Label-Noise Robust Domain Adaptation
Supplementary Material

Xiyu Yu [1] **Tongliang Liu** [2] **Mingming Gong** [3] **Kun Zhang** [4] **Kayhan Batmanghelich** [5] **Dacheng Tao** [2]

Abstract

In this supplementary material, the detailed proofs of theorems in this paper are present.

1. Proof of Porposition 1

Proof. Denote $\pi_{ij} = P^S(Y = j|\hat{Y} = i), \forall i, j \in \{1, 2\}$. Note that as in many label noise methods, we assume $\pi_{11}\pi_{22} - \pi_{12}\pi_{21} \neq 0$. Since $Y \to \hat{Y}$ and $Y \to X$, then X and $\hat{Y}$ are conditional independent given Y. Then, it is easy to derive that

$$\begin{aligned} P^S_{\rho X|Y=1} &= \pi_{11}P^S_{X|Y=1} + \pi_{12}P^S_{X|Y=2} \\ P^S_{\rho X|Y=2} &= \pi_{21}P^S_{X|Y=1} + \pi_{22}P^S_{X|Y=2}. \end{aligned} \tag{A.1}$$

Since $P^T_X = \omega_{\rho 1}P^S_{\rho X|Y=1} + \omega_{\rho 2}P^S_{\rho X|Y=2}$, we have

$$P^T_X = (\omega_{\rho 1}\pi_{11} + \omega_{\rho 2}\pi_{21})P^S_{X|Y=1} + (\omega_{\rho 1}\pi_{12} + \omega_{\rho 2}\pi_{22})P^S_{X|Y=2}. \tag{A.2}$$

We also have $P^T_X = \omega_1 P^S_{X|Y=1} + \omega_2 P^S_{X|Y=2}$. Since $P^S_{X|Y=1}$ and $P^S_{X|Y=2}$ is different, according to Theorem 1 in (Yu et al., 2018), we have

$$\begin{aligned} \omega_1 &= \omega_{\rho 1}\pi_{11} + \omega_{\rho 2}\pi_{21}; \\ \omega_2 &= \omega_{\rho 1}\pi_{12} + \omega_{\rho 2}\pi_{22}. \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} \omega_{\rho 1} &= (\pi_{22}\omega_1 - \pi_{21}\omega_2)/(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}); \\ \omega_{\rho 2} &= (\pi_{11}\omega_2 - \pi_{12}\omega_1)/(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}). \end{aligned} \tag{A.4}$$

Let $\omega_i = \omega_{\rho i}, i = 1, 2$, we have

$$\begin{aligned} \omega_1 &= (\pi_{22}\omega_1 - \pi_{21}\omega_2)/(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}); \\ \omega_2 &= (\pi_{11}\omega_2 - \pi_{12}\omega_1)/(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}). \end{aligned} \tag{A.5}$$

Then, we have

$$\begin{aligned} (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})\omega_1 &= \pi_{22}\omega_1 - \pi_{21}\omega_2; \\ (\pi_{11}\pi_{22} - \pi_{12}\pi_{21})\omega_2 &= \pi_{11}\omega_2 - \pi_{12}\omega_1. \end{aligned} \tag{A.6}$$

Then,

$$\begin{aligned} (\pi_{22} - \pi_{11}\pi_{22} + \pi_{12}\pi_{21})\omega_1 &= \pi_{21}\omega_2; \\ (\pi_{11} - \pi_{11}\pi_{22} + \pi_{12}\pi_{21})\omega_2 &= \pi_{12}\omega_1. \end{aligned} \tag{A.7}$$

Then, we have

$$\pi_{12}\omega_1 = \pi_{21}\omega_2. \tag{A.8}$$

According to $\omega_1 + \omega_2 = 1$, we have $\omega_2 = \pi_{12}/(\pi_{12} + \pi_{21})$ and $\omega_1 = \pi_{21}/(\pi_{12} + \pi_{21})$.

□

2. Proof of Theorem 1

Proof. In this proof, $Y = y$ (resp. $\hat{Y} = y'$) is replaced by y (resp. y') for simplicity. For example, we let $P^S(\hat{Y} = y'|Y = y) = P^S(y'|y)$. We also let $X' = \tau(X)$. According to Eq. (2) in the main paper, we have

$$P^{\text{new}}_{X'} = \sum_y \sum_{y'} \beta_\rho(y') P^S(y'|X', y) P^S(X', y) = \sum_y P^S(X'|y) P^S(y) \sum_{y'} P^S(y'|y)\beta_\rho(y'). \tag{A.9}$$

Because $P^T_{X'} = \sum_y P^T(X'|y)P^T(y)$, then combining with the above equation, we have

$$\sum_y P^T(X'|y)P^T(y) = \sum_y P^S(X'|y)P^S(y) \sum_{y'} P^S(y'|y)\beta_\rho(y'). \tag{A.10}$$

