Supplement to "Domain Adaptation with Conditional Transferable Components"

This supplementary material provides the proofs and some details which are omitted in the submitted paper. The equation numbers in this material are consistent with those in the paper.

S1. Proof of Theorem 1

Proof. Combine (3) and (4), we have

$$\sum_{c=1}^{C} p^{\mathcal{T}}(Y=v_c)p^{\mathcal{T}}(X^{ci}|Y=v_c) = \sum_{c=1}^{C} p^{new}(Y=v_c)p^{\mathcal{S}}(X^{ci}|Y=v_c). \tag{15}$$

If the transformation W is non-trivial, there do not exist non-zero $\gamma_1, ..., \gamma_C$ and $\nu_1, ..., \nu_C$ such that $\sum_{c=1}^{C}\gamma_c p^{\mathcal{T}}(X^{ci}|Y=v_c)=0$ and $\sum_{c=1}^{C}\nu_c p^{\mathcal{S}}(X^{ci}|Y=v_c)=0$. Therefore, we can transform (15) to

$$\sum_{c=1}^{C} P^{\mathcal{T}}(Y=v_c)P^{\mathcal{T}}(X^{ci}|Y=v_c) - P^{new}(Y=v_c)p^{\mathcal{S}}(X^{ci}|Y=v_c) = 0. \tag{16}$$

According to $\mathbf{A^{CIC}}$ in Theorem 1, we have $\forall c$,

$$P^{\mathcal{T}}(Y=v_c)P^{\mathcal{T}}(X^{ci}|Y=v_c) - P^{new}(Y=v_c)P^{\mathcal{S}}(X^{ci}|Y=v_c) = 0. \tag{17}$$

Taking the integral of (17) leads to $P^{new}(Y=v_c)=P^{\mathcal{T}}(Y=v_c)$, which further implies that $P^{\mathcal{S}}(X^{ci}|Y=v_c) = P^{\mathcal{T}}(X^{ci}|Y=v_c)$. □

S2. Proof of Lemma 1

Proof.

$$\begin{aligned}
\epsilon_{\mathcal{T}}(h) =& \epsilon_{\mathcal{T}}(h) + \epsilon_{\text{new}}(h) - \epsilon_{\text{new}}(h) \\
\le& \epsilon_{\text{new}}(h) + \left|\epsilon_{\mathcal{T}}(h) - \epsilon_{\text{new}}(h)\right| \\
\le& \epsilon_{\text{new}}(h) + \int \left|P^{\text{new}}(X^{ci},Y) - P^{\mathcal{T}}(X^{ci},Y)\right|\left|L(Y,h(X^{ci}))\right| dX^{ci}dY \\
\le& \epsilon_{\text{new}}(h) + d_1(p^{\text{new}}(X^{ci},Y), p^{\mathcal{T}}(X^{ci},Y)).
\end{aligned} \tag{18}$$

□

S3. Proof of Theorem 2

Proof. In the binary classification problem, because $Y \in \{0,1\}$ is a discrete variable, we use the Kronecker delta kernel for Y. Then (13) becomes

$$\begin{aligned}
& d_k(p^{\text{new}}(X^{ci},Y), p^{\mathcal{T}}(X^{ci},Y)) \\
&= \sum_{c=0}^{1} \left|\left|P^{\text{new}}(Y=c)\mu_{p^{\mathcal{S}}(X^{ci}|Y=c)}[\psi(X^{ci})] - P^{\mathcal{T}}(Y=c)\mu_{p^{\mathcal{T}}(X^{ci}|Y=c)}[\psi(X^{ci})]\right|\right|^2 \\
&= \left|\left|\Delta_1\right|\right|^2 + \left|\left|\Delta_0\right|\right|^2 \\
&= \left|\left|\Delta_1 + \Delta_0\right|\right|^2 - 2\Delta_1^{\intercal}\Delta_0
\end{aligned}$$

