Supplementary Material for Domain Neural Adaptation

Sentao Chen, Zijie Hong, Mehrtash Harandi, *Member, IEEE,* and Xiaowei Yang

I. Proofs and Statistical test

In this supplementary material, we present the details of the proofs and the statistical test in the main paper.
1, In Part A, we provide the proof for Theorem 1 in Section II-B.
2, In Part B, we provide the proof for Theorem 2 in Section II-C.
3, In Part C, we describe the procedure of the Wilcoxon signed-ranks test conducted in Section V-C.

A. Proof of Theorem 1 in Section II-B of the Main Paper

Theorem 1. *Assume that the loss $\ell \le M$ for some $M > 0$. Then, for any hypothesis $h \in \mathcal{H}$,*

$$\mathrm{E}_t[h] \le \mathrm{E}_s[h] + \frac{\sqrt{2}M}{1-\alpha}\sqrt{\mathrm{RCS}_\alpha(P^s, P^t)}. \tag{1}$$

Proof.

$$\begin{aligned}
\mathrm{E}_t[h] &= \mathrm{E}_s[h] + \mathrm{E}_t[h] - \mathrm{E}_s[h] && (2)\\
&\le \mathrm{E}_s[h] + \big|\mathrm{E}_t[h] - \mathrm{E}_s[h]\big| && (3)\\
&= \mathrm{E}_s[h] + \Big|\int \ell\big(h(\boldsymbol{x}), y\big)\big(P^s(\boldsymbol{x}, y) - P^t(\boldsymbol{x}, y)\big)\Big| d\boldsymbol{x}dy && (4)\\
&\le \mathrm{E}_s[h] + M\int \big|P^s(\boldsymbol{x}, y) - P^t(\boldsymbol{x}, y)\big| d\boldsymbol{x}dy && (5)\\
&= \mathrm{E}_s[h] + \frac{M}{1-\alpha}\int \big|P^s(\boldsymbol{x}, y) - P^\alpha(\boldsymbol{x}, y)\big| d\boldsymbol{x}dy && (6)\\
&\le \mathrm{E}_s[h] + \frac{\sqrt{2}M}{1-\alpha}\sqrt{\int P^s(\boldsymbol{x}, y)\log\frac{P^s(\boldsymbol{x}, y)}{P^\alpha(\boldsymbol{x}, y)} d\boldsymbol{x}dy} && (7)\\
&\le \mathrm{E}_s[h] + \frac{\sqrt{2}M}{1-\alpha}\sqrt{\int P^s(\boldsymbol{x}, y)\frac{P^s(\boldsymbol{x}, y)}{P^\alpha(\boldsymbol{x}, y)} d\boldsymbol{x}dy - 1} && (8)\\
&= \mathrm{E}_s[h] + \frac{\sqrt{2}M}{1-\alpha}\sqrt{\mathrm{RCS}_\alpha(P^s, P^t)}. && (9)
\end{aligned}$$

Inequality (5) is due to the property of integral and the assumption that ℓ is upper bounded by M. Eq. (6) introduces the α-mixture joint distribution $P^\alpha(\boldsymbol{x}, y) = \alpha P^s(\boldsymbol{x}, y) + (1-\alpha)P^t(\boldsymbol{x}, y)$ and equivalently rewrites (5). Inequality (7) is a direct result of the Pinsker's inequality [1]. Inequality (8) holds since $\log x \le x - 1$ for $x > 0$. □

B. Proof of Theorem 2 in Section II-C of the Main Paper

Theorem 2. *Assume that for all $r \in \mathcal{R}$, there exists $\widetilde{r} \in \mathcal{R}$, such that $|\widehat{F}(r) - F(r)| \le |\widehat{F}(\widetilde{r}) - F(\widetilde{r})|$. Then, for any $\delta \in (0, 1)$, there exists a positive integer N_δ, such that when $\min(m_s, m_t) > N_\delta$, with probability at least $1-\delta$,*

$$\begin{aligned}
&\Big|\widehat{\mathrm{RCS}}_\alpha(P^s, P^t) - \mathrm{RCS}_\alpha(P^s, P^t)\Big|\\
&\le 3\Big(\frac{1}{m_s} + \frac{1}{m_t}\Big) + \max_r F(r) - \max_{r\in\mathcal{R}} F(r). \qquad (10)
\end{aligned}$$

Proof. For the convenience of theoretical analysis, we denote $r^* = \mathrm{argmax}_r F(r)$, $r_0 = \mathrm{argmax}_{r\in\mathcal{R}} F(r)$, and $\widehat{r} = \mathrm{argmax}_{r\in\mathcal{R}} \widehat{F}(r)$, respectively. Subsequently, we bound the approximation error as follows

$$\begin{aligned}
&\Big|\widehat{\mathrm{RCS}}_\alpha(P^s, P^t) - \mathrm{RCS}_\alpha(P^s, P^t)\Big|\\
&= |\widehat{F}(\widehat{r}) - F(\widehat{r}) + F(\widehat{r}) - F(r^*)| && (11)\\
&\le |\widehat{F}(\widehat{r}) - F(\widehat{r})| + F(r^*) - F(\widehat{r}) && (12)\\
&= |\widehat{F}(\widehat{r}) - F(\widehat{r})| + F(r^*) - F(r_0) + F(r_0) - F(\widehat{r}) && (13)\\
&= |\widehat{F}(\widehat{r}) - F(\widehat{r})| + F(r^*) - F(r_0) + F(r_0) - \widehat{F}(r_0)\\
&\quad + \widehat{F}(r_0) - F(\widehat{r}) && (14)\\
&= |\widehat{F}(\widehat{r}) - F(\widehat{r})| + F(r^*) - F(r_0) + F(r_0) - \widehat{F}(r_0)\\
&\quad + \widehat{F}(r_0) - \widehat{F}(\widehat{r}) + \widehat{F}(\widehat{r}) - F(\widehat{r}) && (15)\\
&\le |\widehat{F}(\widehat{r}) - F(\widehat{r})| + |F(r_0) - \widehat{F}(r_0)| + |\widehat{F}(\widehat{r}) - F(\widehat{r})|\\
&\quad + F(r^*) - F(r_0) && (16)\\
&\le 3|\widehat{F}(\widetilde{r}) - F(\widetilde{r})| + F(r^*) - F(r_0). && (17)
\end{aligned}$$

