A. Missing Proofs

Lemma 4.1. Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and $\mathcal{D}, \mathcal{D}'$ be two distributions over $\mathcal{X}$. Then $\forall h, h' \in \mathcal{H}$, $|\varepsilon_{\mathcal{D}}(h,h') - \varepsilon_{\mathcal{D}'}(h,h')| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D},\mathcal{D}')$, where $\tilde{\mathcal{H}} := \{\text{sgn}(|h(\mathbf{x}) - h'(\mathbf{x})| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$.

Proof. By definition, for $\forall h, h' \in \mathcal{H}$, we have:

$$\begin{aligned}|\varepsilon_S(h,h') - \varepsilon_T(h,h')| &\leq \sup_{h,h' \in \mathcal{H}} |\varepsilon_S(h,h') - \varepsilon_T(h,h')| \\ &= \sup_{h,h' \in \mathcal{H}} \left|\mathbb{E}_{\mathbf{x}\sim S}[|h(\mathbf{x}) - h'(\mathbf{x})|] - \mathbb{E}_{\mathbf{x}\sim T}[|h(\mathbf{x}) - h'(\mathbf{x})|]\right| \end{aligned} \tag{6}$$

Since $||h||_\infty \leq 1, \forall h \in \mathcal{H}$, then $0 \leq |h(\mathbf{x}) - h'(\mathbf{x})| \leq 1$, $\forall \mathbf{x} \in \mathcal{X}, h, h' \in \mathcal{H}$. We now use Fubini's theorem to bound $\left|\mathbb{E}_{\mathbf{x}\sim S}[|h(\mathbf{x}) - h'(\mathbf{x})|] - \mathbb{E}_{\mathbf{x}\sim T}[|h(\mathbf{x}) - h'(\mathbf{x})|]\right|$:

$$\begin{aligned}&\left|\mathbb{E}_{\mathbf{x}\sim S}[|h(\mathbf{x}) - h'(\mathbf{x})|] - \mathbb{E}_{\mathbf{x}\sim T}[|h(\mathbf{x}) - h'(\mathbf{x})|]\right| \\ &\qquad = \left| \int_0^1 \left(\Pr_S(|h(\mathbf{x}) - h'(\mathbf{x})| > t) - \Pr_T(|h(\mathbf{x}) - h'(\mathbf{x})| > t) \right) dt \right| \\ &\qquad \leq \int_0^1 \left| \Pr_S(|h(\mathbf{x}) - h'(\mathbf{x})| > t) - \Pr_T(|h(\mathbf{x}) - h'(\mathbf{x})| > t) \right| dt \\ &\qquad \leq \sup_{t \in [0,1]} \left| \Pr_S(|h(\mathbf{x}) - h'(\mathbf{x})| > t) - \Pr_T(|h(\mathbf{x}) - h'(\mathbf{x})| > t) \right|\end{aligned}$$

Now in view of (6) and the definition of $\tilde{\mathcal{H}}$, we have:

$$\begin{aligned}&\sup_{h,h' \in \mathcal{H}} \sup_{t \in [0,1]} \left| \Pr_S(|h(\mathbf{x}) - h'(\mathbf{x})| > t) - \Pr_T(|h(\mathbf{x}) - h'(\mathbf{x})| > t) \right| \\ &= \sup_{\tilde{h} \in \tilde{\mathcal{H}}} |\Pr_S(\tilde{h}(\mathbf{x}) = 1) - \Pr_T(\tilde{h}(\mathbf{x}) = 1)| \\ &= \sup_{A \in \mathcal{A}_{\tilde{\mathcal{H}}}} |\Pr_S(A) - \Pr_T(A)| \\ &= d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T)\end{aligned}$$

Combining all the inequalities above finishes the proof. ■

Lemma 4.2. Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and $\mathcal{D}$ be any distribution over $\mathcal{X}$. For any $h, h', h'' \in \mathcal{H}$, we have $\varepsilon_{\mathcal{D}}(h,h') \leq \varepsilon_{\mathcal{D}}(h,h'') + \varepsilon_{\mathcal{D}}(h'',h')$.