$$
\begin{aligned}
&= \Big|\Big| \sum_{c=0}^{1} P^{\text{new}}(Y=c)\mu_{p^{\mathcal{S}}(X^{ci}|Y=c)}[\psi(X^{ci})] - \sum_{c=0}^{1} P^{\mathcal{T}}(Y=c)\mu_{p^{\mathcal{T}}(X^{ci}|Y=c)}[\psi(X^{ci})]\Big|\Big|^2 - 2\Delta_1^\intercal\Delta_0 \\
&= \Big|\Big|\mu_{p^{\text{new}}(X^{ci})}[\psi(X^{ci})] - \mu_{p^{\mathcal{T}}(X^{ci})}[\psi(X^{ci})]\Big|\Big|^2 - 2\Delta_1^\intercal\Delta_0 \\
&= J^{ci} - 2\Delta_1^\intercal\Delta_0. \qquad (19)
\end{aligned}
$$

Clearly, when $0 < \theta \le \pi/2$, we have $\Delta_1^\intercal\Delta_0 \ge 0$. Therefore,

$$
d_k(p^{\text{new}}(X^{ci}, Y), p^{\mathcal{T}}(X^{ci}, Y)) \le J^{ci}. \qquad (20)
$$

When $\pi/2 < \theta \le \pi$, we express J^{ci} as

$$
\begin{aligned}
J^{ci} =& \big|\big|\Delta_1 + \Delta_0\big|\big|^2 \\
=& \big|\big|\Delta_1\big|\big|^2 + \big|\big|\Delta_0\big|\big|^2 + 2\big|\big|\Delta_1\big|\big|\big|\big|\Delta_0\big|\big|\cos\theta \\
=& (\big|\big|\Delta_1\big|\big| + \big|\big|\Delta_0\big|\big|\cos\theta)^2 + \big|\big|\Delta_0\big|\big|^2\sin^2\theta \qquad (21)\\
=& (\big|\big|\Delta_0\big|\big| + \big|\big|\Delta_1\big|\big|\cos\theta)^2 + \big|\big|\Delta_1\big|\big|^2\sin^2\theta. \qquad (22)
\end{aligned}
$$

According to (21) and (22), we have $\big|\big|\Delta_0\big|\big|^2\sin^2\theta \le J^{ci}$ and $\big|\big|\Delta_1\big|\big|^2\sin^2\theta \le J^{ci}$. Thus

$$
d_k(p^{\text{new}}(X^{ci}, Y), p^{\mathcal{T}}(X^{ci}, Y)) = \big|\big|\Delta_1\big|\big|^2 + \big|\big|\Delta_0\big|\big|^2 \le 2\frac{J^{ci}}{sin^2\theta}. \qquad (23)
$$

Combining (20) and (23), we can obtain the results in Theorem 2. □

S4. Proof of Theorem 3

Proof. We have

$$
\begin{aligned}
\hat{J}^{ci}(\boldsymbol{\beta}, W) \quad &= \Big|\Big|\frac{1}{n^{\mathcal{S}}}\psi\left(W^\intercal\mathbf{x}^{\mathcal{S}}\right)\boldsymbol{\beta} - \frac{1}{n^{\mathcal{T}}}\psi(W^\intercal\mathbf{x}^{\mathcal{T}})\mathbf{1}\Big|\Big|^2 \\
&= \Big|\Big|\frac{1}{n^{\mathcal{S}}}\psi\left(W^\intercal\mathbf{x}^{\mathcal{S}}\right)R^{dis}\boldsymbol{\alpha} - \frac{1}{n^{\mathcal{T}}}\psi(W^\intercal\mathbf{x}^{\mathcal{T}})\mathbf{1}\Big|\Big|^2 \\
&= \Big|\Big|[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W^\intercal x_{1i}^{\mathcal{S}}), \ldots, \frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha} - \frac{1}{n^{\mathcal{T}}}\psi(W^\intercal\mathbf{x}^{\mathcal{T}})\mathbf{1}\Big|\Big|^2 \\
&= \hat{J}^{ci}(\boldsymbol{\alpha}, W), \qquad (24)
\end{aligned}
$$

where $x_{ci}^{\mathcal{S}}, c \in \{1, \ldots, C\}$ denotes the i-th observation of the c-th class in the source domain.