Because the transformation τ satisfies that $P(X'|Y = i), i \in \{1, \cdots, c\}$ are linearly independent, there exist **no** such non-zero $\gamma_1, \cdots, \gamma_c$ and $\kappa_1, \cdots, \kappa_c$ that $\sum_{i=1}^c \gamma_i P^S(X'|Y = i) = 0$ and $\sum_{i=1}^c \kappa_i P^T(X'|Y = i) = 0$. According to the assumption in Theorem 1, the elements in the set $\{v_i P^S(X'|Y = i) + \lambda_i P^T(X'|Y = i); i \in \{1, \cdots, c\}; \forall v_i, \lambda_i\ (v_i^2 + \lambda_i^2 \neq 0)\}$ are also linearly independent. Then we have, $\forall y \in \{1, \cdots, c\}$,

$$P^T(X'|y)P^T(y) - P^S(X'|y)P^S(y) \sum_{y'} P^S(y'|y)\beta_\rho(y') = 0. \tag{A.11}$$

Taking the integral of above equation w.r.t. X', we have

$$P^T(y) = P^S(y) \sum_{y'} P^S(y'|y)\beta_\rho(y'), \tag{A.12}$$

which further implies $P^T(X'|y) = P^S(X'|y), \forall y \in \{1, \cdots, c\}$. According to Eq. (A.12), we have $\forall y \in \{1, \cdots, c\}$,

$$\sum_{y'} P^S(y'|y)\beta_\rho(y') = P^T(y)/P^S(y) = \beta(y).$$

The proof of Theorem 1 ends. ■

3. Proof of Theorem 2

Recall the denoising MMD loss, we have

$$\hat{\mathcal{D}}(W, \alpha) = \|\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\|^2.$$

Let

$$\mathcal{D}(W, \alpha) = \|\mathbb{E}\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \mathbb{E}\frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\|^2,$$

where we abuse the training samples $\{(x_1^S, \hat{y}_1^S), \cdots, (x_m^S, \hat{y}_m^S)\}$ and $\{x_1^T, \cdots, x_n^T\}$ as being i.i.d. variables, respectively.

We analyze the convergence property of the learned $\hat{\alpha}$ to the optimal one α^* by analyzing the convergence from the expected objective function $\mathcal{D}(\hat{W}, \hat{\alpha})$ to $\mathcal{D}(\hat{W}, \alpha^*)$.

To prove Theorem 2, we need the following Theorem A.1, Lemma A.1, and Lemma A.2. Theorem A.1 is about concentration inequality (McDiarmid's inequality (Boucheron et al., 2013), also known as the bounded difference inequality). Lemma A.1 shows that the distance $\mathcal{D}(\hat{W}, \hat{\alpha}) - \mathcal{D}(\hat{W}, \alpha^*)$ can be upper bounded even though we do not know the optimal α^*. Lemma A.2 upper bounds the Rademacher-like (Bartlett & Mendelson, 2002) term $\mathbb{E}\sup_{\alpha \in \Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|^2$.

Theorem A.1. *Let $X = [X_1, \cdots, X_n]$ be an independent and identically distributed sample and X^i a new sample with the i-th example in X being replaced by an independent example X_i'. If there exists $b_1, \cdots, b_n > 0$ such that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the following conditions*

$$|f(X) - f(X^i)| \leq b_i, \forall i \in \{1, \cdots, n\}.$$

Then for any $X \in \mathcal{X}^n$ and $\epsilon > 0$, the following inequality holds

$$P(\mathbb{E}f(X) - f(X) \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n b_i^2}\right).$$

Lemma A.1. *We denote* $\Delta \triangleq \{\alpha | \alpha \geq 0, \|\alpha\|_1 = 1\}$ *and*

$$f(\mathbf{x}^S, \mathbf{x}^T, \alpha) \triangleq \mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}. \tag{A.13}$$

Then, we have

$$\begin{aligned} \mathcal{D}(\hat{W}, \hat{\alpha}) - \mathcal{D}(\hat{W}, \alpha^*) &\leq 2 \sup_{\alpha \in \Delta} |\mathcal{D}(\hat{W}, \alpha) - \hat{\mathcal{D}}(\hat{W}, \alpha)| \\ &\leq 4(\wedge_{\hat{Q}} + 1) \wedge_{\hat{W}} \sup_{\alpha \in \Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|. \end{aligned} \tag{A.14}$$

