_{\mathcal{X}}^2 \\
 &= \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} \|\phi^l(\xi_{\Theta_1^s}^{sl}(x^s)) - \phi^l(\xi_{\Theta_2^s}^{sl}(x^s))\|_{\mathcal{X}^l}^2 \\
 &= \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} k^l \left(\xi_{\Theta_1^s}^{sl}(x^s), \xi_{\Theta_1^s}^{sl}(x^s) \right) - 2k^l \left(\xi_{\Theta_1^s}^{sl}(x^s), \xi_{\Theta_2^s}^{sl}(x^s) \right) \\
 &\quad + k^l \left(\xi_{\Theta_2^s}^{sl}(x^s), \xi_{\Theta_2^s}^{sl}(x^s) \right) \\
 274 \quad (\text{SM1.16}) \quad &\leq \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} \left| k^l \left(\xi_{\Theta_1^s}^{sl}(x^s), \xi_{\Theta_1^s}^{sl}(x^s) \right) - k^l \left(\xi_{\Theta_1^s}^{sl}(x^s), \xi_{\Theta_2^s}^{sl}(x^s) \right) \right| \\
 &\quad + \left| k^l \left(\xi_{\Theta_2^s}^{sl}(x^s), \xi_{\Theta_2^s}^{sl}(x^s) \right) - k^l \left(\xi_{\Theta_1^s}^{sl}(x^s), \xi_{\Theta_2^s}^{sl}(x^s) \right) \right| \\
 &\leq \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} 2L_K \|\xi_{\Theta_1^s}^{sl}(x^s) - \xi_{\Theta_2^s}^{sl}(x^s)\|
 \end{aligned}$$

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

$$278 \quad (\text{SM1.17}) \quad \Theta^l = \{\Theta^l = [\mathbf{W}^l \ \mathbf{b}^l] \in \mathbb{R}^{d_l \times (d_{l-1}+1)} : |\mathbf{W}_{ij}^l| \leq A_{\Theta}, |\mathbf{b}_i^l| \leq A_{\Theta}, \forall i, j, l\}.$$

279 Then there exists a cover of Θ^l consisting of open balls around a set $\mathfrak{G}^l = \{\Theta_m^l\}_{m=1}^{\kappa^l}$ of regu-
280 larly sampled grid points, with a distance of δ between adjacent grid centers in each dimen-
281 sion. The maximal overall distance between two adjacent grid centers is then $\delta\sqrt{d_l(d_{l-1}+1)}$.
282 Hence, the distance between any parameter vector $\Theta^l \in \Theta^l$ and the nearest grid center Θ_m^l
283 is at most

$$284 \quad \frac{\delta\sqrt{d_l(d_{l-1}+1)}}{2}$$

285 with the number of balls in the cover being

$$286 \quad \kappa^l = \left(\frac{2A_{\Theta}}{\delta} + 1 \right)^{d_l(d_{l-1}+1)}.$$

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

$$289 \quad (\text{SM1.18}) \quad \mathfrak{G} = \mathfrak{G}^1 \times \dots \times \mathfrak{G}^{L-1} = \{\Theta_k\}_{k=1}^{\kappa^1 \dots \kappa^{L-1}}$$

290 which defines a cover for the overall parameter space

$$291 \quad \Phi = \{\Theta = (\Theta^1, \dots, \Theta^{L-1}) : |\Theta_{ij}^l| \leq A_{\Theta}, \forall i, j, l\}$$

292 consisting of

$$293 \quad \kappa_{\mathfrak{G}} = \prod_{l=1}^{L-1} \kappa^l = \prod_{l=1}^{L-1} \left(\frac{2A_{\Theta}}{\delta} + 1 \right)^{d_l(d_{l-1}+1)}$$

294 balls. Then for any $f_k^s \in \mathcal{F}^s$ with parameters Θ^s , there exists some $f_k^s \in \mathcal{F}^s$ with parameters
295 $\Theta_k = (\Theta_k^1, \Theta_k^2, \dots, \Theta_k^{L-1}) \in \mathfrak{G}$ in the product grid such that

$$296 \quad (\text{SM1.19}) \quad \|\Theta^{sl} - \Theta_k^l\| < \delta \sqrt{d_l(d_{l-1} + 1)}.$$

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

(SM1.20)

$$\begin{aligned} & \|\xi_{\Theta^s}^{sl}(x^s) - \xi_{\Theta_k}^l(x^s)\| = \left\| \eta^l \left(\mathbf{W}^{sl} \xi_{\Theta^s}^{s(l-1)}(x^s) + \mathbf{b}^{sl} \right) - \eta^l \left(\mathbf{W}_k^l \xi_{\Theta_k}^{l-1}(x^s) + \mathbf{b}_k^l \right) \right\| \\ 299 & \leq L_{\eta} \left\| \mathbf{W}^{sl} \xi_{\Theta^s}^{s(l-1)}(x^s) + \mathbf{b}^{sl} - \mathbf{W}_k^l \xi_{\Theta_k}^{l-1}(x^s) - \mathbf{b}_k^l \right\| \\ & = L_{\eta} \left\| \mathbf{W}^{sl} \xi_{\Theta^s}^{s(l-1)}(x^s) - \mathbf{W}^{sl} \xi_{\Theta_k}^{l-1}(x^s) + \mathbf{W}^{sl} \xi_{\Theta_k}^{l-1}(x^s) - \mathbf{W}_k^l \xi_{\Theta_k}^{l-1}(x^s) + \mathbf{b}^{sl} - \mathbf{b}_k^l \right\| \\ & \leq L_{\eta} \|\mathbf{W}^{sl}\| \|\xi_{\Theta^s}^{s(l-1)}(x^s) - \xi_{\Theta_k}^{l-1}(x^s)\| + L_{\eta} \|\mathbf{W}^{sl} - \mathbf{W}_k^l\| \|\xi_{\Theta_k}^{l-1}(x^s)\| + L_{\eta} \|\mathbf{b}^{sl} - \mathbf{b}_k^l\| \end{aligned}$$

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

$$\begin{aligned} & \|\mathbf{W}^{sl}\| \leq \|\mathbf{W}^{sl}\|_F \leq A_{\Theta} \sqrt{d_l d_{l-1}} \\ 303 & \|\mathbf{W}^{sl} - \mathbf{W}_k^l\| \leq \|\mathbf{W}^{sl} - \mathbf{W}_k^l\|_F < \delta \sqrt{d_l d_{l-1}} \\ & \|\mathbf{b}^{sl} - \mathbf{b}_k^l\| < \delta \sqrt{d_l}. \end{aligned}$$