Define $\Delta = \{\boldsymbol{\alpha}|\boldsymbol{\alpha} \ge 0, \sum_{c=1}^{C}\boldsymbol{\alpha}_c = 1\}$. We have

$$
\begin{aligned}
&J^{ci}(\boldsymbol{\alpha}_n, W_n) - J^{ci}(\boldsymbol{\alpha}^*, W_n) \\
&= J^{ci}(\boldsymbol{\alpha}_n, W_n) - \hat{J}^{ci}(\boldsymbol{\alpha}_n, W_n) + \hat{J}^{ci}(\boldsymbol{\alpha}_n, W_n) - \hat{J}^{ci}(\boldsymbol{\alpha}^*, W_n) + \hat{J}^{ci}(\boldsymbol{\alpha}^*, W_n) - J^{ci}(\boldsymbol{\alpha}^*, W_n) \\
&\quad \text{Since } \boldsymbol{\alpha}_n \text{ is the empirical minimizer and thus } \hat{J}^{ci}(\boldsymbol{\alpha}_n, W_n) \le \hat{J}^{ci}(\boldsymbol{\alpha}^*, W_n) \\
&\le J^{ci}(\boldsymbol{\alpha}_n, W_n) - \hat{J}^{ci}(\boldsymbol{\alpha}_n, W_n) + \hat{J}^{ci}(\boldsymbol{\alpha}^*, W_n) - J^{ci}(\boldsymbol{\alpha}^*, W_n) \\
&\le 2\sup_{\boldsymbol{\alpha}\in\Delta}|J^{ci}(\boldsymbol{\alpha}, W_n) - \hat{J}^{ci}(\boldsymbol{\alpha}, W_n)|. \qquad (25)
\end{aligned}
$$

Before upper bounding the above defect on the right hand side, we enable some properties of the kernel. Assume that there exists a $\psi_{\max}$ such that for any $x \in \mathcal{X}$, it holds that $-\psi_{\max} \le \psi(x) \le \psi_{\max}$ and that $\big|\big|\psi_{\max}\big|\big|_2 \le \wedge_2$. Since $\boldsymbol{\alpha} \ge 0$ and $\big|\big|\boldsymbol{\alpha}\big|\big|_1 = 1$, for any $\mathbf{x}^{\mathcal{S}}$, it also holds that $[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^\intercal x_{1i}^{\mathcal{S}}), \ldots, \frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha} \le \psi_{\max}$.

Now, we have the following Lipschitz property of J^{ci}:

$$
|J^{ci}(\boldsymbol{\alpha}, W_n) - \hat{J}^{ci}(\boldsymbol{\alpha}, W_n)|
$$

$$
\begin{aligned}
&\leq |\max_{\boldsymbol{\alpha},\mathbf{x}^{\mathcal{S}}}[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^\intercal x_{1i}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha}+\max_{\mathbf{x}^{\mathcal{S}}}\frac{1}{n^{\mathcal{T}}}\psi(W_n^\intercal \mathbf{x}^{\mathcal{T}})\mathbf{1}|^\intercal|\mathbb{E}_{(Y,X)\sim p^{\mathcal{S}}}[\beta(Y)\psi(W_n^\intercal X)]\\
&\quad -\mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W_n^\intercal X)]-[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^\intercal x_{1i}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha}+\frac{1}{n^{\mathcal{T}}}\psi(W_n^\intercal \mathbf{x}^{\mathcal{T}})\mathbf{1}|\\
&\leq 2|\psi_{\max}|^\intercal|\mathbb{E}_{(Y,X)\sim p^{\mathcal{S}}}[\beta(Y)\psi(W_n^\intercal X)]\\
&\quad -\mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W_n^\intercal X)]-[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^\intercal x_{1i}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha}+\frac{1}{n^{\mathcal{T}}}\psi(W_n^\intercal \mathbf{x}^{\mathcal{T}})\mathbf{1}|. \qquad (26)
\end{aligned}
$$