Proof. We have

$$\begin{aligned} &\mathcal{D}(\hat{W}, \hat{\alpha}) - \mathcal{D}(\hat{W}, \alpha^*) \\ &= \mathcal{D}(\hat{W}, \hat{\alpha}) - \hat{\mathcal{D}}(\hat{W}, \hat{\alpha}) + \hat{\mathcal{D}}(\hat{W}, \hat{\alpha}) - \hat{\mathcal{D}}(\hat{W}, \alpha^*) + \hat{\mathcal{D}}(\hat{W}, \alpha^*) - \mathcal{D}(\hat{W}, \alpha^*) \\ &\leq \mathcal{D}(\hat{W}, \hat{\alpha}) - \hat{\mathcal{D}}(\hat{W}, \hat{\alpha}) + \hat{\mathcal{D}}(\hat{W}, \alpha^*) - \mathcal{D}(\hat{W}, \alpha^*) \\ &\leq 2 \sup_{\alpha \in \Delta} |\mathcal{D}(\hat{W}, \alpha) - \hat{\mathcal{D}}(\hat{W}, \alpha)|, \end{aligned} \tag{A.15}$$

where the first inequality holds because $\hat{\alpha}$ is the empirical minimizer of $\hat{\mathcal{D}}(\hat{W}, \alpha)$ and thus $\hat{\mathcal{D}}(\hat{W}, \hat{\alpha}) \leq \hat{\mathcal{D}}(\hat{W}, \alpha^*)$.

Further, we have

$$\begin{aligned} &|\mathcal{D}(\hat{W}, \alpha) - \hat{\mathcal{D}}(\hat{W}, \alpha)| \\ &= \left(\mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) + \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right)^\top \\ &\left(\mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) \\ &\leq \left\|\mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) + \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right\| \\ &\left\|\mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right\| \\ &\leq 2(\wedge_{\hat{Q}} + 1) \wedge_{\hat{W}} \left\|\mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right\|, \end{aligned} \tag{A.16}$$

where the first inequality holds because of Cauchy-Schwarz inequality.

Since

$$f(\mathbf{x}^S, \mathbf{x}^T, \alpha) \triangleq \mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}, \tag{A.17}$$

we have

$$2 \sup_{\alpha \in \Delta} |\mathcal{D}(\hat{W}, \alpha) - \hat{\mathcal{D}}(\hat{W}, \alpha)| \leq 4(\wedge_{\hat{Q}} + 1) \wedge_{\hat{W}} \sup_{\alpha \in \Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|. \tag{A.18}$$

The proof ends. ■

Lemma A.2. *Given learned* $\hat{Q}$ *and* $\hat{W}$*, let the induced RKHS be universal and upper bounded that* $\|\psi(\tau(x))\| \leq \wedge_{\hat{W}}$ *for all* x *in the source and target domains. Let the entries of* G *be bounded that* $|G_{ij}| \leq \wedge_{\hat{Q}}$ *for all* $i \in \{1, \cdots, m\}, j \in \{1, \cdots, c\}$*. We have*

$$\mathbb{E} \sup_{\alpha \in \Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|^2 \leq 4(\wedge_{\hat{Q}} + 1)^2 \wedge_{\hat{W}}^2 \sqrt{c}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right).$$

Proof. Recall that when $\hat{y}_k = i, \forall k \in \{1, \cdots, m\}$, the k-th row of $G \in \mathbb{R}^{m \times c}$ is $[\frac{\hat{Q}_{i1}^{-1}}{\hat{P}^S(Y=1)}, \cdots, \frac{\hat{Q}_{ic}^{-1}}{\hat{P}^S(Y=c)}]$. Given $\hat{Q}$, $\hat{W}$ and the estimated $\hat{P}^S(Y)$, we assumed that the entries of G is bounded, i.e., $|G_{ij}| \leq \wedge_{\hat{Q}}$, and that RKHS is upper bounded, i.e., $-\psi_{\max} \le \psi(\tau(x)) \le \psi_{\max}$ and $\|\psi_{\max}\| \le \wedge_{\hat{W}}$. Because $\alpha \ge 0$ and $\|\alpha\|_1 = 1$, we can conclude that for any training sample in the source domain, we have

$$\|\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha\| \le \wedge_{\hat{W}} \wedge_{\hat{Q}} .$$

We then have $\|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\| \le 2(\wedge_{\hat{Q}} + 1)\wedge_{\hat{W}}$ and that

$$\|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|^2 \le 2(\wedge_{\hat{Q}} + 1) \wedge_{\hat{W}} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|.$$

Accordingly, we have

$$\mathbb{E} \sup_{\alpha\in\Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|^2 \le 2(\wedge_{\hat{Q}} + 1) \wedge_{\hat{W}} \mathbb{E} \sup_{\alpha\in\Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\|. \tag{A.19}$$