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

$$\begin{aligned} 305 \quad (\text{SM1.21}) \quad & \|\xi_{\Theta^s}^{sl}(x^s) - \xi_{\Theta_k}^l(x^s)\| < L_{\eta} A_{\Theta} \sqrt{d_l d_{l-1}} \|\xi_{\Theta^s}^{s(l-1)}(x^s) - \xi_{\Theta_k}^{l-1}(x^s)\| \\ & + L_{\eta} \delta \sqrt{d_l d_{l-1}} \|\xi_{\Theta_k}^{l-1}(x^s)\| + L_{\eta} \delta \sqrt{d_l}. \end{aligned}$$

306 In order to study (SM1.21), we first obtain an upper bound on the term $\|\xi_{\Theta_k}^l(x^s)\|$. Notice
307 that for the condition (3.14), we simply have

$$\begin{aligned} 308 \quad (\text{SM1.22}) \quad & \|\xi_{\Theta_k}^l(x^s)\| = \|\eta^l \left(\mathbf{W}^l \xi_{\Theta_k}^{l-1}(x^s) + \mathbf{b}^l \right)\| = \left(\sum_{i=1}^{d_l} \left(\eta_i^l (\mathbf{W}^l \xi_{\Theta_k}^{l-1}(x^s) + \mathbf{b}^l) \right)^2 \right)^{1/2} \\ & \leq C_{\eta} \sqrt{d_l}. \end{aligned}$$

309 Next, for the condition (3.15) we have

$$\begin{aligned} & \|\xi_{\Theta_k}^0(x^s)\| = \|x^s\| \leq A_x \\ 310 & \|\xi_{\Theta_k}^1(x^s)\| = \|\eta^1 (\mathbf{W}^1 \xi_{\Theta_k}^0(x^s) + \mathbf{b}^1)\| \leq A_{\eta} \|\mathbf{W}^1 \xi_{\Theta_k}^0(x^s) + \mathbf{b}^1\| \\ & \leq A_{\eta} (\|\mathbf{W}^1\| \|\xi_{\Theta_k}^0(x^s)\| + \|\mathbf{b}^1\|) \leq A_{\eta} A_{\Theta} \sqrt{d_1 d_0} A_x + A_{\eta} A_{\Theta} \sqrt{d_1} \end{aligned}$$

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

$$\begin{aligned} \| \boldsymbol{\xi}_{\Theta_k}^l(x^s) \| &\leq A_\eta \left(\|\mathbf{W}^l\| \|\boldsymbol{\xi}_{\Theta_k}^{l-1}(x^s)\| + \|\mathbf{b}^l\| \right) \\ &\leq A_\eta A_\Theta \sqrt{d_l d_{l-1}} \|\boldsymbol{\xi}_{\Theta_k}^{l-1}(x^s)\| + A_\eta A_\Theta \sqrt{d_l} \\ 313 &\leq (A_\eta A_\Theta)^l (A_x \sqrt{d_0} + 1) \sqrt{d_1} \prod_{k=1}^{l-1} \sqrt{d_{k+1} d_k} \\ &+ \sum_{i=2}^{l-1} (A_\eta A_\Theta)^{l+1-i} \sqrt{d_i} \prod_{k=1}^{l-1} \sqrt{d_{k+1} d_k} + A_\eta A_\Theta \sqrt{d_l}. \end{aligned}$$

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

$$315 \quad (\text{SM1.23}) \quad \| \boldsymbol{\xi}_{\Theta_k}^l(x^s) \| \leq R_l$$

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

$$318 \quad (\text{SM1.24}) \quad \begin{aligned} \| \boldsymbol{\xi}_{\Theta_s}^{sl}(x^s) - \boldsymbol{\xi}_{\Theta_k}^l(x^s) \| &< L_\eta A_\Theta \sqrt{d_l d_{l-1}} \|\boldsymbol{\xi}_{\Theta_s}^{s(l-1)}(x^s) - \boldsymbol{\xi}_{\Theta_k}^{l-1}(x^s)\| \\ &+ L_\eta \delta \sqrt{d_l d_{l-1}} R_{l-1} + L_\eta \delta \sqrt{d_l}. \end{aligned}$$

319 For layer $l = 1$, we have

$$320 \quad \begin{aligned} \| \boldsymbol{\xi}_{\Theta_s}^1(x^s) - \boldsymbol{\xi}_{\Theta_k}^1(x^s) \| &< L_\eta A_\Theta \sqrt{d_1 d_0} \|\boldsymbol{\xi}_{\Theta_s}^0(x^s) - \boldsymbol{\xi}_{\Theta_k}^0(x^s)\| \\ &+ L_\eta \delta \sqrt{d_1 d_0} R_0 + L_\eta \delta \sqrt{d_1} \\ &= L_\eta \delta \sqrt{d_1 d_0} R_0 + L_\eta \delta \sqrt{d_1} \end{aligned}$$

321 since $\boldsymbol{\xi}_{\Theta_s}^0(x^s) = \boldsymbol{\xi}_{\Theta_k}^0(x^s) = x^s$. This relation together with the recursive inequality in (SM1.24)
 322 yields

$$323 \quad (\text{SM1.25}) \quad \begin{aligned} \| \boldsymbol{\xi}_{\Theta_s}^{sl}(x^s) - \boldsymbol{\xi}_{\Theta_k}^l(x^s) \| &< \delta \left((L_\eta R_{l-1} \sqrt{d_l d_{l-1}} + L_\eta \sqrt{d_l}) \right. \\ &+ \sum_{i=1}^{l-1} (L_\eta R_{i-1} \sqrt{d_i d_{i-1}} + L_\eta \sqrt{d_i}) \left. \prod_{k=i+1}^l L_\eta A_\Theta \sqrt{d_k d_{k-1}} \right) \\ &= Q_l \delta \end{aligned}$$

324 for $l = 1, \dots, L - 1$. Hence, we have shown that for any $f^s \in \mathcal{F}^s$ with parameters Θ^s , there
 325 exists some $f_k^s \in \mathcal{F}^s$ with parameters $\Theta_k \in \mathfrak{G}$ in the product grid such that

$$326 \quad \| \boldsymbol{\xi}_{\Theta_s}^{sl}(x^s) - \boldsymbol{\xi}_{\Theta_k}^l(x^s) \| < Q_l \delta$$

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

$$328 \quad (\text{SM1.26}) \quad (\mathfrak{d}_{\mathcal{X}}^s(f^s, f_k^s))^2 \leq \sup_{x^s \in \mathcal{X}^s} \sum_{l=1}^{L-1} 2L_K \|\boldsymbol{\xi}_{\Theta_s}^{sl}(x^s) - \boldsymbol{\xi}_{\Theta_k}^l(x^s)\| < 2L_K \delta \sum_{l=1}^{L-1} Q_l = 2L_K \delta Q.$$

