A Proof of Theorem 1

The proof of Theorem 1 makes use of the following simple inequality which is straight-forward to prove [6]. For any hypotheses $h, h' \in \mathcal{H}$,

$$|\epsilon_S(h, h') - \epsilon_T(h, h')| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \ .$$

The proof also relies heavily on the triangle inequality for classification error [3, 8] which implies that for any labeling functions f_1, f_2, and f_3, $\epsilon_S(f_1, f_2) \leq \epsilon_S(f_1, f_3) + \epsilon_S(f_2, f_3)$. Similarly, for the target domain, for any f_1, f_2, and f_3, $\epsilon_T(f_1, f_2) \leq \epsilon_T(f_1, f_3) + \epsilon_T(f_2, f_3)$.

$$\begin{aligned}
\epsilon_T(h) &\leq \epsilon_T(h^*) + \epsilon_T(h, h^*) \leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + |\epsilon_T(h, h^*) - \epsilon_S(h, h^*)| \\
&\leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\
&\leq \epsilon_T(h^*) + \epsilon_S(h) + \epsilon_S(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\
&= \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda \\
&\leq \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d \log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda
\end{aligned}$$

The last step in the proof is an application of Theorem 3.4 of Ben-David, Gehrke, and Kifer [4], together with the observation that since we can represent every $g \in \mathcal{H}\Delta\mathcal{H}$ as a linear threshold network of depth 2 with 2 hidden units, the VC dimension of $\mathcal{H}\Delta\mathcal{H}$ is at most twice the VC dimension of $\mathcal{H}$ [1]. ■

B Proof of the main theorem

B.1 Proof of Lemma 1

This proof again relies heavily on the triangle inequality for classification error.

$$\begin{aligned}
&|\epsilon_\alpha(h) - \epsilon_T(h)| = (1-\alpha)|\epsilon_S(h) - \epsilon_T(h)| \\
&\quad \leq (1-\alpha)\left[|\epsilon_S(h) - \epsilon_S(h, h^*)| + |\epsilon_S(h, h^*) - \epsilon_T(h, h^*)| + |\epsilon_T(h, h^*) - \epsilon_T(h)|\right] \\
&\quad \leq (1-\alpha)\left[\epsilon_S(h^*) + |\epsilon_S(h, h^*) - \epsilon_T(h, h^*)| + \epsilon_T(h^*)\right] \\
&\quad \leq (1-\alpha)(\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda)
\end{aligned}$$

■

B.2 Proof of Lemma 2

We begin by restating Hoeffding's inequality.

Hoeffding's inequality
If $X_1, X_2, \ldots, X_n$ are independent and $a_i \leq X_i \leq b_i (i = 1, 2, \ldots, n)$, then for $\epsilon > 0$

$$\Pr\left[|\bar{X} - E[\bar{X}]| \geq \epsilon\right] \leq 2e^{-2n^2\epsilon^2 / \sum_{i=1}^{n} (b_i - a_i)^2},$$

where $\bar{X} = (X_1 + \cdots + X_n)/n$.

Let $X_1, \ldots, X_{\beta m}$ be random variables that take on the values $(\alpha/\beta)|h(x) - f_T(x)|$ for the βm instances $x \in S_T$. Similarly, let $X_{\beta m+1}, \ldots, X_m$ be random variables that take on the values

[6] Ben-David et al. [3] incorrectly stated this inequality in the original proof of Theorem 1. They wrote it using $d_{\mathcal{H}}$ instead of $\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}$.

$(1-\alpha)/(1-\beta)|h(x)-f_S(x)|$ for the $(1-\beta)m$ instances $x \in S_S$. Note that $X_1, \ldots, X_{\beta m} \in [0, \alpha/\beta]$ and $X_{\beta m+1}, \ldots, X_m \in [0, (1-\alpha)/(1-\beta)]$. Then

$$\begin{aligned}\hat{\epsilon}_\alpha(h) &= \alpha\hat{\epsilon}_T(h) + (1-\alpha)\hat{\epsilon}_S(h) \\ &= \alpha\frac{1}{\beta m}\sum_{x\in S_T}|h(x)-f_T(x)| + (1-\alpha)\frac{1}{(1-\beta)m}\sum_{x\in S_S}|h(x)-f_S(x)| = \frac{1}{m}\sum_{i=1}^{m}X_i.\end{aligned}$$

Furthermore, by linearity of expectations

$$\begin{aligned}E[\hat{\epsilon}_\alpha(h)] &= \frac{1}{m}\left(\beta m\frac{\alpha}{\beta}\epsilon_T(h) + (1-\beta)m\frac{1-\alpha}{1-\beta}\epsilon_S(h))\right) \\ &= \alpha\epsilon_T(h) + (1-\alpha)\epsilon_S(h) = \epsilon_\alpha(h).\end{aligned}$$

So by Hoeffding's inequality the following holds for every h.

$$\begin{aligned}\Pr\left[|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)| \geq \epsilon\right] &\leq 2\exp\left(\frac{-2m^2\epsilon^2}{\sum_{i=1}^{m}\text{range}^2(X_i)}\right) \\ &= 2\exp\left(\frac{-2m^2\epsilon^2}{\beta m\left(\frac{\alpha}{\beta}\right)^2 + (1-\beta)m\left(\frac{1-\alpha}{1-\beta}\right)^2}\right) \\ &= 2\exp\left(\frac{-2m\epsilon^2}{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\right).\end{aligned}$$