Inequality (12) holds since $|a + b| \le |a| + |b|$ and $F(\widehat{r}) \le F(r^*)$. Inequality (16) is due to the fact that $\widehat{F}(r_0) - \widehat{F}(\widehat{r}) \le 0$. Finally, inequality (17) applies the assumption that $|\widehat{F}(r) - F(r)| \le |\widehat{F}(\widetilde{r}) - F(\widetilde{r})|$ for all $r \in \mathcal{R}$. Furthermore, we know from the law of large numbers and its properties [2] that the empirical mean $\widehat{F}(\widetilde{r})$ converges in probability to the expectation $F(\widetilde{r})$, *i.e.*, $\widehat{F}(\widetilde{r}) \xrightarrow{P} F(\widetilde{r})$. Therefore, for any $\delta \in (0, 1)$, there exists a positive integer N_δ, such that when $\min(m_s, m_t) > N_\delta$, with probability at least $1-\delta$,

$$|\widehat{F}(\widetilde{r}) - F(\widetilde{r})| \le \Big(\frac{1}{m_s} + \frac{1}{m_t}\Big). \tag{18}$$

This work was supported in part by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2020AAA0108404; in part by the National Natural Science Foundation of China under Grant 62106137; and in part by the Shantou University under Grant NTF21035. *(Corresponding author: Sentao Chen.)*

S. Chen is with the Department of Computer Science, Shantou University, Shantou 12386, China (e-mail: sentaochen@yahoo.com).
Z. Hong and X. Yang are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China.
M. Harandi is with the Department of Electrical and Computer Systems Engineering, Monash University, and Data61-CSIRO, Australia.

As a consequence,

$$\begin{aligned}
&\left|\widehat{\mathrm{RCS}}_\alpha(P^s, P^t) - \mathrm{RCS}_\alpha(P^s, P^t)\right| \\
&\leq 3|\widehat{F}(\widetilde{r}) - F(\widetilde{r})| + F(r^*) - F(r_0) && (19)\\
&\leq 3\Big(\frac{1}{m_s} + \frac{1}{m_t}\Big) + F(r^*) - F(r_0) && (20)\\
&= 3\Big(\frac{1}{m_s} + \frac{1}{m_t}\Big) + \max_{r} F(r) - \max_{r\in\mathcal{R}} F(r). && (21)
\end{aligned}$$

□

C. Statistical Test in Section V-C of the Main Paper

We describe the procedure of the Wilcoxon signed-ranks test [3], [4], [5] on the tasks from Table I, Table III, and Table VI in the main paper. The test compares the performance of two methods over multiple tasks. Specifically, in each task the classification accuracy is adopted as the performance measure of the methods. We fix DNA as a control method, and conduct 8 pairs of tests: DAN versus DNA, DANN versus DNA, CDAN versus DNA, DSAN versus DNA, GSDA versus DNA, JAN versus DNA, DeepJDOT versus DNA, and DNA_0 versus DNA. To run the test, we rank the differences in performance of two methods for each task out of N tasks. The differences are ranked according to their absolute values. The smallest absolute value gets the rank of 1, the second smallest gets the rank of 2, and so on. In case of equality, average ranks are assigned. The statistic of the Wilcoxon signed-ranks test is:

$$z(a,b) = \frac{T(a,b) - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}, \tag{22}$$

where $T(a,b) = \min\{R^+(a,b), R^-(a,b)\}$. $R^+(a,b)$ is the sum of ranks for the tasks on which method b outperforms method a and $R^-(a,b)$ is the sum of ranks for the opposite. They are defined as follows:

$$R^+(a,b) = \sum_{\mathrm{diff}_i > 0} \mathrm{rank}(\mathrm{diff}_i) + \frac{1}{2}\sum_{\mathrm{diff}_i = 0} \mathrm{rank}(\mathrm{diff}_i), \tag{23}$$

$$R^-(a,b) = \sum_{\mathrm{diff}_i < 0} \mathrm{rank}(\mathrm{diff}_i) + \frac{1}{2}\sum_{\mathrm{diff}_i = 0} \mathrm{rank}(\mathrm{diff}_i), \tag{24}$$

where diff_i is the difference between the accuracy of two methods on the i-th task out of N tasks, and $\mathrm{rank}(\mathrm{diff}_i)$ is the rank of $|\mathrm{diff}_i|$. We fix b as DNA, and let a vary from DAN to DNA_0 in turn. Based on formulas (22)-(24), we can compute $z(a,b)$ for the 8 pairs of tests.

References

[1] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, Nov 2016.

[2] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[3] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[4] S. Chen, L. Han, X. Liu, Z. He, and X. Yang, "Subspace distribution adaptation frameworks for domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5204–5218, 2020.

[5] S. Chen, M. Harandi, X. Jin, and X. Yang, "Domain adaptation by joint distribution invariant projections," *IEEE Transactions on Image Processing*, vol. 29, pp. 8264–8277, 2020.