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

$$331 \quad \delta = \frac{\epsilon^2}{2L_KQ}$$

332 which provides a grid consisting of

$$333 \quad \prod_{l=1}^{L-1} \kappa^l = \prod_{l=1}^{L-1} \left(\frac{4A_\Theta L_K Q}{\epsilon^2} + 1 \right)^{d_l(d_{l-1}+1)}$$

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

$$335 \quad \mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}_{\mathcal{X}}^s) \leq \prod_{l=1}^{L-1} \left(\frac{4A_\Theta L_K Q}{\epsilon^2} + 1 \right)^{d_l(d_{l-1}+1)}$$

336 for the covering number stated in the lemma. ■

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

$$339 \quad \begin{aligned} \mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s) &\leq \prod_{l=1}^L \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)} \\ \mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \epsilon, \mathfrak{d}^t) &\leq \prod_{l=1}^L \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)}. \end{aligned}$$

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

$$344 \quad (\text{SM1.27}) \quad \mathfrak{G}_{\mathcal{H} \circ \mathcal{F}} = \mathfrak{G}^1 \times \cdots \times \mathfrak{G}^L = \{\boldsymbol{\Theta}_k\}_{k=1}^{\kappa^1 \dots \kappa^L}$$

345 providing a cover for the parameter space

$$346 \quad \Phi_{\mathcal{H} \circ \mathcal{F}} = \{\boldsymbol{\Theta} = (\boldsymbol{\Theta}^1, \dots, \boldsymbol{\Theta}^L) : |\boldsymbol{\Theta}_{ij}^l| \leq A_\Theta, \forall i, j, l\}$$

347 consisting of

$$348 \quad (\text{SM1.28}) \quad \prod_{l=1}^L \kappa^l = \prod_{l=1}^L \left(\frac{2A_\Theta}{\delta} + 1 \right)^{d_l(d_{l-1}+1)}$$

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

$$351 \quad \|\boldsymbol{\Theta}^{sl} - \boldsymbol{\Theta}_k^l\| < \delta \sqrt{d_l(d_{l-1}+1)}$$

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

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

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

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

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

$$358 \quad \mathfrak{d}^s(g^s, g_k^s) = \sup_{x^s \in \mathcal{X}^s} \|g^s(x^s) - g_k^s(x^s)\| = \sup_{x^s \in \mathcal{X}^s} \|\xi_{\Theta^s}^{sL}(x^s) - \xi_{\Theta_k}^L(x^s)\| < Q_L \delta.$$

359 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
 360 a covering radius of ϵ , we set $\epsilon = Q_L \delta$, which results in a cover with

$$361 \quad (\text{SM1.30}) \quad \prod_{l=1}^L \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)}$$

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

$$363 \quad \mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s) \leq \prod_{l=1}^L \left(\frac{2A_\Theta Q_L}{\epsilon} + 1 \right)^{d_l(d_{l-1}+1)} \quad \blacksquare$$

364 stated in the lemma.

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

$$\begin{aligned}
 367 \quad \mathfrak{d}_{\mathcal{V}}^s(v_1^s, v_2^s) &\triangleq \sup_{x^s \in \mathcal{X}^s} |v_1^s(x^s) - v_2^s(x^s)| \\
 &\quad \mathfrak{d}_{\mathcal{V}}^t(v_1^t, v_2^t) \triangleq \sup_{x^t \in \mathcal{X}^t} |v_1^t(x^t) - v_2^t(x^t)|
 \end{aligned}$$

368 where $v_1^s, v_2^s \in \mathcal{V}^s$ and $v_1^t, v_2^t \in \mathcal{V}^t$. Then,

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

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

372 $\kappa^s = \mathcal{N}(\mathcal{V}^s, \epsilon, \mathfrak{d}_{\mathcal{V}}^s), \quad \kappa^t = \mathcal{N}(\mathcal{V}^t, \epsilon, \mathfrak{d}_{\mathcal{V}}^t)$

373 respectively, and the corresponding sets of ball centers as $\{v_k^s\}_{k=1}^{\kappa^s}$ and $\{v_l^t\}_{l=1}^{\kappa^t}$. Then, for any
 374 $v^s \in \mathcal{V}^s$ and any $v^t \in \mathcal{V}^t$ there exist some $v_k^s \in \mathcal{V}^s$ and $v_l^t \in \mathcal{V}^t$ such that

375 (SM1.31)
$$\begin{aligned} \mathfrak{d}_{\mathcal{V}}^s(v^s, v_k^s) &= \sup_{x^s \in \mathcal{X}^s} |v^s(x^s) - v_k^s(x^s)| < \epsilon \\ \mathfrak{d}_{\mathcal{V}}^t(v^t, v_l^t) &= \sup_{x^t \in \mathcal{X}^t} |v^t(x^t) - v_l^t(x^t)| < \epsilon. \end{aligned}$$

376 Let us denote

377
$$\begin{aligned} D(v_k^s, v_l^t) &\triangleq |E[v_k^s(x^s)] - E[v_l^t(x^t)]| \\ \hat{D}(v_k^s, v_l^t) &\triangleq \left| \frac{1}{N_s} \sum_{i=1}^{N_s} v_k^s(x_i^s) - \frac{1}{N_t} \sum_{j=1}^{N_t} v_l^t(x_j^t) \right|. \end{aligned}$$

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

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

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

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

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

385 (SM1.34)
$$|\hat{D}(v_k^s, v_l^t) - \hat{D}_{\Delta}(f^s, f^t)| < 2\epsilon.$$

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

387 (SM1.35)
$$\begin{aligned} &|D(v_k^s, v_l^t) - \hat{D}(v_k^s, v_l^t)| \\ &= \left| |E[v_k^s(x^s)] - E[v_l^t(x^t)]| - \left| \frac{1}{N_s} \sum_{i=1}^{N_s} v_k^s(x_i^s) - \frac{1}{N_t} \sum_{j=1}^{N_t} v_l^t(x_j^t) \right| \right| \\ &\leq \left| E[v_k^s(x^s)] - E[v_l^t(x^t)] - \frac{1}{N_s} \sum_{i=1}^{N_s} v_k^s(x_i^s) + \frac{1}{N_t} \sum_{j=1}^{N_t} v_l^t(x_j^t) \right| \\ &\leq \left| \frac{1}{N_s} \sum_{i=1}^{N_s} v_k^s(x_i^s) - E[v_k^s(x^s)] \right| + \left| \frac{1}{N_t} \sum_{j=1}^{N_t} v_l^t(x_j^t) - E[v_l^t(x^t)] \right|. \end{aligned}$$

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

$$390 \quad P\left(\left|\frac{1}{N_s} \sum_{i=1}^{N_s} v_k^s(x_i^s) - E[v_k^s(x^s)]\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{N_s \epsilon^2}{2C_D^2}\right)$$

391 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
 392 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

$$393 \quad 1 - 2\kappa^s \exp\left(-\frac{N_s \epsilon^2}{2C_D^2}\right) - 2\kappa^t \exp\left(-\frac{N_t \epsilon^2}{2C_D^2}\right)$$