Then, combining (25) and (26), we have

$$
\begin{aligned}
&J^{ci}(\boldsymbol{\alpha}_n,W_n)-J^{ci}(\boldsymbol{\alpha}^*,W_n)\\
&\leq 2\sup_{\boldsymbol{\alpha}\in\Delta}|J^{ci}(\boldsymbol{\alpha},W_n)-\hat{J}^{ci}(\boldsymbol{\alpha},W_n)|\\
&\leq 4\sup_{\boldsymbol{\alpha}\in\Delta}|\psi_{\max}|^\intercal|\mathbb{E}_{(Y,X)\sim p^{\mathcal{S}}}[\beta(Y)\psi(W_n^\intercal X)]\\
&\quad -\mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W_n^\intercal X)]-[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^\intercal x_{1i}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha}+\frac{1}{n^{\mathcal{T}}}\psi(W_n^\intercal \mathbf{x}^{\mathcal{T}})\mathbf{1}|. \qquad (27)
\end{aligned}
$$

Now, we are going to upper bound the defect:

$$
\begin{aligned}
f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}) \quad &\triangleq \mathbb{E}_{(Y,X)\sim p^{\mathcal{S}}}[\boldsymbol{\beta}(Y)\psi(W_n^\intercal X)]\\
&-\mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W_n^\intercal X)]-[\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^\intercal x_{1i}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^\intercal x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha}+\frac{1}{n^{\mathcal{T}}}\psi(W_n^\intercal \mathbf{x}^{\mathcal{T}})\mathbf{1}. \quad (28)
\end{aligned}
$$

We employ the McDiarmid's inequality to upper bound the ℓ_2-norm of the defect.

Theorem 4 (McDiarmid's inequality)**.** *Let $X=(X_1,\ldots,X_n)$ 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 $c_1,\ldots,c_n>0$ such that $f:\mathcal{X}^n\rightarrow\mathbb{R}$ satisfies the following conditions:*

$$
|f(X)-f(X^i)|\leq c_i, \forall i\in\{1,\ldots,n\}. \qquad (29)
$$

Then for any $X\in\mathcal{X}^n$ and $\epsilon>0$, the following inequalities hold:

$$
Pr\{|Ef(X)-f(X)|\geq\epsilon\}\leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right). \qquad (30)
$$

We now check that $f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})=\left|\left|f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})\right|\right|^2$ satisfies the bounded difference property. Let $\mathbf{x}_{ci}^{\mathcal{S}}$ denote the i-th observation belonging to the c-th class. We have

$$
\begin{aligned}
&|f(X,\mathbf{x}_i^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})-f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})|\\
&=|(f^{\psi}(X,\mathbf{x}_i^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})+f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}))^T(f^{\psi}(X,\mathbf{x}_i^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})-f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}))|\\
&\leq 4|\psi_{\max}|^\intercal|f^{\psi}(X,\mathbf{x}_i^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})-f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})|\\
&=4|\psi_{\max}|^\intercal|\frac{\boldsymbol{\alpha}_c}{n_c}(\psi(W_n^\intercal\mathbf{x}_{ci}^{\mathcal{S}})-\psi(W_n^\intercal{\mathbf{x}'}_{ci}^{\mathcal{S}}))|\\
&\leq\frac{8\boldsymbol{\alpha}_c}{n_c}|\psi_{\max}|^\intercal|\psi_{\max}|\leq\frac{8\wedge_2^2\boldsymbol{\alpha}_c}{n_c}. \qquad (31)
\end{aligned}
$$

Similarly, it holds that

$$
|f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}_i^{\mathcal{T}})-f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})|\leq\frac{8\wedge_2^2}{n^{\mathcal{S}}}. \qquad (32)
$$

Employing McDiarmid's inequality, we have that

$$\Pr\{|f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}) - E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}} f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})| \geq \epsilon\} \leq 2\exp\left(\frac{-2\epsilon^2}{64\wedge_2^4 (\sum_{c=1}^{C}\frac{\boldsymbol{\alpha}_c^2}{n_c}+\frac{1}{n^{\mathcal{T}}})}\right). \tag{33}$$