Furthermore, let $\tilde{\mathbf{x}}^S$ and $\tilde{\mathbf{x}}^T$ be i.i.d. copies of $\mathbf{x}^S$ and $\mathbf{x}^T$, respectively. In the literature, $\tilde{\mathbf{x}}^S$ and $\tilde{\mathbf{x}}^T$ are referred as ghost samples (Mohri et al., 2012). We have

$$\begin{aligned}
&\mathbb{E} \sup_{\alpha\in\Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\| \\
&= \mathbb{E} \sup_{\alpha\in\Delta} \left\| \mathbb{E}\left(\frac{1}{m}\psi(\mathbf{x'}^S)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1} \right\| \\
&= \mathbb{E}_{\mathbf{x}^S,\mathbf{x}^T} \sup_{\alpha\in\Delta} \left\| \mathbb{E}_{\tilde{\mathbf{x}}^S,\tilde{\mathbf{x}}^T}\left(\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1} \right\| \\
&\le \mathbb{E}_{\mathbf{x}^S,\mathbf{x}^T,\tilde{\mathbf{x}}^S,\tilde{\mathbf{x}}^T} \sup_{\alpha\in\Delta} \left\| \left(\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1} \right\|,
\end{aligned}$$

where the last inequality holds because of Jensen's inequality and that every norm is a convex function.

Since $\tilde{\mathbf{x}}^S$ and $\tilde{\mathbf{x}}^T$ be i.i.d. copies of $\mathbf{x}^S$ and $\mathbf{x}^T$, respectively, the random variable $\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T)\mathbf{1} - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}$ is a symmetric random variable, which means its density function is even. Let σ_i be independent Rademacher variables, which are uniformly distributed from $\{-1, 1\}$. Let

$$\psi(\mathbf{x'}^S, \sigma) \triangleq [\sigma_1\psi(x'^S_1), \cdots, \sigma_m\psi(x'^S_m)]^\top;$$

and

$$\psi(\mathbf{x'}^T, \sigma) \triangleq [\sigma_1\psi(x'^T_1), \cdots, \sigma_n\psi(x'^T_n)]^\top.$$

We have that the random variable $\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T)\mathbf{1} - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1}$ and the random variable $\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S, \sigma)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T, \sigma)\mathbf{1} - \frac{1}{m}\psi(\mathbf{x'}^S, \sigma)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T, \sigma)\mathbf{1}$ have the same distribution.

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}^S,\mathbf{x}^T,\tilde{\mathbf{x}}^S,\tilde{\mathbf{x}}^T} \sup_{\alpha\in\Delta} \left\| \left(\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T)\mathbf{1} \right\| \\
&= \mathbb{E}_{\mathbf{x}^S,\mathbf{x}^T,\tilde{\mathbf{x}}^S,\tilde{\mathbf{x}}^T,\sigma} \sup_{\alpha\in\Delta} \left\| \left(\frac{1}{m}\psi(\tilde{\mathbf{x}}'^S, \sigma)G\alpha - \frac{1}{n}\psi(\tilde{\mathbf{x}}'^T, \sigma)\mathbf{1}\right) - \frac{1}{m}\psi(\mathbf{x'}^S, \sigma)G\alpha + \frac{1}{n}\psi(\mathbf{x'}^T, \sigma)\mathbf{1} \right\| \\
&\le 2\mathbb{E}_{\mathbf{x}^S,\mathbf{x}^T,\sigma} \sup_{\alpha\in\Delta} \left\| \left(\frac{1}{m}\psi(\mathbf{x'}^S, \sigma)G\alpha - \frac{1}{n}\psi(\mathbf{x'}^T, \sigma)\mathbf{1}\right) \right\| \\
&\le 2\mathbb{E}_{\mathbf{x}^S,\sigma} \sup_{\alpha\in\Delta} \left\| \frac{1}{m}\psi(\mathbf{x'}^S, \sigma)G\alpha \right\| + 2\mathbb{E}_{\mathbf{x}^T,\sigma} \sup_{\alpha\in\Delta} \left\| \frac{1}{n}\psi(\mathbf{x'}^T, \sigma)\mathbf{1} \right\|,
\end{aligned}$$

where the inequalities hold because of the triangle inequality.

