Supplemental Material: Bridging Theory and Algorithm for Domain Adaptation

Yuchen Zhang [*1 2] **Tianle Liu** [*2 3] **Mingsheng Long** [1 2] **Michael I. Jordan** [4]

A. Properties of DD

Proposition A.1. *Let $\mathcal{P}$ denote the space of probability distributions over the domain $\mathcal{X}$. For any $h \in \mathcal{H}$ in the setting of binary classification, the induced Disparity Discrepancy $d_{h,\mathcal{H}}(\cdot,\cdot)$ is a pseudometric on $\mathcal{P}$. More precisely it is a metric on some quotient space of $\mathcal{P}$.*

Proof. Firstly, $d_{h,\mathcal{H}}(P,Q)$ is non-negative since $h \in \mathcal{H}$ and $d_{h,\mathcal{H}}(P,P) = 0$ holds for any $P \in \mathcal{P}$ by definition.

Secondly, $d_{h,\mathcal{H}}(P,Q)$ is symmetric. Otherwise suppose $d_{h,\mathcal{H}}(P,Q) > d_{h,\mathcal{H}}(Q,P)$. Therefore, we can choose $g \in \mathcal{H}$ such that $\mathbb{E}_Q \mathbb{1}[g \neq h] - \mathbb{E}_P \mathbb{1}[g \neq h] > d_{h,\mathcal{H}}(Q,P)$. By our assumption, $1 - g \in \mathcal{H}$. Thus

$$\begin{aligned} & d_{h,\mathcal{H}}(Q,P) \\ \geq & \mathbb{E}_P \mathbb{1}[1-g \neq h] - \mathbb{E}_Q \mathbb{1}[1-g \neq h] \\ = & \mathbb{E}_Q \mathbb{1}[g \neq h] - \mathbb{E}_P \mathbb{1}[g \neq h] \\ > & d_{h,\mathcal{H}}(Q,P). \end{aligned}$$

Contradiction.

Lastly, for any distribution P, Q, R we have

$$\begin{aligned} & d_{h,\mathcal{H}}(P,Q) \\ = & \sup_{h' \in \mathcal{H}} (\text{disp}_Q(h',h) - \text{disp}_P(h',h)) \\ \leq & \sup_{h' \in \mathcal{H}} (\text{disp}_Q(h',h) - \text{disp}_R(h',h)) \\ & \quad + \sup_{h'' \in \mathcal{H}} (\text{disp}_R(h'',h) - \text{disp}_P(h'',h)) \\ = & d_{h,\mathcal{H}}(R,Q) + d_{h,\mathcal{H}}(R,P). \end{aligned}$$

Thus $d_{h,\mathcal{H}}(P,Q)$ is a pseudometric on $\mathcal{P}$.

Note that $d_{h,\mathcal{H}}(P,Q) = 0$ does not imply $P = Q$ in general. However, this equation gives a equivalence relation $\sim_{h,\mathcal{H}}$ on $\mathcal{P}$, which could be easily checked noticing that $d_{h,\mathcal{H}}(P,Q) = 0$ is equivalent to

$$\mathbb{E}_Q \mathbb{1}[h' \neq h] = \mathbb{E}_P \mathbb{1}[h' \neq h]$$

for any $h' \in \mathcal{H}$. Thus $d_{h,\mathcal{H}}(\cdot,\cdot)$ is a metric on the quotient space $\mathcal{P}/\sim_{h,\mathcal{H}}$. □

Next in the binary classification setting we show that there are strong connections between the $\mathcal{H}\Delta\mathcal{H}$-distance and disparity discrepancy. For a hypothesis set $\mathcal{H}$, the *symmetric difference hypothesis set* $\mathcal{H}\Delta\mathcal{H}$ is the set of classifiers

$$\mathcal{H}\Delta\mathcal{H} \triangleq \{h - h' | h, h' \in \mathcal{H}\}. \tag{1}$$

Proposition A.2. *For the binary classification,*

$$d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) = \sup_{h \in \mathcal{H}} d_{h,\mathcal{H}}(P,Q). \tag{2}$$

Proof. By definition,

$$\begin{aligned} \sup_{h \in \mathcal{H}} d_{h,\mathcal{H}}(P,Q) &= \sup_{h,h' \in \mathcal{H}} (\text{disp}_Q(h',h) - \text{disp}_P(h',h)) \\ &= \sup_{g \in \mathcal{H}\Delta\mathcal{H}} (\mathbb{E}_Q \mathbb{1}[g \neq 0] - \mathbb{E}_P \mathbb{1}[g \neq 0]) \\ &= \sup_{g \in \mathcal{H}\Delta\mathcal{H}} (\mathbb{E}_Q g - \mathbb{E}_P g) \\ &= d_{\mathcal{H}\Delta\mathcal{H}}(P,Q). \end{aligned}$$

□

Now we consider when our proposed discrepancy is independent of the selection of h, in which case the disparity discrepancy is actually equivalent to $\mathcal{H}\Delta\mathcal{H}$-distance. A sufficient condition is stated below:

Proposition A.3. *For the binary classification, if the hypothesis set $\mathcal{H}$ is a linear space over the prime field $\mathbb{Z}_2$, in other words $\mathcal{H}\Delta\mathcal{H} = \mathcal{H}$, we have*

$$d_{h,\mathcal{H}}(P,Q) = d_{\mathcal{H}\Delta\mathcal{H}}(P,Q) = d_{\mathcal{H}}(P,Q) \tag{3}$$

for any $h \in \mathcal{H}$.

[*]Equal contribution [1]School of Software [2]Research Center for Big Data, BNRist [3]Department of Mathematical Science, Tsinghua University, China [4]University of California, Berkeley, USA.

[†]Yuchen Zhang <zhangyuc17@mails.tsinghua.edu.cn>. Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Proof. Suppose there exist $h, h' \in \mathcal{H}$ such that

$$d_{h,\mathcal{H}}(P,Q) > d_{h',\mathcal{H}}(P,Q),$$

Then by the definition of $d_{h,\mathcal{H}}(P,Q)$, for any $\epsilon > 0$ there exists $g \in \mathcal{H}$ such that

$$d_{h,\mathcal{H}}(P,Q) - (\mathbb{E}_Q \mathbb{1}[g \neq h] - \mathbb{E}_P \mathbb{1}[g \neq h]) < \epsilon.$$

Let $\epsilon = \frac{1}{2}(d_{h,\mathcal{H}}(P,Q) - d_{h',\mathcal{H}}(P,Q))$, then consider $g' = g - h + h'$

$$\begin{aligned}&\mathbb{E}_Q \mathbb{1}[g' \neq h'] - \mathbb{E}_P \mathbb{1}[g' \neq h'] \\ =&\mathbb{E}_Q \mathbb{1}[g \neq h] - \mathbb{E}_P \mathbb{1}[g \neq h] \\ >&d_{h',\mathcal{H}}(P,Q).\end{aligned}$$

Contradiction. □

However, in the case of neural networks, especially with activations such as the rectified linear unit (ReLU), the condition mentioned above is generally not satisfied and $\mathcal{H}\Delta\mathcal{H}$-distance is often strictly larger than ours. To verify this, we provide the following example:

Example A.4. *Let $\mathcal{X} = \mathbb{R}^2$ and P, Q be two dirac masses on the points $(-1, 1)$ and $(1, -1)$ respectively. Let a, b be two parameters with values in $\mathbb{R}$. Let $r(t) \triangleq \max\{0, t\}$ be the ReLU function. For any input data $x = (x_1, x_2) \in \mathbb{R}^2$, the pseudo label predicted by a hypothesis $h \in \mathcal{H}$ is defined as follows:*

$$h(x) = \begin{cases} 1 & \text{if } r(x_1 - a) \geq r(x_2 - b) \\ 0 & \text{if } r(x_1 - a) < r(x_2 - b) \end{cases}.$$

One can check that there are three kinds of hypotheses with values

$$h_1(P,Q) = (0,0), h_2(P,Q) = (0,1), h_3(P,Q) = (1,1).$$

Then $d_{h_1,\mathcal{H}}(P,Q) = 1$ and $d_{h_3,\mathcal{H}}(P,Q) = 0$, in which case $d_{h_3,\mathcal{H}}(P,Q)$ does not coincide with $d_{\mathcal{H}\Delta\mathcal{H}}(P,Q)$.