Combining (27) and (33), we have that for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}
&J^{ci}(\boldsymbol{\alpha}_n, W_n) - J^{ci}(\boldsymbol{\alpha}^*, W_n)\\
&\leq 2\sup_{\boldsymbol{\alpha}\in\Delta}|J^{ci}(\boldsymbol{\alpha}, W_n) - \hat{J}^{ci}(\boldsymbol{\alpha}, W_n)|\\
&\leq 4\sup_{\boldsymbol{\alpha}\in\Delta}|\psi_{\max}|^{\intercal}|f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})|\\
&\quad \text{Using Cauchy-Schwarz inequality}\\
&\leq 4\sup_{\boldsymbol{\alpha}\in\Delta}||\psi_{\max}||||f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})||\\
&\leq 4\wedge_2\left(E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta} f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}) + \wedge_2^2\sqrt{32\log\frac{2}{\delta}(\sum_{c=1}^{C}\frac{\boldsymbol{\alpha}_c^2}{n_c}+\frac{1}{n^{\mathcal{T}}})}\right)^{\frac{1}{2}}\\
&\leq 4\wedge_2\left(E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta} f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}) + 32\wedge_2^2\sqrt{\frac{1}{2}\log\frac{2}{\delta}(\max_{c\in\{1,\ldots,C\}}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}})}\right)^{\frac{1}{2}}.
\end{aligned} \tag{34}$$

Now we are going to upper bound the term $E_X \sup_{\boldsymbol{\alpha}\in\Delta} f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})$. Let

$$g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}) \triangleq [\frac{1}{n_1}\sum_{i=1}^{n_1}\psi(W_n^{\intercal}x_{1i}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\psi(W_n^{\intercal}x_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha} - \frac{1}{n^{\mathcal{T}}}\psi(W_n^{\intercal}\mathbf{x}^{\mathcal{T}})\mathbf{1} \tag{35}$$

and

$$g(X) \triangleq \mathbb{E}_{(Y,X)\sim p^{\mathcal{S}}}[\boldsymbol{\beta}(Y)\psi(W_n^{\intercal}X)] - \mathbb{E}_{X\sim p^{\mathcal{T}}}[\psi(W_n^{\intercal}X)]. \tag{36}$$

We have that

$$\begin{aligned}
&E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta}||f^{\psi}(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})||^2\\
&= E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta}||g(X) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})||^2\\
&= E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta}||E_{\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}} g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})||^2\\
&\leq E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta}||g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})||^2.
\end{aligned} \tag{37}$$

where $\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}$ are ghost samples which are i.i.d. with $\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}$, respectively.

Since $\mathbf{x}^j,\mathbf{x}'^j, j=\mathcal{S},\mathcal{T}$ are i.i.d. samples, $\sum_{i=1}^{n_c}\psi(W_n^{\intercal}\mathbf{x}_{ci}^j) - \psi(W_n^{\intercal}\mathbf{x}'^j_{ci})$ has a symmetric property, which means it has an even density function. Thus, $\sum_{i=1}^{n_c}\psi(W_n^{\intercal}\mathbf{x}_{ci}^j) - \psi(W_n^{\intercal}\mathbf{x}'^j_{ci})$ and $\sum_{i=1}^{n_c}\sigma_{ci}(\psi(W_n^{\intercal}\mathbf{x}_{ci}^j) - \psi(W_n^{\intercal}\mathbf{x}'^j_{ci}))$ has the same distribution, where σ_{ci} are independent variables uniformly distributed from $\{-1,1\}$. Then, we have

$$E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}}\sup_{\boldsymbol{\alpha}\in\Delta}||g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}})||^2 = E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}}\sup_{\boldsymbol{\alpha}\in\Delta}||g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\boldsymbol{\sigma})||^2, \tag{38}$$

where

$$g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\boldsymbol{\sigma}) \triangleq [\frac{1}{n_1}\sum_{i=1}^{n_c}\sigma_{1i}(\psi(W_n^{\intercal}\mathbf{x}_{ci}^{\mathcal{S}})\ldots\frac{1}{n_C}\sum_{i=1}^{n_C}\sigma_{Ci}(\psi(W_n^{\intercal}\mathbf{x}_{Ci}^{\mathcal{S}})]\boldsymbol{\alpha} - \frac{1}{n^{\mathcal{T}}}\sum_{i=1}^{n^{\mathcal{T}}}\sigma_{\mathcal{T}i}\psi(W_n^{\intercal}\mathbf{x}_i^{\mathcal{T}}). \tag{39}$$