394 we have

$$395 \quad \left|\frac{1}{N_s} \sum_{i=1}^{N_s} v_k^s(x_i^s) - E[v_k^s(x^s)]\right| < \epsilon \quad \text{and} \quad \left|\frac{1}{N_t} \sum_{j=1}^{N_t} v_l^t(x_j^t) - E[v_l^t(x^t)]\right| < \epsilon$$

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

$$397 \quad |D(v_k^s, v_l^t) - \hat{D}(v_k^s, v_l^t)| < 2\epsilon.$$

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

$$399 \quad \begin{aligned} & P\left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, \Delta \in \mathcal{D}} |D_\Delta(f^s, f^t) - \hat{D}_\Delta(f^s, f^t)| \leq 6\epsilon\right) \\ & \geq 1 - 2\kappa^s \exp\left(-\frac{N_s \epsilon^2}{2C_D^2}\right) - 2\kappa^t \exp\left(-\frac{N_t \epsilon^2}{2C_D^2}\right). \end{aligned}$$

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

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

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

$$415 \quad O\left(\sqrt{\frac{\dim_{VC}(\mathcal{H})}{M_s}} + \log(\delta^{-1})\right) + d_{\mathcal{H}}(D_S, D_T) + \lambda$$

416 ignoring the logarithmic factors, where $\dim_{VC}(\mathcal{H})$ denotes the VC-dimension of the hypothesis
 417 space \mathcal{H} , M_s is the number of labeled source samples, and λ is a measure of the proximity
 418 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]
 419 between the source and target distributions D_S and D_T , given by

$$420 \quad d_{\mathcal{H}}(D_S, D_T) = 2 \sup_{A \in \mathcal{A}} |P_{D_S}(A) - P_{D_T}(A)|$$

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

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

$$426 \quad O\left(\sqrt{\frac{\alpha^2}{\gamma} + \frac{(1-\alpha)^2}{1-\gamma}} \sqrt{\frac{\dim_{VC}(\mathcal{H}) + \log(\delta^{-1})}{M}}\right. \\ \left. + (1-\alpha)\left(\sqrt{\frac{\dim_{VC}(\mathcal{H}) \log(\delta^{-1})}{N}} + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \lambda\right)\right).$$

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

433 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
 434 weighted loss towards the target loss, by decreasing the weight $1 - \alpha$ of the source loss.

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

440 Following these pioneering works, many other domain divergence measures have been
 441 proposed in succeeding studies [SM32]. Deng et al. have explored a robust variant of the
 442 discrepancy in [SM26] based on the adversarial Rademacher complexity definition [SM11],
 443 which has been shown to vary with the number of samples M and the network width d at rate
 444 $O(\sqrt{d/M})$ for two-layer ReLU neural networks. Zhang et al. have proposed an alternative
 445 characterization of distribution distance based on the margin disparity discrepancy, leading
 446 to generalization bounds in terms of the Rademacher complexities and the covering numbers
 447 of hypothesis spaces [SM45]. Zellinger et al. have presented performance bounds depending
 448 on the VC-dimension of the function classes by formulating the domain discrepancy in terms
 449 of the difference between the moments of the source and target distributions [SM44]. Other
 450 recent efforts along this line include studies involving margin-aware risks with links to opti-
 451 mal transport distances [SM12], information-theoretic bounds based on mutual information
 452 [SM38, SM42], hypothesis-specific divergence measures [SM39], and risk definitions based on
 453 stochastic predictors [SM33].

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

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

476 method is proposed in [SM46], which learns the similarity between source and target samples
 477 through a linear transformation \mathbf{G} . Assuming the incoherence of the projections correspond-
 478 ing to different tasks, the estimation error of the transformation \mathbf{G} is shown to be bounded by
 479 $O(d_T \sqrt{\log(d_S)/n})$, where d_S and d_T denote the dimensions of the source and target Euclidean
 480 domains, and n is the number of tasks. While this bound is subsequently leveraged in [SM46]
 481 to design suitable classifiers based on the incoherence principle, the scope of their analysis is
 482 limited to linear transformations.

483 A performance analysis of conditional distribution matching is presented in [SM37], show-
 484 ing that the generalization gap in the target domain is bounded by

$$485 \quad O\left(1 + \frac{1}{\sqrt{M_t}} + \sqrt{\frac{\log(\delta^{-1})}{M_s + M_t}}\right)$$

486 when the source domain is mapped to the target domain through a location and scale trans-
 487 form.

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

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

$$499 \quad O\left(\text{dist}_{\mathcal{Y}}(f^s, f^t) + \sqrt{\frac{\log(\delta^{-1})}{M_s + M_t}}\right).$$

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

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

513 $O(\sqrt{\dim_{VC}(\mathcal{H} \circ \mathcal{F})/M_s} + \sqrt{\dim_{VC}(\mathcal{H})/M_t})$, where \mathcal{F} is the space of feature representations
 514 and \mathcal{H} is the space of classifiers. The significance of this result lies in the fact that the number
 515 M_t of labeled target samples should scale with the dimension of only the classifier \mathcal{H} , rather
 516 than the more complex composite hypothesis space $\mathcal{H} \circ \mathcal{F}$. A parallel finding is presented in
 517 [SM35] for the problem of transfer learning in a multi-task setting, demonstrating that the
 518 number of labeled samples for a new task needs to scale only with the complexity of its own
 519 task-specific map, assuming the abundance of the training data for the previous tasks.

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

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

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

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

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

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

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

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

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

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

569 **SM3.1.1. Synthetic data experiments.** We validate our theoretical findings on a synthetic
 570 data set with two classes. The source and target data sets are generated by applying
 571 two different geometric transformations to 400 samples drawn from the standard normal distribution
 572 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

$$580 \quad \min_{\mathbf{w} \in \mathbb{R}^2} \frac{1 - \alpha}{M_s} \sum_{i=1}^{M_s} (\mathbf{w}^T f^s(x_i^s) - \mathbf{y}_i^s)^2 + \frac{\alpha}{M_t} \sum_{j=1}^{M_t} (\mathbf{w}^T f^t(x_j^t) - \mathbf{y}_j^t)^2 + \lambda \|\mathbf{w}\|^2.$$

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

583 The variation of the target misclassification rate with the number M_t of labeled target
 584 samples is shown for different values of the weight α for the target loss. In order to interpret
 585 these results, it is helpful to recall our theoretical analysis: Theorem 2.1 states that
 586 the expected target loss $\mathcal{L}^t(f^t, h)$ deviates from its reference value based on the empirical
 587 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
 588 this with high and fixed probability, the term $M_t \epsilon^2$ in the probability expression must be
 589 constant (ignoring logarithmic factors and assuming that the generic covering numbers grow
 590 at a typical geometric rate). This implies that the expected target loss should decrease at
 591 rate $\epsilon = O(\sqrt{1/M_t})$ as M_t increases. We observe that the decay in the target error with
 592 M_t is consistent with Theorem 2.1. The fitted theoretical rates of decay $O(\sqrt{1/M_t})$ closely
 593 match the experimental data. We can also observe that large M_t values favor larger α values,
 594 while α must be chosen smaller at small M_t values. This aligns with the conclusion that the
 595 parameter α must be chosen as $\alpha = O(\sqrt{M_t})$ in order to control the probability term as M_t
 596 decreases.