B. Generalization Bounds with DD

Lemma B.1 (Rademacher Generalization Bound, Theorem 3.1 of Mohri et al. (2012))**.** *Suppose that $\mathcal{G}$ is a class of function maps $\mathcal{X} \to [0,1]$. For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $g \in \mathcal{G}$:*

$$|\mathbb{E}_D\, g - \mathbb{E}_{\widehat{D}}\, g| \leq 2\mathfrak{R}_{n,D}(\mathcal{G}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \tag{4}$$

Theorem B.2. *For any classifier h*

$$\text{err}_Q(h) \leq \text{err}_P(h) + d_{h,\mathcal{H}}(P,Q) + \lambda, \tag{5}$$

where $\lambda = \lambda(\mathcal{H}, P, Q)$ is independent of h.

Proof. Let h^* be the ideal joint classifier which minimizes the combined error,

$$h^* \triangleq \arg\min_{h \in \mathcal{H}} \{\text{err}_P(h) + \text{err}_Q(h)\}.$$

Set $\lambda = \text{err}_P(h^*) + \text{err}_Q(h^*)$. Then

$$\begin{aligned}&\text{err}_Q(h) = \text{err}_P(h) + \text{err}_Q(h) - \text{err}_P(h) \\ \leq&\text{err}_P(h) + (\mathbb{E}_Q \mathbb{1}[h^* \neq h] - \mathbb{E}_P \mathbb{1}[h^* \neq h]) \\ &+ (\text{err}_P(h^*) + \text{err}_Q(h^*)) \\ \leq&\text{err}_P(h) + \sup_{h' \in \mathcal{H}} (\text{disp}_Q(h', h) - \text{disp}_P(h', h)) + \lambda \\ =&\text{err}_P(h) + d_{h,\mathcal{H}}(P,Q) + \lambda.\end{aligned}$$

□

Theorem B.3. *Suppose $\mathcal{H}$ is a hypothesis space maps $\mathcal{X}$ to $\{0,1\}$. $\widehat{D}$ is empirical distribution corresponding to datasets contains n data points sampled from D. For any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h, h' \in \mathcal{H}$:*

$$\begin{aligned}&|\text{disp}_D(h', h) - \text{disp}_{\widehat{D}}(h', h)| \\ &\qquad \leq 2\mathfrak{R}_{n,D}(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.\end{aligned} \tag{6}$$

Proof.

$$\begin{aligned}&\sup_{h,h' \in \mathcal{H}} |\mathbb{E}_{\widehat{D}} \mathbb{1}[h' \neq h] - \mathbb{E}_D \mathbb{1}[h' \neq h]| \\ =&\sup_{g \in \mathcal{H}\Delta\mathcal{H}} |\mathbb{E}_{\widehat{D}} \mathbb{1}[g \neq \mathbf{1}] - \mathbb{E}_D \mathbb{1}[g \neq \mathbf{1}]| \\ =&\sup_{g \in \mathcal{H}\Delta\mathcal{H}} |\mathbb{E}_{\widehat{D}} g - \mathbb{E}_D g|.\end{aligned}$$

With Lemma B.1, we could know that

$$\sup_{g \in \mathcal{H}\Delta\mathcal{H}} |\mathbb{E}_D\, g - \mathbb{E}_{\widehat{D}}\, g| \leq 2\mathfrak{R}_{n,D}(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

□

Theorem B.4. *For any $\delta > 0$ and binary classifier $h \in \mathcal{H}$, with probability $1 - 3\delta$, we have*

$$\begin{aligned}\text{err}_Q(h) \leq&\ \text{err}_{\widehat{P}}(h) + d_{h,\mathcal{H}}(\widehat{P}, \widehat{Q}) + \lambda \\ &+ 2\mathfrak{R}_{n,P}(\mathcal{H}\Delta\mathcal{H}) + 2\mathfrak{R}_{n,P}(\mathcal{H}) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\ &+ 2\mathfrak{R}_{m,Q}(\mathcal{H}\Delta\mathcal{H}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.\end{aligned} \tag{7}$$

where $\lambda = \min_{h' \in \mathcal{H}}(\text{err}_P(h') + \text{err}_Q(h'))$ is independent with h.

Proof. Consider the difference of expected and empirical terms on the right-hand side.

$$\begin{aligned}
&\sup_{h\in\mathcal{H}}(\text{err}_P(h)+d_{h,\mathcal{H}}(P,Q)-\text{err}_{\widehat{P}}(h)-d_{h,\mathcal{H}}(\widehat{P},\widehat{Q}))\\
=&\sup_{h\in\mathcal{H}}(\text{err}_P(h)-\text{err}_{\widehat{P}}(h)+d_{h,\mathcal{H}}(P,Q)-d_{h,\mathcal{H}}(\widehat{P},\widehat{Q}))\\
\leq&\sup_{h\in\mathcal{H}}(\text{err}_P(h)-\text{err}_{\widehat{P}}(h))+\sup_{h\in\mathcal{H}}(d_{h,\mathcal{H}}(P,Q)-d_{h,\mathcal{H}}(\widehat{P},\widehat{Q})).
\end{aligned}$$

First by Lemma B.1, $\forall\ \delta>0$, with probability $1-\delta$,

$$\sup_{h\in\mathcal{H}}(\text{err}_P(h)-\text{err}_{\widehat{P}}(h))\leq 2\mathfrak{R}_{n,P}(\mathcal{H})+\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Then we bound the difference between $d_{h,\mathcal{H}}(P,Q)$ and $d_{h,\mathcal{H}}(\widehat{P},\widehat{Q})$:

$$\begin{aligned}
&d_{h,\mathcal{H}}(P,Q)-d_{h,\mathcal{H}}(\widehat{P},\widehat{Q})\\
=&\sup_{h'\in\mathcal{H}}(\text{disp}_Q(h',h)-\text{disp}_P(h',h))\\
&\quad-\sup_{h''\in\mathcal{H}}(\text{disp}_{\widehat{Q}}(h'',h)-\text{disp}_{\widehat{P}}(h'',h))\\
\leq&\sup_{h'\in\mathcal{H}}\big(\text{disp}_Q(h',h)-\text{disp}_P(h',h)\\
&\quad-\text{disp}_{\widehat{Q}}(h',h)+\text{disp}_{\widehat{P}}(h',h)\big)\\
\leq&\sup_{h'\in\mathcal{H}}(\text{disp}_Q(h',h)-\text{disp}_{\widehat{Q}}(h',h))\\
&\quad+\sup_{h''\in\mathcal{H}}(\text{disp}_{\widehat{P}}(h'',h)-\text{disp}_P(h'',h)).
\end{aligned}$$