According to Talagrand contraction Lemma, we have

$$
\begin{aligned}
& E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha}\in\Delta} \left|\left|g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\boldsymbol{\sigma})\right|\right|^2 \\
& \leq 2E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha}\in\Delta} |\psi_{\max}|^\intercal |g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\boldsymbol{\sigma})| \\
& \leq 4E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha}\in\Delta} |\psi_{\max}|^\intercal g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\boldsymbol{\sigma}) \\
& = 4E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha}\in\Delta} |\psi_{\max}|^\intercal \\
& \left\langle [\boldsymbol{\alpha}^\intercal,-1]^\intercal, [\frac{1}{n_1}\sum_{i=1}^{n_c}\sigma_{1i}(\psi(W_n^\intercal \mathbf{x}_{ci}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\sigma_{Ci}(\psi(W_n^\intercal \mathbf{x}_{Ci}^{\mathcal{S}}),\frac{1}{n^{\mathcal{T}}}\sum_{i=1}^{n^{\mathcal{T}}}\sigma_{\mathcal{T}i}\psi(W_n^\intercal \mathbf{x}_i^{\mathcal{T}})]^\intercal \right\rangle. \qquad (40)
\end{aligned}
$$

Let

$$
\boldsymbol{v} \triangleq [\frac{1}{n_1}\sum_{i=1}^{n_c}\sigma_{1i}(\psi(W_n^\intercal \mathbf{x}_{ci}^{\mathcal{S}}),\ldots,\frac{1}{n_C}\sum_{i=1}^{n_C}\sigma_{Ci}(\psi(W_n^\intercal \mathbf{x}_{Ci}^{\mathcal{S}}),\frac{1}{n^{\mathcal{T}}}\sum_{i=1}^{n^{\mathcal{T}}}\sigma_{\mathcal{T}i}\psi(W_n^\intercal \mathbf{x}_i^{\mathcal{T}})]^\intercal. \qquad (41)
$$

Since $\left|\left|[\boldsymbol{\alpha}^\intercal,-1]^\intercal\right|\right|_2 \leq 2$, using Cauchy-Schwarz inequality again, we have

$$
\begin{aligned}
& E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha}\in\Delta} \left|\left|g_n(\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}) - g_n(\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\boldsymbol{\sigma})\right|\right|^2 \\
& \leq 4E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} \sup_{\boldsymbol{\alpha}\in\Delta} |\psi_{\max}|^\intercal \left\langle [\boldsymbol{\alpha}^\intercal,-1]^\intercal, \boldsymbol{v}\right\rangle \\
& \leq 8E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}},\boldsymbol{\sigma}} |\psi_{\max}|^\intercal \sqrt{\boldsymbol{v}^\intercal\boldsymbol{v}} \\
& \leq 8E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}} |\psi_{\max}|^\intercal \sqrt{E_{\boldsymbol{\sigma}}\boldsymbol{v}^\intercal\boldsymbol{v}} \\
& = 8E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}},\mathbf{x}'^{\mathcal{S}},\mathbf{x}'^{\mathcal{T}}} |\psi_{\max}|^\intercal \sqrt{\sum_{c=1}^{C}\frac{1}{n_c^2}\sum_{i=1}^{n_c}(\psi(W_n^\intercal \mathbf{x}_{ci}^{\mathcal{S}}))^2 + \frac{1}{(n^{\mathcal{T}})^2}\sum_{i=1}^{n^{\mathcal{T}}}(\psi(W_n^\intercal \mathbf{x_i}^{\mathcal{T}}))^2} \\
& \leq 8|\psi_{\max}|^\intercal|\psi_{\max}|\sqrt{\sum_{c=1}^{C}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}}} \\
& \leq 8\wedge_2^2\sqrt{\sum_{c=1}^{C}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}}}. \qquad (42)
\end{aligned}
$$