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

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

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

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

629 In our experiments, the MNIST handwritten digit data set [SM24] is used as the source
 630 data set, which consists of 60000 images. The target data set is taken as MNIST-M [SM16],
 631 which contains 59000 handwritten digit images with colored backgrounds. We train the neural
 632 networks with labeled and unlabeled training samples from the source and target domains,
 633 and then evaluate the target accuracy of the learnt models, defined as the correct classification
 634 rate of test samples from the target domain. In all experiments, algorithm hyperparameters
 635 and fixed variables are chosen to keep the neural network in the overfitting regime, enabling
 636 the characterization of the sample complexity of the models under consideration.

637 **SM3.2.1. MMD-based domain adaptation networks.** In our analysis of MMD-based
 638 domain adaptation networks, we consider the architecture proposed in the pioneering study
 639 [SM25] as our benchmark. We build on our previous experimental study [SM23] and employ a
 640 neural network structure similar to the baseline model in [SM25], beginning with convolutional
 641 layers and followed by several fully connected MMD layers. The MMD layer parameters are
 642 coupled between the source and target domains. The dimensions (widths) of all MMD layers
 643 are set as equal. Batch normalization is applied after each layer in order to stabilize the
 644 performance. We use the PyTorch implementation of the network available in [SM8] and

645 adapt it for the minimization of the objective function

$$646 \quad \frac{1-\alpha}{M_s} \sum_{i=1}^{M_s} \ell(h \circ f(x_i^s), \mathbf{y}_i^s) + \frac{\alpha}{M_t} \sum_{i=1}^{M_t} \ell(h \circ f(x_j^t), \mathbf{y}_j^t) + \beta \sum_{l=1}^{L-1} (\hat{D}^l)^2(f^l, f^l)$$

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

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

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

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

678 **SM3.2.2. Adversarial domain adaptation networks.** In order to experimentally evaluate
679 our findings in Section 3.3.2, we adopt the model proposed in [SM17], which is a well-known
680 representative of adversarial domain adaptation architectures. We use the PyTorch imple-
681 mentation of this model available in [SM18], by adapting it to the semi-supervised setting

682 studied in our analysis. We train the adversarial network to minimize the objective function

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

684 where the label loss $\ell(\cdot, \cdot)$ and the domain discriminator loss $\ell_{\mathcal{D}}(\cdot, \cdot)$ are selected as the negative
685 log likelihood function, and the source and target feature extractor networks are coupled as
686 $f^s = f^t = f$.

687 The feature extractor network contains only convolutional layers, while the label predictor
688 and domain discriminator networks consist of fully connected layers in the implementation
689 in [SM18]. In order to adapt our experiments to this structure, when analyzing the sample
690 complexity of labeled data (M_s), we set the number of layers in the feature extractor and
691 label predictor networks as equal, which is represented by the parameter L . Likewise, when
692 studying the sample complexity of all data (N_s), the number of layers in the feature extractor
693 and domain discriminator networks are equated and denoted as L . We use a similar strategy
694 to adjust the network width, where we scale the number of convolutional channels and the
695 fully connected layer width in the original paper [SM17] with the same factor d . Hence, the
696 number of convolutional channels is scaled proportionally to the width of the label predictor
697 and the domain discriminator networks, respectively, when studying the sample complexities
698 of M_s and N_s . Batch normalization and ReLU layers are included after each convolutional or
699 fully connected layer, following standard practice.

700 The sample complexities of the number of source samples with the network depth L and
701 width d are presented in the figures. Similarly to the MMD experiments, we show the variation
702 of the target accuracy with L or d at different M_s and N_s values, and then investigate the
703 smallest M_s and N_s values ensuring a reference target accuracy level as L or d increases. The
704 results of these experiments align with the theoretical bounds in Theorem 3.7, confirming the
705 quadratic growth in the sample complexities $M_s, N_s = O(L^2)$ and $M_s, N_s = O(d^2)$ as the
706 network depth L and width d increase.

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

711 Overall, our experimental findings are in line with the theoretical bounds presented in
712 Theorems 3.6 and 3.7, supporting our sample complexity and optimal weight choice analyses
713 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 $\eta : \mathbb{R}^{d_l} \rightarrow \mathbb{R}^{d_l}$ represent an activation function in layer l giving the output $\zeta = \eta(\xi)$ for the input $\xi \in \mathbb{R}^{d_l}$.

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

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

where $\zeta = \eta_R(\xi)$, and the notation $(\cdot)(k)$ denotes the k -th entry of a vector. For two vectors $\xi_1, \xi_2 \in \mathbb{R}^{d_l}$, we have

$$\begin{aligned} \| \eta_R(\xi_1) - \eta_R(\xi_2) \|^2 &= \sum_{k=1}^{d_l} (\max\{0, \xi_1(k)\} - \max\{0, \xi_2(k)\})^2 \\ (SM4.2) \quad &\leq \sum_{k=1}^{d_l} (\xi_1(k) - \xi_2(k))^2 = \| \xi_1 - \xi_2 \|^2 \end{aligned}$$

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

$$\| \eta_R(\xi_1) - \eta_R(\xi_2) \| \leq \| \xi_1 - \xi_2 \|$$

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

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

given by

$$(SM4.3) \quad \zeta(k) = \log(1 + e^{\xi(k)})$$

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

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

for all $t \in \mathbb{R}$. Then for $\zeta_1 = \eta_{SP}(\xi_1)$ and $\zeta_2 = \eta_{SP}(\xi_2)$ with $\xi_1, \xi_2 \in \mathbb{R}^{d_l}$, from the mean value theorem we get

$$(SM4.5) \quad |\zeta_1(k) - \zeta_2(k)| \leq |\xi_1(k) - \xi_2(k)|$$

which implies

$$(SM4.6) \quad \| \eta_{SP}(\xi_1) - \eta_{SP}(\xi_2) \| \leq \| \xi_1 - \xi_2 \|.$$

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

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

$$741 \quad \eta_{SM}(\boldsymbol{\xi}) = [\eta_{SM}^1(\boldsymbol{\xi}) \ \eta_{SM}^2(\boldsymbol{\xi}) \ \cdots \ \eta_{SM}^{d_l}(\boldsymbol{\xi})]^T$$