We then upper bound $\mathbb{E}_{\mathbf{x}^S,\sigma}\sup_{\alpha\in\Delta}\left\|\frac{1}{m}\psi(\mathbf{x}'^S,\sigma)G\alpha\right\|$ and $\mathbb{E}_{\mathbf{x}^T,\sigma}\left\|\frac{1}{n}\psi(\mathbf{x}'^T,\sigma)\mathbf{1}\right\|$, respectively. For example, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}^S,\sigma}\sup_{\alpha\in\Delta}\left\|\frac{1}{m}\psi(\mathbf{x}'^S,\sigma)G\alpha\right\| \\
&=\mathbb{E}_{\mathbf{x}^S,\sigma}\sup_{\alpha\in\Delta}\left\|\frac{1}{m}\left\langle G^\top[\sigma_1\psi(x_1'^S),\cdots,\sigma_m\psi(x_m'^S)]^\top,\alpha\right\rangle\right\| \\
&\leq \mathbb{E}_{\mathbf{x}^S,\sigma}\sup_{\alpha\in\Delta}\frac{1}{m}\|G^\top[\sigma_1\psi(x_1'^S),\cdots,\sigma_m\psi(x_m'^S)]^\top\|\|\alpha\| \\
&\leq \mathbb{E}_{\mathbf{x}^S,\sigma}\sup_{\alpha\in\Delta}\frac{1}{m}\|G^\top[\sigma_1\psi(x_1'^S),\cdots,\sigma_m\psi(x_m'^S)]^\top\|\|\alpha\|_1 \\
&\leq \mathbb{E}_{\mathbf{x}^S,\sigma}\frac{1}{m}\|G^\top[\sigma_1\psi(x_1'^S),\cdots,\sigma_m\psi(x_m'^S)]^\top\| \\
&\leq \frac{\wedge_{\hat{Q}}\wedge_{\hat{W}}}{m}\mathbb{E}_\sigma\sqrt{c(\sum_{i=1}^m\sigma_i)^2} \\
&\leq \frac{\wedge_{\hat{Q}}\wedge_{\hat{W}}}{m}\sqrt{c\mathbb{E}_\sigma(\sum_{i=1}^m\sigma_i)^2} \\
&=\frac{\wedge_{\hat{Q}}\wedge_{\hat{W}}\sqrt{c}}{\sqrt{m}},
\end{aligned}
$$

where $G\in\mathbb{R}^{m\times c}$, c is the number of classes. The first inequality holds because of Cauchy-Schwarz inequality. The second inequality holds because $\|\alpha\|\leq\|\alpha\|_1$. The fourth inequality holds because of the Talagrand Contraction Lemma (Ledoux & Talagrand, 2013). And the last inequality holds because of the Jensen's inequality and that the function sqrt is a concave function. Similarly, we can prove that

$$
\mathbb{E}_{\mathbf{x}^T,\sigma}\left\|\frac{1}{n}\psi(\mathbf{x}'^T,\sigma)\mathbf{1}\right\|\leq\frac{\wedge_{\hat{W}}}{\sqrt{n}}. \tag{A.20}
$$

Combining Eq. (A.19), Eq. (A.20), Eq. (A.20), Eq. (A.20), and Eq. (A.20), we have

$$
\begin{aligned}
&\mathbb{E}\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2 \\
&\leq 4(\wedge_{\hat{Q}}+1)\wedge_{\hat{W}}\left(\frac{\wedge_{\hat{Q}}\wedge_{\hat{W}}\sqrt{c}}{\sqrt{m}}+\frac{\wedge_{\hat{W}}}{\sqrt{n}}\right) \\
&\leq 4(\wedge_{\hat{Q}}+1)^2\wedge_{\hat{W}}^2\sqrt{c}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right).
\end{aligned} \tag{A.21}
$$

The proof of Lemma A.2 ends. ■

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. According to Lemma A.1, we have

$$
\begin{aligned}
\mathcal{D}(\hat{W},\hat{\alpha})-\mathcal{D}(\hat{W},\alpha^*)&\leq 2\sup_{\alpha\in\Delta}|\mathcal{D}(\hat{W},\alpha)-\hat{\mathcal{D}}(\hat{W},\alpha)| \\
&\leq 4(\wedge_{\hat{Q}}+1)\wedge_{\hat{W}}\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|.
\end{aligned} \tag{A.22}
$$

Since $\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|\geq 0$, it holds that

$$
\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|=\sqrt{\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2}. \tag{A.23}
$$

Then, we will employ McDiarmid's inequality to upper bound the defect $\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2$. We now check its bounded difference property.

Let $\mathbf{x}^{Si}$ be a new sample in the source domain with the i-th example in $\mathbf{x}^S$ being replaced by an independent example $\tilde{x}_i^S$, where $i \in \{1, \cdots, m\}$, and $\mathbf{x}^{Ti}$ be a new sample in the target domain with the i-th example in $\mathbf{x}^T$ being replaced by an independent example $\tilde{x}_i^T$, where $i \in \{1, \cdots, n\}$.