Take supremum over $h\in\mathcal{H}$, we have:

$$\begin{aligned}
&\sup_{h\in\mathcal{H}}(d_{h,\mathcal{H}}(P,Q)-d_{h,\mathcal{H}}(\widehat{P},\widehat{Q}))\\
\leq&\sup_{h,h'\in\mathcal{H}}|\text{disp}_Q(h',h)-\text{disp}_{\widehat{Q}}(h',h)|+\\
&\sup_{h,h''\in\mathcal{H}}|\text{disp}_{\widehat{P}}(h'',h)-\text{disp}_P(h'',h)|.
\end{aligned}$$

From Theorem B.2, we directly get:

$$\begin{aligned}
&\sup_{h\in\mathcal{H}}(d_{h,\mathcal{H}}(P,Q)-d_{h,\mathcal{H}}(\widehat{P},\widehat{Q}))\\
\leq&2\mathfrak{R}_{n,P}(\mathcal{H}\Delta\mathcal{H})+2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+2\mathfrak{R}_{m,Q}(\mathcal{H}\Delta\mathcal{H})+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.
\end{aligned}$$

Combine the two parts of inequality, we get the final result. □

We introduce theory of VC-dimension here to further measure the generalization ability.

Definition B.5 (VC-Dimension)**.** *The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be fully shattered by $\mathcal{H}$. Let*

$$\Pi(n,\mathcal{H})\triangleq\max_{x_1,\ldots,x_n}\big|\{h(x_1),\ldots,h(x_n)\big|h\in\mathcal{H}\}\big|. \tag{8}$$

Then

$$\text{VC}(\mathcal{H})\triangleq\max\{m\big|\Pi(m,\mathcal{H})=2^m\}. \tag{9}$$

Lemma B.6 (Corollary 3.1 & 3.3 of Mohri et al. (2012))**.** *Suppose $\mathcal{G}$ takes value in $\{0,1\}$ and d is the VC-dimension of $\mathcal{G}$. Then the Rademacher complexity of $\mathcal{G}$ has the following holds for all $h\in\mathcal{H}$,*

$$\mathfrak{R}_{n,D}(\mathcal{G})\leq\frac{1}{2}\sqrt{\frac{2d\log\frac{en}{d}}{n}}. \tag{10}$$

Theorem B.7. *For any $\delta>0$ and $h\in\mathcal{H}$, with probability $1-3\delta$, we have*

$$\begin{aligned}
\text{err}_Q(h)\leq&\,\text{err}_{\widehat{P}}(h)+d_{h,\mathcal{H}}(\widehat{P},\widehat{Q})+\lambda\\
&+C_1\sqrt{\frac{d\log\frac{en}{d}}{n}}+C_2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\\
&+C_3\sqrt{\frac{d\log\frac{em}{d}}{m}}+C_4\sqrt{\frac{\log\frac{2}{\delta}}{2m}}+\lambda,
\end{aligned} \tag{11}$$

where C_1,C_2 are constants independent of $\mathcal{H},P,Q$.

Proof. 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 $\mathcal{H}$ (Anthony & Bartlett, 2009). Let $g\triangleq 1+h-h'$, then $g\in\mathcal{H}\Delta\mathcal{H}$ and $h\neq h'$ is equivalent to $g\neq 1$. Thus by Lemma B.6 we have

$$\mathfrak{R}_{n,P}(\mathcal{H}\Delta\mathcal{H})\leq\ \frac{1}{2}\sqrt{\frac{2d\log\frac{en}{4d}}{n}}.$$

To summarize, with probability $1-3\delta$,

$$\begin{aligned}
\text{err}_Q(h)&\leq\text{err}_P(h)+d_{h,\mathcal{H}}(P,Q)+\lambda\\
&\leq\text{err}_{\widehat{P}}(h)+d_{h,\mathcal{H}}(\widehat{P},\widehat{Q})\\
&\quad+4\sqrt{\frac{d\log\frac{en}{d}}{n}}+2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\\
&\quad+2\sqrt{\frac{d\log\frac{em}{d}}{m}}+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}+\lambda.
\end{aligned}$$

□

C. Generalization Bounds with MDD

Lemma C.1. *For any distribution D and any f, we have*

$$\mathrm{disp}_D^{(\rho)}(f', f) \le \mathrm{err}_D^{(\rho)}(f') + \mathrm{err}_D^{(\rho)}(f). \tag{12}$$

Proof. We prove that for any (x_i, y_i),

$$\Phi_\rho \circ \rho_{f'}(x_i, h_f(x_i)) \le \Phi_\rho \circ \rho_{f'}(x_i, y_i) + \Phi_\rho \circ \rho_f(x_i, y_i),$$

If $h_f(x_i) \neq y_i$ or $h_{f'}(x_i) \neq y_i$, the right side of above equation will reach 1, which is a trivial upper bound for the left part. Otherwise $h_f(x_i) = h_{f'}(x_i) = y_i$, and

$$\begin{aligned}&\Phi_\rho \circ \rho_{f'}(x_i, h_f(x_i))\\ \le&\Phi_\rho \circ \rho_{f'}(x_i, h_f(x_i)) + \Phi_\rho \circ \rho_f(x_i, y_i)\\ =&\Phi_\rho \circ \rho_{f'}(x_i, y_i) + \Phi_\rho \circ \rho_f(x_i, y_i).\end{aligned}$$

Take expectation on distribution P and we get the result. □

Theorem C.2 (Proposition 3.3)**.** *For any scoring function f,*

$$\mathrm{err}_Q(h_f) \le \mathrm{err}_P^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(P, Q) + \lambda, \tag{13}$$

where $\lambda = \lambda(\rho, \mathcal{F}, P, Q)$ is a constant independent of f.

Proof. Let f^* be the ideal joint hypothesis which minimizes the combined margin loss,

$$f^* \triangleq \arg\min_{f \in \mathcal{H}} \{\mathrm{err}_P^{(\rho)}(f) + \mathrm{err}_Q^{(\rho)}(f)\}.$$

Set $\lambda = \mathrm{err}_P^{(\rho)}(f^*) + \mathrm{err}_Q^{(\rho)}(f^*)$. Then by Lemma C.1,

$$\begin{aligned}\mathrm{err}_Q(f) &\le \mathbb{E}_Q \mathbb{1}[h_f \neq h_{f^*}] + \mathbb{E}_Q \mathbb{1}[h_{f^*} \neq y]\\ &\le \mathrm{err}_P^{(\rho)}(f) - \mathrm{err}_P^{(\rho)}(f)\\ &\quad + \mathrm{disp}_Q^{(\rho)}(f^*, f) + \mathrm{err}_Q^{(\rho)}(f^*)\\ &\le \mathrm{err}_P^{(\rho)}(f) + \mathrm{err}_P^{(\rho)}(f^*) - \mathrm{disp}_P^{(\rho)}(f^*, f)\\ &\quad + \mathrm{disp}_Q^{(\rho)}(f^*, f) + \mathrm{err}_Q^{(\rho)}(f^*)\\ &\le \mathrm{err}_P^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(P, Q) + \lambda\end{aligned}$$

□

Definition C.3. *Given a class of scoring functions $\mathcal{F}$ and a class of the induced classifiers $\mathcal{H}$, we define $\Pi_{\mathcal{H}}\mathcal{F}$ as*

$$\Pi_{\mathcal{H}}\mathcal{F} = \{x \mapsto f(x, h(x)) | h \in \mathcal{H}, f \in \mathcal{F}\}. \tag{14}$$