At the end, combining (34), (37) and (42), with probability at least $1-\delta$, we have

$$
\begin{aligned}
& J^{ci}(\boldsymbol{\alpha}_n, W_n) - J^{ci}(\boldsymbol{\alpha}^*, W_n) \\
& \leq 4\wedge_2 \left(E_{\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}} \sup_{\boldsymbol{\alpha}\in\Delta} f(X,\mathbf{x}^{\mathcal{S}},\mathbf{x}^{\mathcal{T}}) + 32\wedge_2\sqrt{\frac{1}{2}\log\frac{2}{\delta}(\max_{c\in\{1,\ldots,C\}}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}})}\right)^{\frac{1}{2}} \\
& = 4\wedge_2 \left(8\wedge_2^2\sqrt{\sum_{c=1}^{C}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}}} + 32\wedge_2^2\sqrt{\frac{1}{2}\log\frac{2}{\delta}(\max_{c\in\{1,\ldots,C\}}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}})}\right)^{\frac{1}{2}} \\
& \leq 8\wedge_2^2 \left(2\sqrt{\sum_{c=1}^{C}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}}} + 8\sqrt{\frac{1}{2}\log\frac{2}{\delta}(\max_{c\in\{1,\ldots,C\}}\frac{1}{n_c}+\frac{1}{n^{\mathcal{T}}})}\right)^{\frac{1}{2}}.
\end{aligned}
$$

The proof ends. □

S5. Derivatives used in Sec. 2.5

The gradient of $\hat{J}^{ct}$ w.r.t. $\tilde{K}^{\mathcal{S}}$, $\tilde{K}^{\mathcal{T},\mathcal{S}}$, and $K^{\mathcal{T}}$ is

$$\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{S}}} = \frac{1}{n^{\mathcal{S}2}}\boldsymbol{\beta}\boldsymbol{\beta}^{\intercal}, \frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{T},\mathcal{S}}} = -\frac{2}{n^{\mathcal{S}}n^{\mathcal{T}}}\mathbf{1}\boldsymbol{\beta}^{\intercal}, \text{and} \frac{\partial \hat{J}^{ct}}{\partial K^{\mathcal{T}}} = \frac{1}{n^{\mathcal{T}2}}\mathbf{1}\mathbf{1}^{\intercal}.$$

The gradient of $\mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]$ w.r.t. $\tilde{K}^{\mathcal{S}}$ is

$$\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial \tilde{K}^{\mathcal{S}}} = -\varepsilon(\tilde{K}^{\mathcal{S}} + n^{\mathcal{S}}\varepsilon I)^{-1}L(\tilde{K}^{\mathcal{S}} + n^{\mathcal{S}}\varepsilon I)^{-1}.$$

Using the chain rule, we further have the gradient of $\hat{J}^{ct}$ w.r.t. the entries of W, $\mathbf{G}$, and $\mathbf{H}$:

$$\frac{\partial \hat{J}^{ct}}{\partial W_{pq}} = Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{S}}}\right)^{\intercal}(\mathbf{D}^1_{pq})\right] - Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{T},\mathcal{S}}}\right)^{\intercal}(\mathbf{D}^2_{pq})\right] + Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{T}}}\right)^{\intercal}(\mathbf{D}^3_{pq})\right], \tag{43}$$

$$\frac{\partial \hat{J}^{ct}}{\partial \mathbf{G}_{pq}} = Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{S}}}\right)^{\intercal}(\mathbf{E}^1_{pq})\right] - Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{T},\mathcal{S}}}\right)^{\intercal}(\mathbf{E}^2_{pq})\right], \tag{44}$$