Proof.

$$\begin{aligned}\varepsilon_{\mathcal{D}}(h,h') &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|h(\mathbf{x}) - h'(\mathbf{x})|] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|h(\mathbf{x}) - h''(\mathbf{x}) + h''(\mathbf{x}) - h'(\mathbf{x})|] \\ &\leq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[|h(\mathbf{x}) - h''(\mathbf{x})| + |h''(\mathbf{x}) - h'(\mathbf{x})|] = \varepsilon_{\mathcal{D}}(h,h'') + \varepsilon_{\mathcal{D}}(h'',h')\end{aligned}$$

■

Theorem 4.1. Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains, respectively. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\begin{aligned}\varepsilon_T(h) \leq \varepsilon_S(h) &+ d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) \\ &+ \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}.\end{aligned}$$

Proof. On one hand, with Lemma 4.1 and Lemma 4.2, we have $\forall h \in \mathcal{H}$:

$$\varepsilon_T(h) = \varepsilon_T(h, f_T) \leq \varepsilon_S(h, f_T) + d_{\tilde{H}}(\mathcal{D}_S, \mathcal{D}_T) \leq \varepsilon_S(h) + \varepsilon_S(f_S, f_T) + d_{\tilde{H}}(\mathcal{D}_S, \mathcal{D}_T).$$

On the other hand, by changing the order of two triangle inequalities, we also have:

$$\varepsilon_T(h) = \varepsilon_T(h, f_T) \leq \varepsilon_T(h, f_S) + \varepsilon_T(f_S, f_T) \leq \varepsilon_S(h) + \varepsilon_T(f_S, f_T) + d_{\tilde{H}}(\mathcal{D}_S, \mathcal{D}_T).$$

Realize that by definition $\varepsilon_S(f_S, f_T) = \mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|]$ and $\varepsilon_T(f_S, f_T) = \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]$. Combining the above two inequalities completes the proof. ■

Lemma 4.3. Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, then for all $\delta > 0$, w.p. at least $1-\delta$, the following inequality holds for all $h \in \mathcal{H}$: $\varepsilon_S(h) \leq \widehat{\varepsilon}_S(h) + 2\text{Rad}_{\mathbf{S}}(\mathcal{H}) + 3\sqrt{\log(2/\delta)/2n}$.

Proof. Consider the source domain $\mathcal{D}_S$. For $\forall h \in \mathcal{H}$, define the loss function $\ell : \mathcal{X} \to [0,1]$ as $\ell(\mathbf{x}) := |h(\mathbf{x}) - f_S(\mathbf{x})|$. First, we know that $\text{Rad}_{\mathbf{S}}(\mathcal{H} - f_S) = \text{Rad}_{\mathbf{S}}(\mathcal{H})$ where we slightly abuse the notation $\mathcal{H} - f_S$ to mean the family of functions $\{h - f_S \mid \forall h \in \mathcal{H}\}$:

$$\begin{aligned}
\text{Rad}_{\mathbf{S}}(\mathcal{H} - f_S) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h' \in \mathcal{H} - f_S} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h'(\mathbf{x}_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i (h(\mathbf{x}_i) - f_S(\mathbf{x}_i))\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right] + \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i f_S(\mathbf{x}_i)\right] \\
&= \text{Rad}_{\mathbf{S}}(\mathcal{H})
\end{aligned}$$

Observe that the function $\phi : t \to |t|$ is 1-Lipschitz continuous, then by Ledoux-Talagrand's contraction lemma, we can conclude that

$$\text{Rad}_{\mathbf{S}}(\phi \circ (\mathcal{H} - f_S)) \leq \text{Rad}_{\mathbf{S}}(\mathcal{H} - f_S) = \text{Rad}_{\mathbf{S}}(\mathcal{H})$$

Using Lemma B.1 with the above arguments and realize that $\varepsilon_S(h) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_S}[|h(\mathbf{x}) - f_S(\mathbf{x})|]$ finishes the proof. ■

Lemma 4.4. Let $\tilde{\mathcal{H}}$, $\mathcal{D}$ and $\widehat{\mathcal{D}}$ be defined above, then for all $\delta > 0$, w.p. at least $1-\delta$, the following inequality holds for all $h \in \tilde{\mathcal{H}}$: $\mathbb{E}_{\mathcal{D}}[\mathbb{I}_h] \leq \mathbb{E}_{\widehat{\mathcal{D}}}[\mathbb{I}_h] + 2\text{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 3\sqrt{\log(2/\delta)/2n}$.