There is a geometric interpretation of the set $\Pi_{\mathcal{H}}\mathcal{F}$ (Galbis & Maestre, 2012). Assuming $\mathcal{X}$ is a manifold, assigning a vector space $\mathbb{R}^k$ to each point in $\mathcal{X}$ yields a vector bundle $\mathcal{B}$. Now regarding the values of $\mathcal{H}$ as one-hot vectors in $\mathbb{R}^k$, $\mathcal{F}$ and $\mathcal{H}$ are both sets of sections of $\mathcal{B}$ containing (probably piecewise continuous) vector fields. $\Pi_{\mathcal{H}}\mathcal{F}$ can be seen as the space of inner products of vector fields from $\mathcal{H}$ and $\mathcal{F}$,

$$\Pi_{\mathcal{H}}\mathcal{F} = \langle \mathcal{H}, \mathcal{F} \rangle = \{\langle h, f \rangle \big| h \in \mathcal{H}, f \in \mathcal{F}\}. \tag{15}$$

Lemma C.4 (A modified version of Theorem 8.1, Mohri et al. (2012))**.** *Suppose $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ is the hypothesis set of scoring functions with $\mathcal{Y} = \{1, 2, \ldots, k\}$. Let*

$$\Pi_1\mathcal{F} \triangleq \{x \mapsto f(x, y) | y \in \mathcal{Y}, f \in \mathcal{F}\}. \tag{16}$$

Fix $\rho > 0$. Then for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$:

$$|err_D^{(\rho)}(f) - err_{\widehat{D}}^{(\rho)}(f)| \le \frac{2k^2}{\rho}\mathfrak{R}_{n,D}(\Pi_1\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{17}$$

Note that a simple corollary of this lemma is the margin bound for multi-class classification:

$$\begin{aligned}\mathrm{err}_D(h_f) &\le \mathrm{err}_D^{(\rho)}(f)\\ &\le \mathrm{err}_{\widehat{D}}^{(\rho)}(f) + \frac{2k^2}{\rho}\mathfrak{R}_{n,D}(\Pi_1\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.\end{aligned} \tag{18}$$

Lemma C.5 (Talagrand's lemma, Talagrand (2014); Mohri et al. (2012))**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an ℓ-Lipschitz. Then for any hypothesis set $\mathcal{F}$ of real-valued functions, and any sample $\widehat{D}$ of size n, the following inequality holds:*

$$\widehat{\mathfrak{R}}_{\widehat{D}}(\Phi \circ \mathcal{F}) \le \ell\, \widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{F}) \tag{19}$$

Lemma C.6 (Lemma 8.1 of Mohri et al. (2012))**.** *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be k hypothesis sets in $\mathbb{R}^{\mathcal{X}}$, $k > 1$. $\mathcal{G} = \{\max\{f_1, \ldots, f_k\} : f_i \in \mathcal{F}, i \in \{1, \ldots, k\}\}$, Then for any sample $\widehat{D}$ of size n, we have*

$$\widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{G}) \le \sum_{i=1}^{k} \widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{F}_i) \tag{20}$$

Theorem C.7 (Lemma 3.6)**.** *Let $\mathcal{F} \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ is a hypothesis set. Let $\mathcal{H}$ be the set of classifiers (mapping $\mathcal{X}$ to $\mathcal{Y}$) corresponding to $\mathcal{F}$. For any $\delta > 0$, with probability $1 - 2\delta$, the following holds simultaneously for any scoring function f,*

$$\begin{aligned}&|d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) - d_{f,\mathcal{F}}^{(\rho)}(P, Q)|\\ \le&\frac{k}{\rho}\mathfrak{R}_{n,P}(\Pi_{\mathcal{H}}\mathcal{F}) + \frac{k}{\rho}\mathfrak{R}_{m,Q}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.\end{aligned} \tag{21}$$

Proof. For $\forall f, f' \in \mathcal{F}$, define the τ_f-transform of f' to be

$$\tau_f f'(x, y) = \begin{cases} f'(x, 1) & \text{if } y = h_f(x)\\ f'(x, h_f(x)) & \text{if } y = 1\\ f'(x, y) & \text{else}\end{cases}$$

where h_f is the induced classifier mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{G} = \{\tau_f f' | f, f' \in \mathcal{F}\}$, $\tilde{\mathcal{G}} = \{(x, y) \mapsto \rho_g(x, y) | g \in \mathcal{G}\}$.

Consider the family of functions $\Phi_\rho \circ \tilde{\mathcal{G}}$ which takes values in $[0,1]$. By Lemma C.4, with probability at least $1-\delta$, for $\forall g \in \mathcal{G}$.

$$\begin{aligned}&|\mathrm{err}_P^{(\rho)}(g) - \mathrm{err}_{\widehat{P}}^{(\rho)}(g)|\\ =&|\mathbb{E}\Phi_\rho \circ \rho_g(x,y) - \frac{1}{n}\sum_{i=1}^n \Phi_\rho \circ \rho_g(x_i,y_i)|\\ \leq& 2\mathfrak{R}_{n,D}(\Phi_\rho \circ \tilde{\mathcal{G}}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}\end{aligned}$$

Regard all the data as from the same class 1. Define:

$$\mathfrak{R}^0_{n,D}(\mathcal{G}) = \mathbb{E}_{(x_i,1),x_i\sim D^n}\widehat{\mathfrak{R}}_{\widehat{D}}(\mathcal{G})$$

Then the above equations becomes

$$\begin{aligned}&|\mathrm{disp}_P^{(\rho)}(g,1) - \mathrm{disp}_{\widehat{P}}^{(\rho)}(g,1)|\\ =&|\mathbb{E}\Phi_\rho \circ \rho_g(x,1) - \frac{1}{n}\sum_{i=1}^n \Phi_\rho \circ \rho_g(x_i,1)|\\ \leq& 2\mathfrak{R}^0_{n,D}(\Phi_\rho \circ \tilde{\mathcal{G}}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}\end{aligned}$$

For any $f, f' \in \mathcal{F}$, let $g = \tau_f f'$. Then $g \in \mathcal{G}$ and

$$\mathrm{disp}_P^{(\rho)}(g,1) = \mathrm{disp}_P^{(\rho)}(f',f),\ \mathrm{disp}_{\widehat{P}}^{(\rho)}(g,1) = \mathrm{disp}_{\widehat{P}}^{(\rho)}(f',f)$$

Thus,

$$\begin{aligned}&\sup_{f,f'\in\mathcal{F}}|\mathrm{disp}_P^{(\rho)}(f',f) - \mathrm{disp}_{\widehat{P}}^{(\rho)}(f',f)|\\ \leq&\sup_{g\in\mathcal{G}}|\mathrm{disp}_P^{(\rho)}(g,1) - \mathrm{disp}_{\widehat{P}}^{(\rho)}(g,1)|\\ \leq& 2\mathfrak{R}^0_{n,D}(\Phi_\rho \circ \tilde{\mathcal{G}}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}\end{aligned}$$