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

$$744 \quad (\text{SM4.7}) \quad \eta_{SM}^k(\boldsymbol{\xi}) = \frac{e^{\boldsymbol{\xi}(k)}}{\sum_{n=1}^{d_l} e^{\boldsymbol{\xi}(n)}}.$$

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

$$748 \quad \eta_{SM}^k(\boldsymbol{\xi}_1) - \eta_{SM}^k(\boldsymbol{\xi}_2) = (\nabla \eta_{SM}^k(\boldsymbol{\xi}))^T (\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2)$$

749 where $\nabla \eta_{SM}^k(\boldsymbol{\xi}) \in \mathbb{R}^{d_l}$ denotes the gradient of η_{SM}^k at $\boldsymbol{\xi}$. The following inequality is then
 750 obtained

$$751 \quad (\text{SM4.8}) \quad |\eta_{SM}^k(\boldsymbol{\xi}_1) - \eta_{SM}^k(\boldsymbol{\xi}_2)| \leq \sup_{\boldsymbol{\xi} \in \mathbb{R}^{d_l}} \|\nabla \eta_{SM}^k(\boldsymbol{\xi})\| \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|.$$

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

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

$$756 \quad \frac{\partial \eta_{SM}^k(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(n)} = \frac{\partial}{\partial \boldsymbol{\xi}(n)} \left(\frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right) = - \frac{e^{\boldsymbol{\xi}(k)} e^{\boldsymbol{\xi}(n)}}{\left(\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)} \right)^2}.$$

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

$$759 \quad (\text{SM4.9}) \quad \left| \frac{\partial \eta_{SM}^k(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(n)} \right| \leq \frac{1}{4}.$$

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

$$761 \quad \frac{\partial \eta_{SM}^k(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(k)} = \frac{\partial}{\partial \boldsymbol{\xi}(k)} \left(\frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right) = \left(\frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right) \left(1 - \frac{e^{\boldsymbol{\xi}(k)}}{\sum_{r=1}^{d_l} e^{\boldsymbol{\xi}(r)}} \right).$$

762 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
 763 of the function $\alpha(1 - \alpha)$ in the interval $\alpha \in [0, 1]$ is $1/4$, we get

$$764 \quad (\text{SM4.10}) \quad \left| \frac{\partial \eta_{SM}^k(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}(k)} \right| \leq \frac{1}{4}.$$

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

766
$$\|\nabla \eta_{SM}^k(\xi)\| \leq \frac{\sqrt{d_l}}{4}$$

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

768
$$|\eta_{SM}^k(\xi_1) - \eta_{SM}^k(\xi_2)| \leq \frac{\sqrt{d_l}}{4} \|\xi_1 - \xi_2\|$$

769 for any $\xi_1, \xi_2 \in \mathbb{R}^{d_l}$, which implies

770
$$\|\eta_{SM}(\xi_1) - \eta_{SM}(\xi_2)\| \leq \frac{d_l}{4} \|\xi_1 - \xi_2\|.$$

771 Defining

772
$$d_{\max} = \max_{l=1,\dots,L} d_l$$

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

774 **SM5. Proof of Corollary 3.5.**

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

$$778 \quad R_l = C_\eta \sqrt{d_l} = O(d^{1/2}).$$

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

$$780 \quad R_l = O((cd)^l).$$

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

$$782 \quad Q_l = O(c^{l-1} d^{l+\frac{1}{2}})$$

783 which results in

$$784 \quad (\text{SM5.1}) \quad Q = O(c^{L-2} d^{L-\frac{1}{2}}).$$

785 Meanwhile, condition (3.15) yields

$$786 \quad Q_l = O((l-1) c^{l-1} d^l)$$

787 resulting in

$$788 \quad (\text{SM5.2}) \quad Q = O((L-2) c^{L-2} d^{L-1}).$$

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

$$791 \quad Q = O(L c^{L-2} d^L)$$

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

$$794 \quad \mathcal{N}(\mathcal{F}^s, \epsilon, \mathfrak{d}_{\mathcal{X}}^s) = O\left(\left(\frac{cQ}{\epsilon^2}\right)^{d^2 L}\right) = O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2}\right)$$

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

$$797 \quad \mathcal{N}(\mathcal{F}^t, \epsilon, \mathfrak{d}_{\mathcal{X}}^t) = O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2}\right).$$

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

799 For condition (3.14), we have

$$800 \quad Q_L = O(c^{L-1} d^{L+\frac{1}{2}})$$

801 which gives from Lemma SM1.8

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

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

805
$$Q_L = O((L-1)c^{L-1}d^L)$$

806 resulting in

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

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

809
$$\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \epsilon, \mathfrak{d}^s) = O\left(\left(\frac{L}{\epsilon}\right)^{d^2L}(cd)^{d^2L^2}\right)$$

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

811
$$\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \epsilon, \mathfrak{d}^t) = O\left(\left(\frac{L}{\epsilon}\right)^{d^2L}(cd)^{d^2L^2}\right). \blacksquare$$

812 SM6. Proof of Theorem 3.6.

813 *Proof.* We first notice that, owing to Lemma SM1.5, we can analyze MMD-based domain
814 adaptation networks within the setting of Theorem 2.7. The compactness of the function
815 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.
816 Assumptions 2.3 and 2.6 are thereby satisfied; hence, the statement of Theorem 2.7 applies
817 to the current setting in consideration.

818 We recall from Theorem 2.7 that the expected target loss in (3.16) is attained with prob-
819 ability at least

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

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

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

827
$$a_s(N_s, \epsilon) = \theta(N_s \epsilon^2)$$

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

830
$$\begin{aligned} \mathcal{N}(\mathcal{F}^s, \frac{\epsilon}{8}, \mathfrak{d}_{\mathcal{X}}^s) \exp(-a_s(N_s, \epsilon)) &= O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2} \exp(-N_s \epsilon^2)\right) \\ &= O\left(\exp\left(d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(cd) - N_s \epsilon^2\right)\right). \end{aligned}$$

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

833
$$N_s = O\left(\frac{d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}{\epsilon^2}\right),$$

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

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

838
$$\begin{aligned} \mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} &= O\left(\left(\frac{L\alpha}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2} \exp\left(-\frac{M_t \epsilon^2}{\alpha^2}\right)\right) \\ &= O\left(\exp\left(d^2 L \log\left(\frac{L\alpha}{\epsilon}\right) + d^2 L^2 \log(cd) - \frac{M_t \epsilon^2}{\alpha^2}\right)\right). \end{aligned}$$

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

841
$$\alpha = O\left(\left(\frac{M_t \epsilon^2}{d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}\right)^{1/2}\right)$$

842 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.