For any $i \in \{1, \cdots, m\}$, we have

$$\begin{aligned}
&\left|\sup_{\alpha\in\Delta}\|f(\mathbf{x}^{Si},\mathbf{x}^T,\alpha)\|^2 - \sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2\right| \\
&\le \sup_{\alpha\in\Delta}\left|(f(\mathbf{x}^{Si},\mathbf{x}^T,\alpha)+f(\mathbf{x}^S,\mathbf{x}^T,\alpha))^\top\left(f(\mathbf{x}^{Si},\mathbf{x}^T,\alpha)-f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\right)\right| \\
&\le \sup_{\alpha\in\Delta}\left|4(\wedge_{\hat{Q}}+1)\psi_{\max}^\top\left(f(\mathbf{x}^{Si},\mathbf{x}^T,\alpha)-f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\right)\right| \\
&= \sup_{\alpha\in\Delta}\left|4(\wedge_{\hat{Q}}+1)\psi_{\max}^\top\left(\frac{1}{m}\psi({\mathbf{x}'}^{Si})G\alpha-\frac{1}{m}\psi({\mathbf{x}'}^{S})G\alpha\right)\right| \\
&\le \frac{8\wedge_{\hat{Q}}(\wedge_{\hat{Q}}+1)}{m}|\psi_{\max}|^\top|\psi_{\max}| \\
&\le \frac{8(\wedge_{\hat{Q}}+1)^2\wedge_{\hat{W}}^2}{m}.
\end{aligned} \tag{A.24}$$

Similarly, for any $i \in \{1, \cdots, n\}$, we have

$$\begin{aligned}
&\left|\sup_{\alpha\in\Delta}\|f(\mathbf{x}^{S},\mathbf{x}^{Ti},\alpha)\|^2 - \sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2\right| \\
&\le \sup_{\alpha\in\Delta}\left|\left(f(\mathbf{x}^{S},\mathbf{x}^{Ti},\alpha)+f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\right)^\top\left(f(\mathbf{x}^{S},\mathbf{x}^{Ti},\alpha)-f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\right)\right| \\
&\le \sup_{\alpha\in\Delta}\left|4(\wedge_{\hat{Q}}+1)\psi_{\max}^\top\left(f(\mathbf{x}^{S},\mathbf{x}^{Ti},\alpha)-f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\right)\right| \\
&= \sup_{\alpha\in\Delta}\left|4(\wedge_{\hat{Q}}+1)\psi_{\max}^\top\left(\frac{1}{n}\psi({\mathbf{x}'}^{Ti})\mathbf{1}-\frac{1}{n}\psi({\mathbf{x}'}^{T})\mathbf{1}\right)\right| \\
&\le \frac{8(\wedge_{\hat{Q}}+1)}{n}|\psi_{\max}|^\top|\psi_{\max}| \\
&\le \frac{8(\wedge_{\hat{Q}}+1)\wedge_{\hat{W}}^2}{n}.
\end{aligned} \tag{A.25}$$

Employing McDiarmid's inequality, we have that

$$\begin{aligned}
&P(\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2 - \mathbb{E}_{\mathbf{x}^S,\mathbf{x}^T}\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2 \ge \epsilon) \\
&\le \exp\left(\frac{-\epsilon^2}{32(\wedge_{\hat{Q}}+1)^4\wedge_{\hat{W}}^4(\frac{1}{m}+\frac{1}{n})}\right).
\end{aligned} \tag{A.26}$$

Let

$$\delta = \exp\left(\frac{-\epsilon^2}{32(\wedge_{\hat{Q}}+1)^4\wedge_{\hat{W}}^4(\frac{1}{m}+\frac{1}{n})}\right).$$

For any $\delta > 0$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\| \\
&\le \sqrt{\mathbb{E}\sup_{\alpha\in\Delta}\|f(\mathbf{x}^S,\mathbf{x}^T,\alpha)\|^2 + 8(\wedge_{\hat{Q}}+1)^2\wedge_{\hat{W}}^2\sqrt{\frac{1}{2}(\frac{1}{m}+\frac{1}{n})\log\frac{1}{\delta}}}. \\
&\le (\wedge_{\hat{Q}}+1)\wedge_{\hat{W}}\sqrt{4\sqrt{c}(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}})+\sqrt{32(\frac{1}{m}+\frac{1}{n})\log\frac{1}{\delta}}}
\end{aligned} \tag{A.27}$$

Table 1. Classification accuracies and their standard deviations for WiFi localization dataset.

	Softmax	TCA	DIP	CIC	DCIC
t1 → t2	60.73 ± 0.66	70.80 ± 1.66	71.40 ± 0.83	75.50 ± 1.02	**79.28 ± 0.56**
t1 → t3	55.20 ± 1.22	67.43 ± 0.55	64.65 ± 0.32	69.05 ± 0.28	**70.75 ± 0.91**
t2 → t3	54.38 ± 2.01	63.58 ± 1.33	66.71 ± 2.63	70.92 ± 3.86	**77.28 ± 2.87**
hallway1	40.81 ± 12.05	42.78 ± 7.69	44.31 ± 8.34	51.83 ± 8.73	**59.31 ± 12.30**
hallway2	27.98 ± 10.28	43.68 ± 11.07	44.61 ± 5.94	43.96 ± 6.20	**60.50 ± 8.68**
hallway3	24.94 ± 9.89	31.44 ± 5.47	33.50 ± 2.58	32.00 ± 3.88	**33.89 ± 5.94**