By Lemma C.5, $\mathfrak{R}^0_{n,D}(\Phi_\rho \circ \tilde{\mathcal{G}}) \leq \frac{1}{\rho}\mathfrak{R}^0_{n,D}(\tilde{\mathcal{G}})$

$$\begin{aligned}&\mathfrak{R}^0_{n,D}(\tilde{\mathcal{G}}) = \frac{1}{n}\mathbb{E}_{S,\sigma}(\sup_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i\rho_g(x_i,1))\\ =&\frac{1}{n}\mathbb{E}_{S,\sigma}(\sup_{g\in\mathcal{G}}\sum_{i=1}^n \sigma_i(g(x_i,1) - \max_{y\neq 1} g(x_i,y))\\ =&\frac{1}{n}\mathbb{E}_{S,\sigma}(\sup_{f,f'\in\mathcal{F}}\sum_{i=1}^n \sigma_i(f'(x_i,h_f(x_i)) - \max_{y\neq h_f(x_i)} f'(x_i,y))\\ \leq&\frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\mathcal{F},h\in\mathcal{H}}\sum_{i=1}^n \sigma_i f(x_i,h(x_i))\\ +&\frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\mathcal{F},h\in\mathcal{H}}\sum_{i=1}^n \sigma_i(-\max_{y\neq h(x_i)} f'(x_i,y))\\ =&\mathfrak{R}_{n,D}(\Pi_\mathcal{H}\mathcal{F}) + \frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\mathcal{F},h\in\mathcal{H}}\sum_{i=1}^n \sigma_i\max_{y\neq h(x_i)} f(x_i,y)\end{aligned}$$

Define the permutation

$$\xi(i) = \begin{cases} i+1 & i = 1,\dots,k-1\\ 1 & i = k\end{cases}$$

By our assumption of $\mathcal{H}$, we have the result that $\forall h \in \mathcal{H}$, $\xi^j h \in \mathcal{H},\ j = 1,2,\dots,k-1$.

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\mathcal{F},h\in\mathcal{H}}\sum_{i=1}^n \sigma_i\max_{y\neq h(x_i)} f(x_i,y)\\ =&\frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\mathcal{F},h\in\mathcal{H}}\sum_{i=1}^n \sigma_i\max_{j\in\{1,\dots,k-1\}} f(x_i,\xi^j h(x_i))\end{aligned}$$

Let $\Pi_\mathcal{H}\mathcal{F}^{(k-1)} = \{\max\{f_1,\dots,f_{k-1}\} | f_i \in \Pi_\mathcal{H}\mathcal{F}, i = 1,\dots,k-1\}$. Then applying Lemma C.6:

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f,h}\sum_{i=1}^n \sigma_i\max_{j\in\{1,\dots,k-1\}} f(x_i,\xi^j h(x_i))\\ =&\frac{1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\Pi_\mathcal{H}\mathcal{F}^{(k-1)}}\sum_{i=1}^n \sigma_i f(x_i)\\ \leq&\frac{k-1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\Pi_\mathcal{H}\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathfrak{R}^0_{n,D}(\tilde{\mathcal{G}}) \leq&\mathfrak{R}_{n,D}(\Pi_\mathcal{H}\mathcal{F}) + \frac{k-1}{n}\mathbb{E}_{S,\sigma}\sup_{f\in\Pi_\mathcal{H}\mathcal{F}}\sum_{i=1}^n \sigma_i f(x_i)\\ \leq& k\mathfrak{R}_{n,D}(\Pi_\mathcal{H}\mathcal{F}),\\ &\sup_{f,f'\in\mathcal{F}}|\mathrm{disp}_P^{(\rho)}(f',f) - \mathrm{disp}_{\widehat{P}}^{(\rho)}(f',f)|\\ &\leq \frac{2k}{\rho}\mathfrak{R}_{n,P}(\Pi_\mathcal{H}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.\end{aligned}$$

Similarly

$$\begin{aligned}&\sup_{f,f'\in\mathcal{F}}|\mathrm{disp}_Q^{(\rho)}(f',f) - \mathrm{disp}_{\widehat{Q}}^{(\rho)}(f',f)|\\ &\leq \frac{2k}{\rho}\mathfrak{R}_{m,Q}(\Pi_\mathcal{H}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.\end{aligned}$$

Therefore, we conclude

$$\begin{aligned}&\sup_{f\in\mathcal{F}}|d^{(\rho)}_{f,\mathcal{F}}(\widehat{P},\widehat{Q}) - d^{(\rho)}_{f,\mathcal{F}}(P,Q)|\\ \leq&\sup_{f,f'\in\mathcal{F}}|\mathrm{disp}_Q^{(\rho)}(f',f) - \mathrm{disp}_{\widehat{Q}}^{(\rho)}(f',f)|\\ &+\sup_{f,f'\in\mathcal{F}}|\mathrm{disp}_P^{(\rho)}(f',f) - \mathrm{disp}_{\widehat{P}}^{(\rho)}(f',f)|\\ \leq&\frac{2k}{\rho}\mathfrak{R}_{n,P}(\Pi_\mathcal{H}\mathcal{F}) + \frac{2k}{\rho}\mathfrak{R}_{m,Q}(\Pi_\mathcal{H}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.\end{aligned}$$

□

Theorem C.8 (Theorem 3.7)**.** *For any $\delta > 0$, with probability $1 - 3\delta$, we have the following uniform generalization bound for all scoring functions f*

$$\begin{aligned}\mathrm{err}_Q(f) \leq & \mathrm{err}_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P}, \widehat{Q}) + \lambda \\ & + \frac{2k^2}{\rho}\mathfrak{R}_{n,P}(\Pi_1 \mathcal{F}) + \frac{2k}{\rho}\mathfrak{R}_{n,P}(\Pi_{\mathcal{H}}\mathcal{F}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\ & + \frac{2k}{\rho}\mathfrak{R}_{m,Q}(\Pi_{\mathcal{H}}\mathcal{F}) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.\end{aligned} \tag{22}$$

Proof. This is the result of combining Theorem C.2, Equation (18) and Theorem C.7. □

Example C.9 (Linear Classifiers)**.** *Let*

$$S \subseteq \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^s | \|\mathbf{x}\|_2 \leq r\}$$

be a sample of size m and suppose

$$\begin{aligned}\mathcal{F} = \big\{ f : \mathcal{X} \times \{\pm 1\} \to \mathbb{R} \;\big|\; f(\mathbf{x}, y) = \\ \mathrm{sgn}(y)\ \mathbf{w} \cdot \mathbf{x},\ \|\mathbf{w}\|_2 \leq \Lambda \big\}, \\ \mathcal{H} = \big\{ h \mid h(\mathbf{x}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}),\ \|\mathbf{w}\|_2 \leq \Lambda \big\}.\end{aligned}$$

Then the empirical Rademacher complexity of $\Pi_{\mathcal{H}}\mathcal{F}$ can be bounded as follows:

$$\widehat{\mathfrak{R}}_S(\Pi_{\mathcal{H}}\mathcal{F}) \leq 2\Lambda r\sqrt{\frac{d\log\frac{em}{d}}{m}},$$

where d is the VC-dimension of $\mathcal{H}$. If we further suppose

$$\min_{x \in S} |\mathbf{w} \cdot \mathbf{x}| = 1 \wedge \|w\|_2,$$

then

$$\widehat{\mathfrak{R}}_S(\Pi_{\mathcal{H}}\mathcal{F}) \leq 2\Lambda^2 r^2\sqrt{\frac{\log em}{m}}.$$

To prove this we need two lemmas.

Lemma C.10 (Propostion 6 of Maurer (2016))**.** *Let ξ_i be the Rademacher random variables. For any vector $\mathbf{v} \in \mathbb{R}^s$, the following holds:*

$$\|\mathbf{v}\|_2 \leq \sqrt{2}\mathop{\mathbb{E}}_{\xi_i \sim \{\pm 1\}, i \in \{1,2,\ldots,s\}} |\langle \boldsymbol{\xi}, \mathbf{v}\rangle|.$$

Lemma C.11 (Theorem 4.2 of Mohri et al. (2012))**.** *Let $S \subseteq \{\mathbf{x} : \|\mathbf{x}\| \leq r\}$. Then, the VC-dimension d of the set of canonical hyperplanes*

$$\big\{x \mapsto \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}) : \min_{x \in S} |\mathbf{w} \cdot \mathbf{x}| = 1 \wedge \|\mathbf{w}\|_2 \leq \Lambda\big\}$$

verifies

$$d \leq r^2\Lambda^2.$$

Now we present the proof of Example C.9.