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

844
$$\begin{aligned} \mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}} &= O\left(\left(\frac{L(1-\alpha)}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2} \exp\left(-\frac{M_s \epsilon^2}{(1-\alpha)^2}\right)\right) \\ &= O\left(\exp\left(d^2 L \log\left(\frac{L(1-\alpha)}{\epsilon}\right) + d^2 L^2 \log(cd) - \frac{M_s \epsilon^2}{(1-\alpha)^2}\right)\right). \end{aligned}$$

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

846

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

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

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

850 We first discuss the magnitude bound A_ℓ for the widely used cross-entropy loss function. Let

851 $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y} \subset \mathbb{R}^m$ be two nonnegative label vectors in the label set $\mathcal{Y} = [0, 1] \times \dots \times [0, 1] \subset \mathbb{R}^m$.

852 In its naïve form, the cross-entropy loss between \mathbf{y}_1 and \mathbf{y}_2 is given by

853 (SM7.1)

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

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

857

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

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

859

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

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

861

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

862 We next derive the Lipschitz constant L_ℓ of the cross-entropy loss function. For any
863 $\mathbf{y}, \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ we have

864 (SM7.2)

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

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

866

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

867 which gives

868

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

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

$$870 \quad |\ell(\mathbf{y}_1, \mathbf{y}) - \ell(\mathbf{y}_2, \mathbf{y})| \leq \sum_{k=1}^m \delta^{-1} |\mathbf{y}_2(k) - \mathbf{y}_1(k)| \leq \delta^{-1} \sqrt{m} \|\mathbf{y}_2 - \mathbf{y}_1\|$$

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

$$873 \quad L_\ell \triangleq \delta^{-1} \sqrt{m}.$$

874 SM8. Proof of Theorem 3.11.

875 *Proof.* We begin by bounding the expected target loss as

$$876 \quad \mathcal{L}^t(f^t, h) \leq \mathcal{L}^s(f^s, h) + R_A D_\Delta(f^s, f^t)$$

877 using Assumption 3.10. It follows that

$$878 \quad (\text{SM8.1}) \quad \begin{aligned} \mathcal{L}^t(f^t, h) &= \alpha \mathcal{L}^t(f^t, h) + (1 - \alpha) \mathcal{L}^t(f^t, h) \\ &\leq \alpha \mathcal{L}^t(f^t, h) + (1 - \alpha) (\mathcal{L}^s(f^s, h) + R_A D_\Delta(f^s, f^t)) \\ &= \mathcal{L}_\alpha(f^s, f^t, h) + (1 - \alpha) R_A D_\Delta(f^s, f^t). \end{aligned}$$

879 We next aim to upper bound the expected loss $\mathcal{L}_\alpha(f^s, f^t, h)$ and the expected distribution
880 distance $D_\Delta(f^s, f^t)$ in terms of their empirical counterparts. It follows from Assumptions 3.1
881 and 3.8 that the source hypothesis space $\mathcal{G}^s = \mathcal{H} \circ \mathcal{F}^s$, the target hypothesis space $\mathcal{G}^t = \mathcal{H} \circ \mathcal{F}^t$,
882 the source domain discriminator space $\mathcal{V}^s = \mathcal{D} \circ \mathcal{F}^s$ and the target domain discriminator space
883 $\mathcal{V}^t = \mathcal{D} \circ \mathcal{F}^t$ are compact with respect to the metrics $\mathfrak{d}^s, \mathfrak{d}^t, \mathfrak{d}_{\mathcal{V}}^s, \mathfrak{d}_{\mathcal{V}}^t$, respectively, which can be
884 shown by following similar steps as in the proof of Lemma SM1.6 in Appendix 6.

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

$$887 \quad (\text{SM8.2}) \quad \begin{aligned} P \left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, h \in \mathcal{H}} |\mathcal{L}_\alpha(f^s, f^t, h) - \hat{\mathcal{L}}_\alpha(f^s, f^t, h)| \leq \epsilon \right) \\ \geq 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}} \end{aligned}$$

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

$$890 \quad (\text{SM8.3}) \quad \begin{aligned} P \left(\sup_{f^s \in \mathcal{F}^s, f^t \in \mathcal{F}^t, \Delta \in \mathcal{D}} |D_\Delta(f^s, f^t) - \hat{D}_\Delta(f^s, f^t)| \leq \epsilon \right) \\ \geq 1 - 2\mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}_{\mathcal{V}}^s) \exp \left(-\frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2} \right) - 2\mathcal{N}(\mathcal{V}^t, \frac{\epsilon}{6}, \mathfrak{d}_{\mathcal{V}}^t) \exp \left(-\frac{N_t \epsilon^2}{72 C_{\mathcal{D}}^2} \right) \end{aligned}$$

891 due to Lemma SM1.9.

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

$$894 \quad \begin{aligned} & 1 - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}} - 2\mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}} \\ & - 2\mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}_\mathcal{V}^s) \exp\left(-\frac{N_s \epsilon^2}{72C_\mathcal{D}^2}\right) - 2\mathcal{N}(\mathcal{V}^t, \frac{\epsilon}{6}, \mathfrak{d}_\mathcal{V}^t) \exp\left(-\frac{N_t \epsilon^2}{72C_\mathcal{D}^2}\right) \end{aligned}$$

895 the expected target loss is bounded as

$$896 \quad \mathcal{L}^t(f^t, h) \leq \hat{\mathcal{L}}_\alpha(f^s, f^t, h) + (1-\alpha)R_A \hat{D}_\Delta(f^s, f^t) + (1-\alpha)R_A \epsilon + \epsilon.$$

897 In the sequel, we examine each one of the terms in the probability expression in (SM8.4).
 898 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
 899 the result in Lemma 3.4 applies to this setting as well, which implies that the rate of growth
 900 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

$$901 \quad O\left(\left(\frac{L}{\epsilon}\right)^{d^2 L} (cd)^{d^2 L^2}\right)$$

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

$$904 \quad \alpha = O\left(\left(\frac{M_t \epsilon^2}{d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}\right)^{1/2}\right),$$

905 together with scaling M_s at rate

$$906 \quad M_s = O\left(\frac{d^2 L \log\left(\frac{L}{\epsilon}\right) + d^2 L^2 \log(d)}{\epsilon^2}\right)$$

907 ensures an upper bound on the terms

$$908 \quad \mathcal{N}(\mathcal{H} \circ \mathcal{F}^s, \frac{\epsilon}{8(1-\alpha)L_\ell}, \mathfrak{d}^s) e^{-\frac{M_s \epsilon^2}{8(1-\alpha)^2 A_\ell^2}}$$

909 and

$$910 \quad \mathcal{N}(\mathcal{H} \circ \mathcal{F}^t, \frac{\epsilon}{8\alpha L_\ell}, \mathfrak{d}^t) e^{-\frac{M_t \epsilon^2}{8\alpha^2 A_\ell^2}}$$

911 in the probability expression in (SM8.4).