Proof. Note that $\mathbb{I}_h \in \{0,1\}$, hence this lemma directly follows Lemma B.1. ■

Lemma 4.5. Let $\tilde{\mathcal{H}}$, $\mathcal{D}, \mathcal{D}'$ and $\widehat{\mathcal{D}}, \widehat{\mathcal{D}}'$ be defined above, then for $\forall \delta > 0$, w.p. at least $1-\delta$, for $\forall h \in \tilde{\mathcal{H}}$:

$$d_{\tilde{\mathcal{H}}}(\mathcal{D}, \mathcal{D}') \leq d_{\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}, \widehat{\mathcal{D}}') + 4\text{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 6\sqrt{\log(4/\delta)/2n}.$$

Proof. By the triangular inequality of $d_{\tilde{\mathcal{H}}}(\cdot,\cdot)$, we have:

$$d_{\tilde{\mathcal{H}}}(\mathcal{D}, \mathcal{D}') \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}, \widehat{\mathcal{D}}) + d_{\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}, \widehat{\mathcal{D}}') + d_{\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}', \mathcal{D}').$$

Now with Lemma 4.4, we know that with probability $\geq 1 - \delta/2$, we have:

$$d_{\tilde{\mathcal{H}}}(\mathcal{D}, \widehat{\mathcal{D}}) \leq 2\text{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 3\sqrt{\log(4/\delta)/2n}.$$

Similarly, with probability $\geq 1 - \delta/2$, the following inequality also holds:

$$d_{\tilde{\mathcal{H}}}(\mathcal{D}', \widehat{\mathcal{D}}') \leq 2\text{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) + 3\sqrt{\log(4/\delta)/2n}.$$

A union bound to combine the above two inequalities then finishes the proof. ■

Theorem 4.2. Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains, and let $\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T$ be the empirical source and target distributions constructed from sample $\mathbf{S} = \{\mathbf{S}_S, \mathbf{S}_T\}$, each of size n. Then for any $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ and $\forall h \in \mathcal{H}$:

$$\begin{aligned}
\varepsilon_T(h) \leq \widehat{\varepsilon}_S(h) &+ d_{\tilde{\mathcal{H}}}(\widehat{\mathcal{D}}_S, \widehat{\mathcal{D}}_T) + 2\text{Rad}_{\mathbf{S}}(\mathcal{H}) + 4\text{Rad}_{\mathbf{S}}(\tilde{\mathcal{H}}) \\
&+ \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\} \\
&+ O\left(\sqrt{\log(1/\delta)/n}\right),
\end{aligned}$$

where $\tilde{\mathcal{H}} := \{\text{sgn}(|h(\mathbf{x}) - h'(\mathbf{x})| - t) | h, h' \in \mathcal{H}, t \in [0,1]\}$.

Proof. By Theorem 4.1, the following inequality holds:

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}.$$

To get probabilistic bounds for both $\varepsilon_S(h)$ and $d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T)$, we apply Lemma 4.3 and Lemma 4.5, respectively. The final step, again, is to use a union bound to combine all the inequalities above, which completes the proof. ■

Lemma 4.6. Let $\mathcal{D}_S^Z$ and $\mathcal{D}_T^Z$ be two distributions over $\mathcal{Z}$ and let $\mathcal{D}_S^Y$ and $\mathcal{D}_T^Y$ be the induced distributions over $\mathcal{Y}$ by function $h : \mathcal{Z} \mapsto \mathcal{Y}$, then

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z). \tag{3}$$

Proof. Let B be a uniform random variable taking value in $\{0, 1\}$ and let the random variable Y_B with distribution $\mathcal{D}_B^Y$ (resp. Z_B with distribution $\mathcal{D}_B^Z$) be the mixture of $\mathcal{D}_S^Y$ and $\mathcal{D}_T^Y$ (resp. $\mathcal{D}_S^Z$ and $\mathcal{D}_T^Z$) according to B. We know that:

$$D_{\text{JS}}(\mathcal{D}_S^Z \mid\mid \mathcal{D}_T^Z) = I(B; Z_B), \quad \text{and} \quad D_{\text{JS}}(\mathcal{D}_S^Y \mid\mid \mathcal{D}_T^Y) = I(B; Y_B). \tag{7}$$