Proof. By the definition of empirical Rademacher complexity and Cauchy-Schwartz inequality

$$\begin{aligned}m\widehat{\mathfrak{R}}_S(\Pi_{\mathcal{H}}\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{f,h} \sum_{i=1}^m \sigma_i f(\mathbf{x}_i, h(\mathbf{x}_i)) \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{w},h} \sum_{i=1}^m \sigma_i h(\mathbf{x}_i)\langle \mathbf{w}, \mathbf{x}_i\rangle \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{w},h} \langle \mathbf{w}, \sum_{i=1}^m \sigma_i h(\mathbf{x}_i)\mathbf{x}_i\rangle \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}} \sup_{h} \Lambda \| \sum_{i=1}^m \sigma_i h(\mathbf{x}_i)\mathbf{x}_i\|_2\end{aligned}$$

Applying Lemma C.10, we get

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\sigma}} \sup_h \Lambda \| \sum_{i=1}^m h(\mathbf{x}_i)\sigma_i\mathbf{x}_i\|_2 \\ \leq & \sqrt{2}\Lambda\mathbb{E}_{\boldsymbol{\sigma}} \sup_h \mathbb{E}_{\boldsymbol{\xi}\sim\{\pm 1\}^s} |\langle \boldsymbol{\xi}, \sum_{i=1}^m h(\mathbf{x_i})\sigma_i \boldsymbol{x_i}\rangle| \\ \leq & \sqrt{2}\Lambda\mathbb{E}_{\boldsymbol{\sigma}}\mathbb{E}_{\boldsymbol{\xi}} \sup_h |\langle \boldsymbol{\xi}, \sum_{i=1}^m h(\mathbf{x}_i)\sigma_i\mathbf{x}_i\rangle| \\ = & \sqrt{2}\Lambda\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \sup_h \langle \boldsymbol{\xi}, \sum_{i=1}^m h(\mathbf{x}_i)\sigma_i\mathbf{x}_i\rangle \\ = & \sqrt{2}\Lambda\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \sup_h \sum_{i=1}^m \sum_{j=1}^s \xi_j\sigma_i h(\mathbf{x}_i)x_{ij}.\end{aligned}$$

Let

$$A = \{(h(\mathbf{x}_1), \ldots, h(\mathbf{x}_m)) \mid h \in \mathcal{H}\}.$$

By Jensen's inequality, for any $t > 0$

$$\begin{aligned}&\exp(t\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \sup_h \sum_{i=1}^m \sum_{j=1}^s \xi_j\sigma_i h(\mathbf{x}_i)x_{ij}) \\ \leq & \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \exp(t \sup_h \sum_{i=1}^m \sum_{j=1}^s \xi_j\sigma_i h(\mathbf{x}_i)x_{ij}) \\ \leq & \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \sum_{\mathbf{a}\in A} \exp(t \sum_{i=1}^m \sum_{j=1}^s \xi_j\sigma_i a_i x_{ij}) \\ = & \sum_{\mathbf{a}\in A} \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \prod_{i=1}^m \prod_{j=1}^s \exp(t\xi_j\sigma_i a_i x_{ij}) \\ = & \sum_{\mathbf{a}\in A} \prod_{i=1}^m \prod_{j=1}^s \mathbb{E}_{\sigma_i,\xi_j} \exp(t\xi_j\sigma_i a_i x_{ij}) \\ \leq & \sum_{\mathbf{a}\in A} \prod_{i=1}^m \prod_{j=1}^s \exp(\frac{t^2(a_i x_{ij})^2}{2}) \\ = & \sum_{\mathbf{a}\in A} \exp(\frac{t^2\sum_{i=1}^m \|\mathbf{x}_i\|_2^2}{2}) \leq |A| \exp(\frac{t^2 r^2 m}{2}).\end{aligned}$$

Thus

$$\sup_h \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \sum_{i=1}^{m}\sum_{j=1}^{s} \xi_j \sigma_i h(\mathbf{x}_i) x_{ij} \leq \frac{\log|A|}{t} + \frac{tr^2 m}{2}.$$

Take

$$t = \sqrt{\frac{2\log|A|}{r^2 m}},$$

We get

$$\sup_h \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\xi}} \sum_{i=1}^{m}\sum_{j=1}^{s} \xi_j \sigma_i h(\mathbf{x}_i) x_{ij} \leq \sqrt{2r^2 m \log|A|}.$$

Note that

$$\log|A| = \log(\Pi(m,\mathcal{H})) \leq d\log\frac{\mathrm{e}m}{d},$$

We conclude

$$\widehat{\mathfrak{R}}_S(\Pi_{\mathcal{H}}\mathcal{F}) \leq 2\Lambda r\sqrt{\frac{d\log\frac{\mathrm{e}m}{d}}{m}}.$$

Now if the extra condition is satisfied, by Lemma C.11

$$\widehat{\mathfrak{R}}_S(\Pi_{\mathcal{H}}\mathcal{F}) \leq 2\Lambda^2 r^2\sqrt{\frac{\log \mathrm{e}m}{m}}.$$

□

Definition C.12 (Covering Number)**.** *Let (M,d) be a metric space. A subset $\widehat{T} \subseteq M$ is called an ϵ cover of $T \subseteq M$ if for every $t \in T$, there exists an $t' \in \widehat{T}$ such that $\rho(t,t') \leq \epsilon$. The covering number of T is the cardinality of the smallest ϵ cover of T, that is*

$$\mathcal{N}(\epsilon,T,d) \triangleq \min\left\{|\widehat{T}| \;\middle|\; \widehat{T} \text{ is an } \epsilon \text{ cover of } T\right\}. \quad (23)$$

Let $(\mathcal{F}_{x_1,\ldots,x_n}, \mathcal{L}_2(\widehat{D}))$ stand for the data-dependent $\mathcal{L}_2$ metric space given by metric

$$d(f,f') \triangleq \|f-f'\|_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-f'(x_i))^2} \quad (24)$$

where $x_1,\ldots,x_n$ are a sample from space $\mathcal{X}$ and $\mathcal{F}_{x_1,\ldots,x_n}$ stands for the restriction of (real-valued) function class $\mathcal{F}$ to that sample. Denote the $\mathcal{L}_2$ covering number by

$$\mathcal{N}_2(\epsilon,\mathcal{F}) \triangleq \mathcal{N}(\epsilon,\mathcal{F},\mathcal{L}_2(\widehat{D})). \quad (25)$$

The covering number can be interpreted as a measure of the richness of the class $\mathcal{F}$ at the scale ϵ. For a fixed value of ϵ, this covering number, and in particular how rapidly it grows with n, indicate how much the set $\mathcal{F}_{x_1,\ldots,x_n}$ "fills up" $\mathbb{R}^n$, when we examine it at the scale ϵ.

First we show that the $\mathcal{L}_2$ covering number of $\Pi_{\mathcal{H}}\mathcal{F}$ can be bounded by that of $\Pi_1\mathcal{F}$ and $\Pi_1\mathcal{H}$.