Combining the above inequality with those in Lemma A.1 and Lemma A.2, we have

$$\begin{aligned}
&\mathcal{D}(\hat{W}, \hat{\alpha}) - \mathcal{D}(\hat{W}, \alpha^*) \\
&\leq 2 \sup_{\alpha \in \Delta} |\mathcal{D}(\hat{W}, \alpha) - \hat{\mathcal{D}}(\hat{W}, \alpha)| \\
&\leq 4(\wedge_{\hat{Q}} + 1) \wedge_{\hat{W}} \sup_{\alpha \in \Delta} \|f(\mathbf{x}^S, \mathbf{x}^T, \alpha)\| \\
&\leq 8(\wedge_{\hat{Q}} + 1)^2 \wedge_{\hat{W}}^2 \sqrt{\frac{\sqrt{c}}{\sqrt{m}} + \frac{\sqrt{c}}{\sqrt{n}} + \sqrt{2(\frac{1}{m} + \frac{1}{n}) \log \frac{1}{\delta}}},
\end{aligned} \tag{A.28}$$

which concludes the proof of Theorem 2. ■

4. Additional Description of Methods

4.1. Linear Model

Linear model is a two-stage model in which we first identify invariant representations and $P^T(Y)$ and then train the classifier according to the importance reweighting framework. In linear model, $\tau(x_i) = x_i' = W^\top x_i$. To avoid the trivial solution, W is constrained to be orthogonal. Then, according to Eq. (5) in the main paper, we have

$$\begin{aligned}
&\min_{W,\alpha} \hat{\mathcal{D}}(W, \alpha) = \|\frac{1}{m}\psi(W^\top \mathbf{x}^S) G\alpha - \frac{1}{n}\psi(W^\top \mathbf{x}^T)\mathbf{1}\|^2, \\
&\text{s.t.} \quad W^\top W = I; \sum_{i=1}^{c} \alpha_i = 1; \\
&\qquad \alpha_i \geq 0, \forall i \in \{1, \cdots, c\}.
\end{aligned}$$

Note that even though the objective function has similar form with that in (Gong et al., 2016), it is essentially different. This is because in this objective function, the source data is noisily labeled and G is carefully designed to relate $P^S_\rho(X, Y)$ and $P^T(X)$ such that conditional invariant components and $P^T(Y)$ can be identified from noisy source data and unlabeled target data.

The alternating optimization method is applied to update W and α. Specifically, we apply the conjugate gradient algorithm on the Grassmann manifold to optimize W, and use the quadratic programming to optimize α. After identifying the invariant representations and P^T_Y by solving above problem, we can then use them to train a classifier for the target data by minimizing Eq. (8) in the main paper.

4.2. Structure of Deep Model

Figure 1 illustrates the pipeline of our end-to-end deep domain adaptation model.

5. Additional Experiments

In this section, we give some additional experimental results on WiFi Localization dataset.

WiFi Localization Dataset. We further compare our linear DCIC model with DIP, TCA, and CIC on the cross-domain indoor WiFi localization dataset (Zhang et al., 2013).

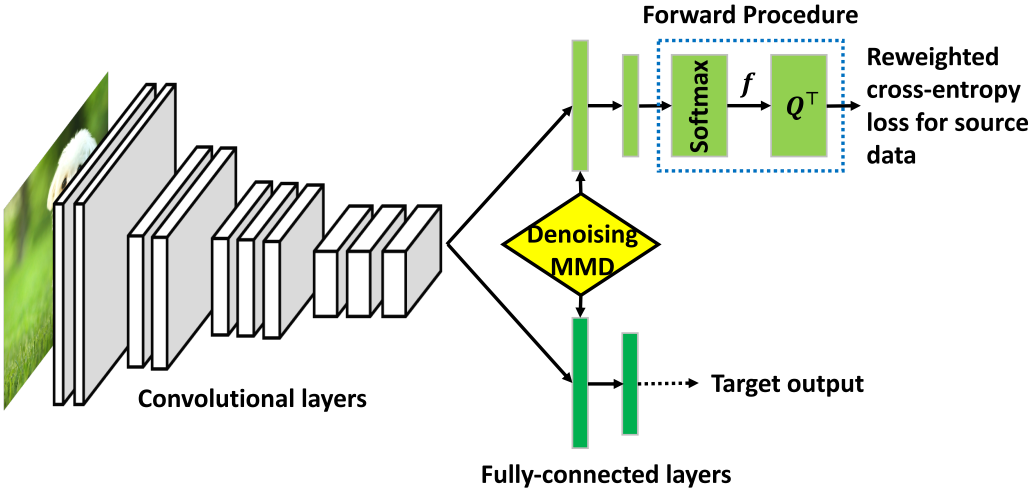


Figure 1. An overview of the proposed end-to-end deep domain adaptation model.