The remainder of the proof for hypothesis classes of finite VC dimension follows a standard argument. In particular, the reduction to a finite hypothesis class using the growth function does not change [16, 1]. This, combined with the union bound, gives us the probability that there exists *any* hypothesis $h \in \mathcal{H}$, $|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)| \geq \epsilon$. Substituting δ for the probability and solving gives us

$$\epsilon = \sqrt{\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)\frac{d\log(2m) - \log\delta}{2m}}.$$

■

B.3 Proof of Theorem 2

The proof follows the standard set of steps for proving learning bounds [1], using Lemma 1 to bound the difference between target and weighted errors and Lemma 2 for the uniform convergence of empirical and true weighted errors. Below we use L1, L2, and Thm1 to indicate that a line of the

proof follows by application of Lemma 1, Lemma 2, or Theorem 1 respectively.

$$\epsilon_T(\hat{h}) \leq \epsilon_\alpha(\hat{h}) + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \text{ (L1)}$$

$$\leq \hat{\epsilon}_\alpha(\hat{h}) + \sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{d\log(2m)-\log\delta}{2m}} + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \text{ (L2)}$$

$$\leq \hat{\epsilon}_\alpha(h_T^*) + \sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{d\log(2m)-\log\delta}{2m}} + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right)$$

$$\leq \epsilon_\alpha(h_T^*) + 2\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{d\log(2m)-\log\delta}{2m}} + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \text{ (L2)}$$

$$\leq \epsilon_T(h_T^*) + 2\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{d\log(2m)-\log\delta}{2m}} + 2(1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \text{ (L1)}$$

$$\leq \epsilon_T(h_T^*) + 2\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{d\log(2m)-\log\delta}{2m}} \quad +$$

$$2(1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S,\mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m')+\log(\frac{4}{\delta})}{m'}}+\lambda\right) \text{ (Thm 1)}$$

■

C Proof of Theorem 3

Lemma 3 *Let h be a hypothesis in class $\mathcal{H}$. Then $|\epsilon_{\boldsymbol{\alpha}}(h) - \epsilon_T(h)| \leq d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\boldsymbol{\alpha}},\mathcal{D}_T) + \gamma_{\boldsymbol{\alpha}}$.*

Proof:

$$\begin{aligned}|\epsilon_{\boldsymbol{\alpha}}(h) - \epsilon_T(h)| &\leq [|\epsilon_{\boldsymbol{\alpha}}(h) - \epsilon_{\boldsymbol{\alpha}}(h,h^*)| + |\epsilon_{\boldsymbol{\alpha}}(h,h^*) - \epsilon_T(h,h^*)| + |\epsilon_T(h,h^*) - \epsilon_T(h)|] \\ &\leq [\epsilon_{\boldsymbol{\alpha}}(h^*) + |\epsilon_{\boldsymbol{\alpha}}(h,h^*) - \epsilon_T(h,h^*)| + \epsilon_T(h^*)] \\ &\leq (\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\boldsymbol{\alpha}},\mathcal{D}_T) + \gamma_{\boldsymbol{\alpha}})\end{aligned}$$

■

Lemma 4 *Let $\mathcal{H}$ be a hypothesis space of VC-dimension d. If a random labeled sample of size m is generated by drawing $\beta_j m$ points from $\mathcal{D}_j$, and labeling them according to f_j, then with probability at least $1-\delta$ (over the choice of the samples), for every $h \in \mathcal{H}$:*

$$|\hat{\epsilon}_{\boldsymbol{\alpha}}(h) - \epsilon_{\boldsymbol{\alpha}}(h)| < \sqrt{\sum_j \frac{\alpha_j^2}{\beta_j}}\sqrt{\frac{d\log(2m)-\log\delta}{2m}}$$

Proof: Because of its similarity to the proof of Lemma 2 (in Appendix B.2), we will omit some details of this proof. Let $X_1,\ldots,X_{\beta_j m}$ be random variables that take on the values $(\alpha_j/\beta_j)|h(x) - f_j(x)|$ for the $\beta_j m$ instances $x \in S_j$. Note that $X_1,\ldots,X_{\beta_j m} \in [0, \alpha_j/\beta_j]$. Then

$$\hat{\epsilon}_{\boldsymbol{\alpha}}(h) = \sum_{j=1}^N \alpha_j \hat{\epsilon}_j(h) = \sum_{j=1}^N \alpha_j \frac{1}{\beta_j m}\sum_{x\in S_j}|h(x) - f_j(x)| = \frac{1}{m}\sum_{i=1}^m X_i.$$

By linearity of expectations again, we have $E[\hat{\epsilon}_{\boldsymbol{\alpha}}(h)] = \epsilon_{\boldsymbol{\alpha}}(h)$.

By Hoeffding's inequality the following holds for every h.

$$\begin{aligned}\Pr\left[|\hat{\epsilon}_{\boldsymbol{\alpha}}(h) - \epsilon_{\boldsymbol{\alpha}}(h)| \geq \epsilon\right] &\leq 2\exp\left(\frac{-2m^2\epsilon^2}{\sum_{i=1}^{m}\operatorname{range}^2(X_i)}\right) \\ &= 2\exp\left(\frac{-2m\epsilon^2}{\sum_j \frac{\alpha_j^2}{\beta_j}}\right).\end{aligned}$$

The remainder of the proof is identical to the proof of Lemma 2. ■

The proof of Theorem 3 combines Lemmas 3 and 4, following an identical argument to the proof of Theorem 2.