Lemma C.13. *Suppose the value of $f \in \Pi_1\mathcal{F}$ is bounded by $L < \infty$, i.e.*

$$\|f\|_2 \leq L. \quad (26)$$

Then we have

$$\mathcal{N}_2(\epsilon,\Pi_{\mathcal{H}}\mathcal{F}) \leq \mathcal{N}_2^k(\frac{\epsilon}{2k},\Pi_1\mathcal{F})\cdot\mathcal{N}_2^k(\frac{\epsilon}{2kL},\Pi_1\mathcal{H}) \quad (27)$$

Proof. For any $g \in \Pi_{\mathcal{H}}\mathcal{F}$, $g = \langle h,f\rangle$, choose $\widehat{h}_i, \widehat{f}_i$ from the $\mathcal{N}_2^k(\frac{\epsilon}{2kL},\Pi_1\mathcal{H})$ and $\mathcal{N}_2(\frac{\epsilon}{2k},\Pi_1\mathcal{F})$ cover of $\Pi_1\mathcal{H}$ and $\Pi_1\mathcal{F}$ according to the components h_i, f_i of h, f ($i = 1,\ldots k$). Let $\widehat{g} = \sum_{i=1}^k \widehat{h}_i\widehat{f}_i$. Then the choices of $\hat{g}$ is at most $\mathcal{N}_2^k(\frac{\epsilon}{2k},\Pi_1\mathcal{F})\cdot\mathcal{N}_2^k(\frac{\epsilon}{2kL},\Pi_1\mathcal{H})$. By Minkowski inequality and Hölder inequality we have

$$\begin{aligned}
\|g-\widehat{g}\|_2 &= \|\sum_{i=1}^{k}(h_i f_i - \widehat{h}_i\widehat{f}_i)\|_2 \\
&= \|\sum_{i=1}^{k}(h_i(f_i-\widehat{f}_i) + \widehat{f}_i(h_i-\widehat{h}_i)\|_2 \\
&\leq \sum_{i=1}^{k}(\|h_i\|_2\,\|f_i-\widehat{f}_i\|_2 + \|\widehat{f}_i\|_2\,\|h_i-\widehat{h}_i\|_2) \\
&\leq \sum_{i=1}^{k}(\|f_i-\widehat{f}_i\|_2 + L\|h_i-\widehat{h}_i\|_2) \leq \epsilon.
\end{aligned}$$

□

Lemma C.14 (Dudley's Entropy Bound, Talagrand (2014))**.** *For any function class $\mathcal{F}$ containing functions $f : \mathcal{X} \to \mathbb{R}$, we have*

$$\mathfrak{R}_{n,D}(\mathcal{F}) \leq \inf_{\epsilon\geq 0}\left\{4\epsilon + \frac{12}{\sqrt{n}}\int_{\epsilon}^{\sup_{f\in\mathcal{F}}\|f\|_2}\sqrt{\log\mathcal{N}_2(\tau,\mathcal{F})}\mathrm{d}\tau\right\} \quad (28)$$

Theorem C.15 (Theorem 3.8)**.** *With the same conditions in Theorem C.8, further suppose $\Pi_1\mathcal{F}$ is bounded in $\mathcal{L}_2$ by L. For $\delta > 0$, with probability $1-3\delta$, we have the following uniform generalization bound for all scoring functions f,*

$$\begin{aligned}
\mathrm{err}_Q(f) \leq\; & \mathrm{err}_{\widehat{P}}^{(\rho)}(f) + d_{f,\mathcal{F}}^{(\rho)}(\widehat{P},\widehat{Q}) + \lambda + 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \\
& + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \frac{16k^2\sqrt{k}}{\rho}\inf_{\epsilon\geq 0}\Big\{\epsilon + 3\big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\big) \\
& \Big(\int_{\epsilon}^{L}\sqrt{\log\mathcal{N}_2(\tau,\Pi_1\mathcal{F})}\mathrm{d}\tau + L\int_{\epsilon/L}^{1}\sqrt{\log\mathcal{N}_2(\tau,\Pi_1\mathcal{H})}\mathrm{d}\tau\Big)\Big\}
\end{aligned} \quad (29)$$

Proof. Directly combine Theorem C.7, Lemma C.13 and Lemma C.14 and put together similar terms noticing that

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b},$$

and the change of variable

$$\int_a^b f(x)\mathrm{d}x = \int_{a/t}^{b/t} tf(tx)dx.$$

□

Definition C.16 (Fat-Shattering Dimension)**.** *We say that $\mathcal{F}$ shatters $x_1, \ldots, x_n$ at scale γ, if there exists witness $s_1, \ldots, s_n$ such that, for every $\epsilon \in \{\pm 1\}^n$, there exists $f_\epsilon \in \mathcal{F}$ such that $\forall t \in \{1, \ldots, n\}$*

$$\epsilon_t \cdot (f_\epsilon(x_t) - s_t) \geq \gamma/2. \tag{30}$$

Then define the fat-shattering dimension

$$\begin{aligned}\mathrm{Fat}_\gamma(\mathcal{F}) \triangleq \max\{n \big| \exists\, x_1, \ldots, x_n \in \mathcal{X}\ s.t.\\ \mathcal{F}\ \gamma\text{-shatters } x_1, \ldots, x_n\}.\end{aligned} \tag{31}$$

Lemma C.17. *For any $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ and any $\gamma \in (0, 1)$*

$$\mathcal{N}_2(\gamma, \mathcal{F}) \leq \left(\frac{2}{\gamma}\right)^{K\ \mathrm{Fat}_{c\gamma}(\mathcal{F})} \tag{32}$$

where in the above c and K are universal constants.

Corollary C.18. *For any $\mathcal{F} \subseteq \{-1, 1\}^{\mathcal{X}}$,*

$$\mathrm{Fat}_{c\gamma}(\mathcal{F}) = \mathrm{Fat}_0(\mathcal{F}) = \mathrm{VC}(\mathcal{F}). \tag{33}$$

For detailed proofs of these two propositions, refer to Mendelson & Vershynin (2003); Rakhlin & Sridharan (2014).

Theorem C.19. *Let $\mathrm{Fat}_\gamma(\Pi_1\mathcal{F})$ be the fat-shattering dimension of $\Pi_1\mathcal{F}$ with scale γ and $\mathrm{VC}(\Pi_1\mathcal{H})$ be the VC-dimension of $\Pi_1\mathcal{H}$. Then there exist constants $C_1, C_2, C_3 > 0, 0 < c < 1$ independent of n, m and $\mathcal{F}, \mathcal{H}$ such that for $\delta > 0$, with probability $1 - 3\delta$,*

$$\begin{aligned}&\mathrm{err}_Q(f) \leq \mathrm{err}^{(\rho)}_{\widehat{P}}(f) + d^{(\rho)}_{f,\mathcal{F}}(\widehat{P}, \widehat{Q}) + \lambda\\ &+ \frac{k^2\sqrt{k}}{\rho} C_1 L\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\sqrt{\mathrm{VC}(\Pi_1\mathcal{H})}\\ &+ \frac{k^2\sqrt{k}}{\rho} \inf_{\epsilon \geq 0}\Big\{C_2\epsilon + C_3\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\Big(\int_\epsilon^L \sqrt{\mathrm{Fat}_{c\tau}(\Pi_1\mathcal{F})\log\frac{2}{\tau}}\mathrm{d}\tau\Big)\Big\}\\ &+ 2\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}.\end{aligned} \tag{34}$$

This results from a direct computation after putting Lemma C.17 and Corollary C.18 into Theorem C.15.