The problem is to learn the function between signals X and locations Y. Here, we view it as a classification problem, where each location space is assigned with a discrete label. In the prediction stage, the label is then converted to the location information. We resample the training set to simulate the changes in P_Y.

To ensure that the class ratio is not a vector of all ones, we resample the source training examples. We randomly select $c/2$ classes and let their class ratios be 2.5. For the other $c/2$ classes, we set their $P(Y)$ to be equal. The flip rate from one label to another is set to $\frac{\rho}{c-1}$.

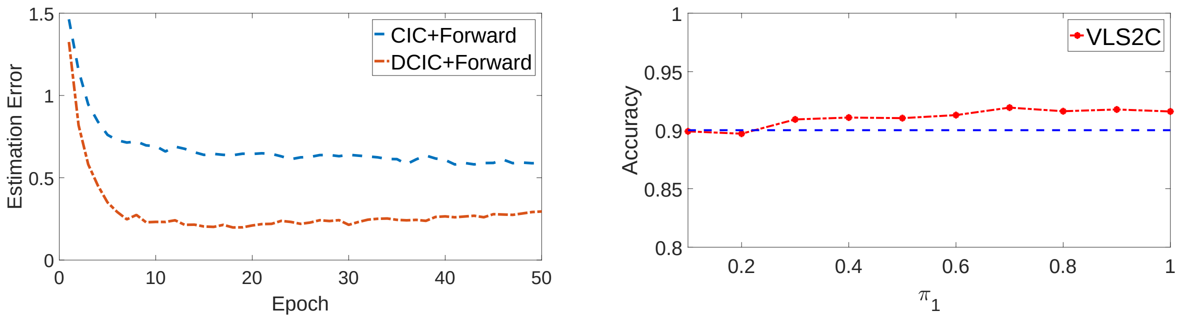
We first learn the linear transformation $W \in \mathbb{R}^{d \times d'}$ ($d' = 10$) and extract the invariant components. A neural network with one hidden layer is trained by minimizing Eq. (7) in the main paper and then obtain the classifier for the signals in target domain according to Eq. (8) in the main paper. The output layer is a softmax with the cross-entropy loss. The activation function in the hidden layer is the Rectified Linear Unit (ReLU). The number of neurons in the hidden layer is set to 800. During training, learning rate is fixed to 0.1. After training, as in (Gong et al., 2016), we report the percentage of examples on which the difference between the predicted and true locations is within 3 meters. Here, we train a neural network with the raw features as the baseline. All the experiments are repeated 10 times and the average performances are reported. In Table 1, the three upper rows present the transfer across different time periods $t1, t2$, and $t3$, where $\rho = 0.4$. The lower part shows the transfer across different devices, where $\rho = 0.2$. We can see that all the results show DCIC can better transfer the invariant knowledge than other methods.

See the results in the lower parts, since the input features in two domains are too complex in these cases, the invariant components cannot be well identified by a simple linear transformation, which finally results in the degraded performances. Therefore, for data with complex features, we would like to introduce our deep denoising models to extract invariant components and to correct the shift. The experiments on deep models are shown in the following subsections.

5.1. More Discussions

5.1.1. Convergence analysis

In order to verify the effectiveness of the proposed method to estimate P_Y^T, in Figure 2 (a), we show the convergence of the estimation errors $\frac{\|\alpha^* - \hat{\alpha}\|_2}{\|\alpha^*\|_2}$ of our "DCIC + Forward $\hat{Q}$" method and the "CIC + Forward $\hat{Q}$" method, where α^* is the true class prior and $\hat{\alpha}$ is the estimated one. The experiment is conducted on the mnist2usps dataset. We can see that our proposed method can find a better solution for P_Y^T after using our denoising MMD loss.



(a)

(b)

Figure 2. (a) The convergence of class prior estimation in target domain. (b) The sensitivity analysis of the parameter π_1.

5.1.2. Parameter sensitivity

Here, we check the sensitivity of the trade-off parameter π_1 of our denoising MMD loss. Figure 2 (b) shows the classification accuracies with respect to different values of π_1, which ranges from 0.1 to 1.0 with step 0.1. This task is evaluated on VLS2C dataset. We can see, the overall performance is not very sensitive to the choice of π_1. In our experiments, we find $\pi_1 = 1.0$ works well on all other datasets.

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press, 2013.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *ICML*, pp. 2839–2848, 2016.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and processes.* Springer Science & Business Media, 2013.

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

Yu, X., Liu, T., Gong, M., Batmanghelich, K., and Tao, D. An efficient and provable approach for mixture proportion estimation using linear independence assumption. In *CVPR*, 2018.

Zhang, K., Zheng, V., Wang, Q., Kwok, J., Yang, Q., and Marsic, I. Covariate shift in hilbert space: A solution via sorrogate kernels. In *ICML*, pp. 388–395, 2013.