Since $\mathcal{D}_S^Y$ (resp. $\mathcal{D}_T^Y$) is induced by the function $h : \mathcal{Z} \mapsto \mathcal{Y}$ from $\mathcal{D}_S^Z$ (resp. $\mathcal{D}_T^Z$), by linearity, we also have $\mathcal{D}_B^Y$ is induced by h from $\mathcal{D}_B^Z$. Hence $Y_B = h(Z_B)$ and the following Markov chain holds:

$$B \to Z_B \to Y_B.$$

Apply the data processing inequality (Lemma B.4), we have

$$D_{\text{JS}}(\mathcal{D}_S^Z \mid\mid \mathcal{D}_T^Z) = I(B; Z_B) \geq I(B; Y_B) = D_{\text{JS}}(\mathcal{D}_S^Y \mid\mid \mathcal{D}_T^Y).$$

Taking square root on both sides of the above inequality completes the proof. ■

Lemma 4.7. Let $Y = f(X) \in \{0, 1\}$ where $f(\cdot)$ is the labeling function and $\hat{Y} = h(g(X)) \in \{0, 1\}$ be the prediction function, then $d_{\text{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}}) \leq \sqrt{\varepsilon(h \circ g)}$.

Proof.

$$\begin{aligned}
d_{\text{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}}) &= \sqrt{D_{\text{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}})} \\
&\leq \sqrt{||\mathcal{D}^Y - \mathcal{D}^{\hat{Y}}||_1/2} && \text{(Lemma B.3)} \\
&= \sqrt{\Big(|\Pr(Y=0) - \Pr(\hat{Y}=0)| + |\Pr(Y=1) - \Pr(\hat{Y}=1)|\Big)/2} \\
&= \sqrt{|\Pr(Y=1) - \Pr(\hat{Y}=1)|} \\
&= \sqrt{|\mathbb{E}_X[f(X)] - \mathbb{E}_X[h(g(X))]|} \\
&\leq \sqrt{\mathbb{E}_X[|f(X) - h(g(X))|]} \\
&= \sqrt{\varepsilon(h \circ g)}
\end{aligned}$$

■

Lemma 4.8. Suppose the Markov chain $X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$ holds, then

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) + \sqrt{\varepsilon_S(h \circ g)} + \sqrt{\varepsilon_T(h \circ g)}.$$

Proof. Since $X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$ forms a Markov chain, by Lemma 4.6, the following inequality holds:

$$d_{\text{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) \leq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z).$$

On the other hand, since $d_{\text{JS}}(\cdot, \cdot)$ is a distance metric, we also have:

$$d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\text{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) + d_{\text{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y) \leq d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) + d_{\text{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y).$$

Applying Lemma 4.7 to both $d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}})$ and $d_{\text{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y)$ then finishes the proof. ■

Theorem 4.3. Suppose the condition in Lemma 4.8 holds and $d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then:

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2.$$

Proof. In view of the result in Theorem 4.8, applying the AM-GM inequality, we have:

$$\sqrt{\varepsilon_S(h \circ g)} + \sqrt{\varepsilon_T(h \circ g)} \leq \sqrt{2\left(\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g)\right)}.$$

Now since $d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, simple algebra shows

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\text{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\text{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2.$$

■

B. Technical Tools

The following lemma is particularly useful to provide data-dependent guarantees in terms of the empirical Rademacher complexity:

Lemma B.1 (Bartlett & Mendelson (2002))**.** Let $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, then for $\forall \delta > 0$, w.p.b. at least $1-\delta$, the following inequality holds for $\forall h \in \mathcal{H}$:

$$\mathbb{E}[h(\mathbf{x})] \leq \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i) + 2\text{Rad}_{\mathbf{S}}(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} \tag{8}$$

Ledoux-Talagrand's contraction lemma is a useful technique in upper bounding the Rademacher complexity of function compositions:

Lemma B.2 (Ledoux-Talagrand's contraction lemma)**.** Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be a Lipschitz function with parameter L, i.e., $\forall a, b \in \mathbb{R}$, $|\phi(a) - \phi(b)| \leq L|a-b|$. Then,

$$\text{Rad}_{\mathbf{S}}(\phi \circ \mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \phi(h(\mathbf{x}_i))\right] \leq L\, \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right] = L\, \text{Rad}_{\mathbf{S}}(\mathcal{H}),$$

where $\phi \circ \mathcal{H} := \{\phi \circ h \mid h \in \mathcal{H}\}$ is the class of composite functions.