D. Analysis of Algorithm

In this section, we show that γ controls the margin ρ and that the minimization of this loss will lead to consistency between source and target domain using similar methods with Goodfellow et al. (2014).

Proposition D.1. *Consider the optimization problem we have defined*

$$\max_{f'}\ \gamma\mathbb{E}_{\widehat{P}}\log(\sigma_{h_f} \circ f') + \mathbb{E}_{\widehat{Q}}\log(1 - \sigma_{h_f} \circ f'). \tag{35}$$

Assume that there is no restriction for the choice of f' and $\gamma > 0$. Fixing a single-output classifier h_f, we have the following two results:

1. *The optimal value of $\sigma_{h_f} \circ f'$ on data x is*

$$\frac{\gamma p(x)}{\gamma p(x) + q(x)} \tag{36}$$

where $p(x)$ and $q(x)$ are the density functions.

2. *Problem (35) is equivalent to a γ-balanced Jensen-Shannon Divergence:*

$$\gamma\mathrm{KL}\Big(P\Big\|\frac{\gamma P + Q}{\gamma + 1}\Big) + \mathrm{KL}\Big(Q\Big\|\frac{\gamma P + Q}{\gamma + 1}\Big) \tag{37}$$

and has the global minimum at $P = Q$.

Proof.

$$\begin{aligned}&\gamma\mathbb{E}_P\log(\sigma_{h_f} \circ f') + \mathbb{E}_Q\log(1 - \sigma_{h_f} \circ f')\\ &= \int_{x\in\mathcal{X}} \gamma p(x)\log(\sigma_{h_f} \circ f') + q(x)\log(1 - \sigma_{h_f} \circ f')\,\mathrm{d}x\\ &= \int_{x\in\mathcal{X}} \mathrm{J}(x, \sigma_{h_f} \circ f')\,\mathrm{d}x.\end{aligned}$$

For there is no restriction on f', we could find the f' that reaches the maximum on each $x \in \mathcal{X}$. Simple calculus gives the result that for $\forall x \in \mathcal{X}$, $\mathrm{J}(x, \sigma_{h_f} \circ f')$ is the largest when

$$\sigma_{h_f} \circ f' = \frac{\gamma p(x)}{\gamma p(x) + q(x)}.$$

So the first conclusion is proved. At this time,

$$\begin{aligned}&\mathrm{J}(x, \sigma_{h_f} \circ f')\\ =&\gamma p(x)\log(\sigma_{h_f} \circ f') + q(x)\log(1 - \sigma_{h_f} \circ f')\\ =&\gamma p(x)\log\Big(\frac{\gamma p(x)}{\gamma p(x) + q(x)}\Big) + q(x)\log\Big(\frac{q(x)}{\gamma p(x) + q(x)}\Big)\\ =&\gamma p(x)\log\Big(\frac{p(x)}{\frac{\gamma p(x) + q(x)}{\gamma + 1}}\Big) + q(x)\log\Big(\frac{q(x)}{\frac{\gamma p(x) + q(x)}{\gamma + 1}}\Big)\\ &+ \gamma\log\gamma\, p(x) - (\gamma p(x) + q(x))\log(\gamma + 1).\end{aligned}$$

Notice that $\frac{\gamma p(x) + q(x)}{\gamma + 1}$ is density of mixed distribution

$$\frac{\gamma P + Q}{\gamma + 1}.$$

Integrate $\mathrm{J}(x, \sigma_{h_f} \circ f')$ on the $\mathcal{X}$,

$$
\begin{aligned}
&\int_{x\in\mathcal{X}} \mathrm{J}(x, \sigma_{h_f} \circ f') \,\mathrm{d}x \\
=&\int_{x\in\mathcal{X}} [\gamma p(x) \log(\frac{p(x)}{\frac{\gamma p(x)+q(x)}{\gamma+1}}) + q(x) \log(\frac{q(x)}{\frac{\gamma p(x)+q(x)}{\gamma+1}}) \,\mathrm{d}x \\
&+ \gamma \log\gamma\, p(x) - (\gamma p(x) + q(x)) \log(\gamma+1)] \\
=&\gamma \mathrm{KL}(P\|\frac{\gamma P+Q}{\gamma+1}) + \mathrm{KL}(Q\|\frac{\gamma P+Q}{\gamma+1}) \\
&+ \gamma\log\gamma - (\gamma+1)\log(\gamma+1)] \\
=&\gamma \mathrm{KL}(P\|\frac{\gamma P+Q}{\gamma+1}) + \mathrm{KL}(Q\|\frac{\gamma P+Q}{\gamma+1}) + C(\gamma),
\end{aligned}
$$

where $C(\gamma)$ is a constant only depending on γ. This derivation shows that the second conclusion holds. □

This proposition implies that different choices of γ does not lead to mismatch between P and Q.

Next we show that γ decides the margin ρ at equilibrium by analyzing the training process, which ensures the optimality of the margin disparity discrepancy that we achieved after training.

During the training session, the discrepancy between source and target features decreases and converges to a value close to zero, indicating $\psi(P) \approx \psi(Q)$. If $\gamma = 1$, the value of $\sigma_{h_f} \circ f'$ converges to a number around $\frac{1}{2}$ on both the source and target domains, in which case the output of f' for the class predicted by f is probably the largest among all classes. However, the margin of f' might still be close to zero as there might exist a prediction for another class approaching $\frac{1}{2}$ from below. As a result, the value of the margin disparity discrepancy measured by f' does not reach minimization for any ρ. For $\gamma > 1$, after some calculation, we conclude that the value of $\sigma_{h_f} \circ f'$ will reach $\frac{\gamma}{\gamma+1}$ and the margin of f' will be around $\log\gamma$ at equilibrium as shown in the proposition below.

Proposition D.2. *For any $j \in \{1, 2, \ldots, k\}$ if $\sigma_j \circ f > \mu > \frac{1}{2}$, then f is a classifier with margin $\log\frac{\mu}{1-\mu}$.*

Proof. For any $r \neq j$, $r \in 1, \ldots, k$

$$
\begin{aligned}
\mu &< \frac{\sigma_j}{\sum_{i=1}^k \sigma_i} \\
&\leq \frac{\sigma_j}{\sigma_j + \sigma_r} \\
&= \frac{1}{1 + e^{f(x,r) - f(x,j)}}.
\end{aligned}
$$

Thus

$$
f(x, j) - f(x, r) > \log(\frac{\mu}{1-\mu}).
$$

□

E. Additional Experiments

We also test our algorithm by minimizing the original MDD loss. Since gradient saturation using the margin losses is fatal in the early training stage, we implement by switching to the margin losses after 2000 steps. The results on Office-31 with the margin $\rho = \log 4$ (equivalent to $\gamma = 4$) are reported in Table 1.

Table 1. Accuracy (%) on Office-31 with original MDD loss.

Task	Accuracy
A → W	94.1
A → D	91.8
D → W	100
W → D	98.2
D → A	73.7
W → A	71.7
Average	88.3

References

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Galbis, A. and Maestre, M. *Vector analysis versus vector calculus*. Springer Science & Business Media, 2012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17, 2016.

Mendelson, S. and Vershynin, R. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152(1): 37–55, 2003.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. Foundations of machine learning. 2012.

Rakhlin, A. and Sridharan, K. Statistical learning and sequential prediction. *Book Draft*, 2014.

Talagrand, M. *Upper and lower bounds for stochastic processes: modern methods and classical problems*, volume 60. Springer Science & Business Media, 2014.