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

915 that the function space $\mathcal{V}^s = \mathcal{D} \circ \mathcal{F}^s$ has an identical construction to the function space
 916 $\mathcal{G}^s = \mathcal{H} \circ \mathcal{F}^s$, if the metric

$$917 \quad \mathfrak{d}^s(g_1^s, g_2^s) = \sup_{x^s \in \mathcal{X}^s} \|g_1^s(x^s) - g_2^s(x^s)\|$$

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

$$919 \quad \mathfrak{d}_{\mathcal{V}}^s(v_1^s, v_2^s) = \sup_{x^s \in \mathcal{X}^s} |v_1^s(x^s) - v_2^s(x^s)|$$

920 which uses the Euclidean distance in \mathbb{R} instead. Hence, the latter is a special case of the
 921 former that can be obtained by setting $m = 1$. Consequently, the analysis of the covering
 922 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
 923 by replacing the number of layers L with the total number of layers $L + K - 1$ in the cascade
 924 network formed by the combination of the feature extractor and the domain discriminator
 925 networks. We thus get

$$926 \quad \mathcal{N}(\mathcal{V}^s, \epsilon, \mathfrak{d}_{\mathcal{V}}^s) = O \left(\left(\frac{L + K}{\epsilon} \right)^{d^2(L+K)} (cd)^{d^2(L+K)^2} \right)$$

927 which yields

$$928 \quad (\text{SM8.5}) \quad \begin{aligned} & \mathcal{N}(\mathcal{V}^s, \frac{\epsilon}{6}, \mathfrak{d}_{\mathcal{V}}^s) \exp \left(-\frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2} \right) \\ &= O \left(\left(\frac{L + K}{\epsilon} \right)^{d^2(L+K)} (cd)^{d^2(L+K)^2} \exp \left(-\frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2} \right) \right) \\ &= O \left(\exp \left(d^2(L + K) \log \left(\frac{L + K}{\epsilon} \right) + d^2(L + K)^2 \log(cd) - \frac{N_s \epsilon^2}{72 C_{\mathcal{D}}^2} \right) \right). \end{aligned}$$

929 We thus conclude that the sample complexity

$$930 \quad N_s = O \left(\frac{d^2(L + K) \log \left(\frac{L + K}{\epsilon} \right) + d^2(L + K)^2 \log(d)}{\epsilon^2} \right)$$

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

$$933 \quad N_t = O \left(\frac{d^2(L + K) \log \left(\frac{L + K}{\epsilon} \right) + d^2(L + K)^2 \log(d)}{\epsilon^2} \right)$$

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

- [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. RAKHIN, *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. DHOUB, 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. RAKHIN, 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:

- 988 <http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html>.
- 989 [28] D. McNAMARA AND M. BALCAN, *Risk bounds for transferring representations with and without fine-*
990 *tuning*, in Proc. Int. Conf. Machine Learning,, vol. 70, 2017, pp. 2373–2381.
- 991 [29] M. MOHRI AND A. M. MEDINA, *New analysis and algorithm for learning with drifting distributions*, in
992 Int. Conf. Algorithmic Learning Theory, vol. 7568, 2012, pp. 124–138.
- 993 [30] B. NEYSHABUR, S. BHOJANAPALLI, AND N. SREBRO, *A PAC-bayesian approach to spectrally-normalized*
994 *margin bounds for neural networks*, in Int. Conf. Learning Representations, 2018.
- 995 [31] B. NEYSHABUR, R. TOMIOKA, AND N. SREBRO, *Norm-based capacity control in neural networks*, in Prof.
996 28th Conference on Learning Theory, vol. 40, 2015, pp. 1376–1401.
- 997 [32] I. REDKO, E. MORVANT, A. HABRARD, M. SEBBAN, AND Y. BENNANI, *A survey on domain adaptation*
998 *theory*, arXiv preprint: <http://arxiv.org/abs/2004.11829>, (2020).
- 999 [33] A. SICILIA, K. ATWELL, M. ALIKHANI, AND S. J. HWANG, *PAC-Bayesian domain adaptation bounds for*
1000 *multiclass learners*, in Proc. Conf. Uncertainty in Artificial Intelligence, vol. 180, 2022, pp. 1824–1834.
- 1001 [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.
- 1002 [35] N. TRIPURANENI, M. I. JORDAN, AND C. JIN, *On the theory of transfer learning: The importance of task*
1003 *diversity*, in Advances in Neural Information Processing Systems, 2020.
- 1004 [36] G. VARDI, O. SHAMIR, AND N. SREBRO, *The sample complexity of one-hidden-layer neural networks*, in
1005 Advances in Neural Information Processing Systems 35, 2022.
- 1006 [37] X. WANG AND J. SCHNEIDER, *Generalization bounds for transfer learning under model shift*, in Proc.
1007 Conf. Uncertainty in Artificial Intelligence, 2015, pp. 922–931.
- 1008 [38] Z. WANG AND Y. MAO, *Information-theoretic analysis of unsupervised domain adaptation*, in Int. Conf.
1009 Learning Representations, 2023.
- 1010 [39] Z. WANG AND Y. MAO, *On f -divergence principled domain adaptation: An improved framework*, in
1011 Advances in Neural Information Processing Systems, 2024.
- 1012 [40] B. WANG ET AL., *Gap minimization for knowledge sharing and transfer*, J. Mach. Learn. Res., 24 (2023),
1013 pp. 33:1–33:57.
- 1014 [41] C. WEI AND T. MA, *Data-dependent sample complexity of deep neural networks via Lipschitz augmentation*,
1015 in Advances in Neural Information Processing Systems 32, 2019, pp. 9722–9733.
- 1016 [42] X. WU, J. H. MANTON, U. AICKELIN, AND J. ZHU, *On the generalization for transfer learning: An*
1017 *information-theoretic analysis*, IEEE Trans. Inf. Theory, 70 (2024), pp. 7089–7124.
- 1018 [43] V. V. YURINSKII, *Exponential inequalities for sums of random vectors*, Journal of Multivariate Analysis,
1019 6 (1976), pp. 473–499.
- 1020 [44] W. ZELLINGER, B. A. MOSER, AND S. SAMINGER-PLATZ, *On generalization in moment-based domain*
1021 *adaptation*, Ann. Math. Artif. Intell., 89 (2021), pp. 333–369.
- 1022 [45] Y. ZHANG, T. LIU, M. LONG, AND M. I. JORDAN, *Bridging theory and algorithm for domain adaptation*,
1023 in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 7404–
1024 7413.
- 1025 [46] J. T. ZHOU, I. W. TSANG, S. J. PAN, AND M. TAN, *Multi-class heterogeneous domain adaptation*,
1026 Journal of Machine Learning Research, 20 (2019), pp. 1–31.
- 1027