$$\frac{\partial \hat{J}^{ct}}{\partial \mathbf{H}_{pq}} = Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{S}}}\right)^{\intercal}(\mathbf{F}^1_{pq})\right] - Tr\left[\left(\frac{\partial \hat{J}^{ct}}{\partial \tilde{K}^{\mathcal{T},\mathcal{S}}}\right)^{\intercal}(\mathbf{F}^2_{pq})\right], \tag{45}$$

and the gradient of $\mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]$ w.r.t. the entries of W, $\mathbf{G}$, and $\mathbf{H}$:

$$\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial W_{pq}} = Tr\left[\left(\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial \tilde{K}^{\mathcal{S}}}\right)^{\intercal}(\mathbf{D}^1_{pq})\right], \tag{46}$$

$$\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial \mathbf{G}_{pq}} = Tr\left[\left(\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial \tilde{K}^{\mathcal{S}}}\right)^{\intercal}(\mathbf{E}^1_{pq})\right], \tag{47}$$

$$\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial \mathbf{H}_{pq}} = Tr\left[\left(\frac{\partial \mathbb{T}\text{r}[\hat{\Sigma}_{YY|X^{ct}}]}{\partial \tilde{K}^{\mathcal{S}}}\right)^{\intercal}(\mathbf{F}^1_{pq})\right], \tag{48}$$

where

$$[\mathbf{D}^1_{pq}]_{ij} = -\frac{\tilde{k}^s(x^s_i, x^s_j)}{\sigma^2}\big[\sum_{k=1}^{D} w_{kq}(a_{qi}x^s_{ik} - a_{qj}x^s_{jk})(a_{qi}x^s_{ip} - a_{qj}x^s_{jp}) + (a_{qi}x^s_{ip} - a_{qj}x^s_{jp})(b_{qi} - b_{qj})\big],$$

$$[\mathbf{D}^2_{pq}]_{ij} = -\frac{\tilde{k}^{t,s}(x^t_i, x^s_j)}{\sigma^2}\big[\sum_{k=1}^{D} w_{kq}(x^t_{ik} - a_{qj}x^s_{jk})(x^t_{ip} - a_{qj}x^s_{jp}) + a_{qj}x^s_{jp}b_{qj}\big],$$

$$[\mathbf{D}^3_{pq}]_{ij} = -\frac{\tilde{k}^t(x^t_i, x^t_j)}{\sigma^2}\big[\sum_{k=1}^{D} w_{kq}(x^t_{ik} - x^t_{jk})(x^t_{ip} - x^t_{jp})\big],$$

$$[\mathbf{E}^1_{pq}]_{ij} = -\frac{\tilde{k}^s(x^s_i, x^s_j)}{\sigma^2}(x^{ct}_{jq} - x^{ct}_{iq})(x^s_{jq}R^{dis}_{jp} - x^s_{iq}R^{dis}_{ip}),$$

$$[\mathbf{E}^2_{pq}]_{ij} = -\frac{\tilde{k}^{t,s}(x^t_i, x^s_j)}{\sigma^2}x^s_{jq}R^{dis}_{jp}(x^{ct}_{jq} - x^t_{iq}),$$

$$[\mathbf{F}^1_{pq}]_{ij} = -\frac{\tilde{k}^s(x^s_i, x^s_j)}{\sigma^2}(x^{ct}_{jq} - x^{ct}_{iq})(R^{dis}_{jp} - R^{dis}_{ip}),$$
$$[\mathbf{F}^2_{pq}]_{ij} = -\frac{\tilde{k}^{t,s}(x^t_i, x^s_j)}{\sigma^2}R^{dis}_{jp}(x^{ct}_{jq} - x^t_{iq}).$$

The derivative of J^{reg} w.r.t. $\mathbf{G}$ and $\mathbf{H}$ is

$$\frac{\partial J^{reg}}{\partial \mathbf{G}} = \frac{2\lambda_S}{n^S}R^{dis\intercal}(\mathbf{A}^\intercal - \mathbf{1}_{n^S \times d}), \text{ and}$$
$$\frac{\partial J^{reg}}{\partial \mathbf{H}} = \frac{2\lambda_L}{n^S}R^{dis\intercal}\mathbf{B}^\intercal.$$