Lin's lemma gives an upper bound of the JS divergence between two distributions via the L_1 distance (total variation distance).

Lemma B.3 (Theorem. 3, (Lin, 1991))**.** Let $\mathcal{D}$ and $\mathcal{D}'$ be two distributions, then $D_{\text{JS}}(\mathcal{D}, \mathcal{D}') \leq \frac{1}{2}||\mathcal{D} - \mathcal{D}'||_1$.

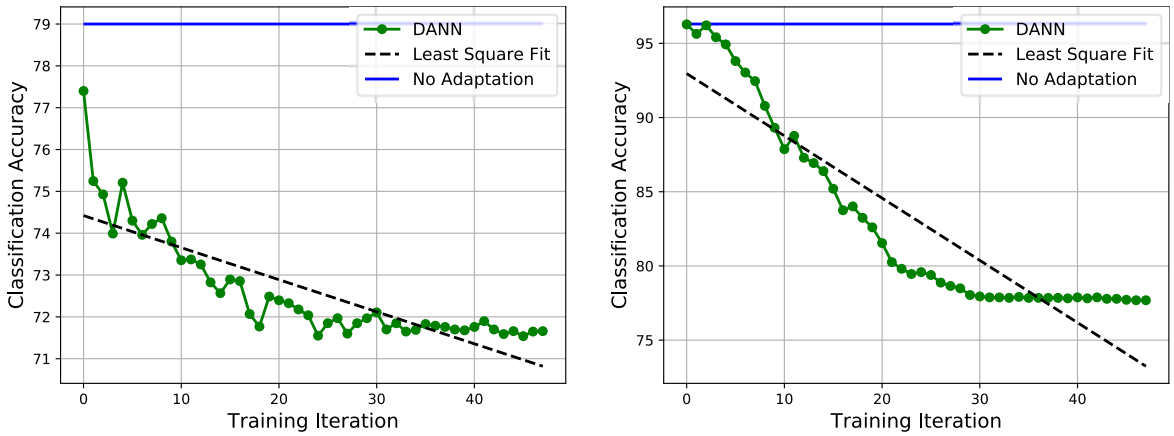
Lemma B.4 (Data processing inequality)**.** Let $X \to Z \to Y$ be a Markov chain, then $I(X;Z) \geq I(X;Y)$, where $I(\cdot;\cdot)$ is the mutual information.

C. Additional Experiments

In order to further validate our claims, we artificially unbalance the label distribution on the source domain by removing samples from the dataset. We perform two such modifications:

- **Unbalanced digits** In our first experiment, the source domain is MNIST, from which we randomly remove 70% of the first five classes (corresponding to digits 0 through 4) while leaving the other classes untouched. The target domain is the full USPS dataset.
- **Unbalanced zeros and ones** In our second experiment, the source domain is still MNIST. We remove 70% of the 0 class and all the classes above 2 entirely. We still target the USPS dataset, but also remove digits 2 to 9 in that dataset.

The results of the DANN domain adaptation algorithm on those tasks are plotted in Figure 4. They confirm the theoretical and experimental findings from the main text. The effect is however enhanced due to a much larger discrepancy between the label distributions (a fact predicted by our theory). Those plots are the mean across 5 seeds, the standard deviation over those 5 runs is significantly lower than the observed trend.



(a) Unbalanced digits

(b) Unbalanced zeros and ones

Figure 4. Digit classification on the unbalanced MNIST to USPS domain adaptation tasks described above. The horizontal solid line corresponds to the target domain test accuracy without adaptation. The green solid line is the target domain test accuracy under domain adaptation with DANN. We also plot the least square fit (dashed line) of the DANN adaptation results to emphasize the negative slope.