Supplementary Material of "Learning with Augmented Features for Supervised and Semi-supervised Heterogeneous Domain Adaptation"

Wen Li, Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang

In this Supplementary Material, we provide the details for:

- The proof of Theorem 2.
- The proof of Proposition 1.

I. Proof of Theorem 2

In the main text, we have formulated our HFA as an infinite kernel learning problem as follows (*i.e.*, the Eq. (11) in the main text),

$$\min_{\boldsymbol{\theta}\in\mathcal{D}_{\boldsymbol{\theta}}}\max_{\boldsymbol{\alpha}\in\mathcal{A}}\ \mathbf{1}'\boldsymbol{\alpha}-\frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})'\sum_{r=1}^{\infty}\theta_r\mathbf{K}_r(\boldsymbol{\alpha}\circ\mathbf{y}),\quad(1)$$

where $\mathbf{K}_r=\mathbf{K}^{\frac{1}{2}}(\lambda\mathbf{M}_r+\mathbf{I})\mathbf{K}^{\frac{1}{2}}$, $\mathcal{A}=\{\boldsymbol{\alpha}|\mathbf{y}'\boldsymbol{\alpha}=0,\mathbf{0}\leq\boldsymbol{\alpha}\leq C\mathbf{1}\}$ and $\mathcal{D}_{\boldsymbol{\theta}}=\{\boldsymbol{\theta}|\mathbf{1}'\boldsymbol{\theta}\leq 1,\boldsymbol{\theta}\geq 0\}$. As shown in the main text, the dual form of (1) can be written as follows by introducing a dual variable τ for $\boldsymbol{\theta}$:

$$\begin{aligned}\max_{\tau,\boldsymbol{\alpha}\in\mathcal{A}}\quad&\mathbf{1}'\boldsymbol{\alpha}-\tau,\\ \text{s.t.}\quad&\frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})'\mathbf{K}_r(\boldsymbol{\alpha}\circ\mathbf{y})\leq\tau,\quad\forall r.\end{aligned}\quad(2)$$

Actually, the problem in (1) can also be deemed as the dual form of (2) by considering θ_r as the dual variable for the r-th constraint in (2).

In the main text, we have defined $F(\boldsymbol{\alpha},\boldsymbol{\theta})=\mathbf{1}'\boldsymbol{\alpha}-\frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})'\sum_{r=1}^{\infty}\theta_r\mathbf{K}_r(\boldsymbol{\alpha}\circ\mathbf{y})$, and denoted the optimal solution to (1) as $(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)=\arg\min_{\boldsymbol{\theta}\in\mathcal{D}_{\boldsymbol{\theta}}}\max_{\boldsymbol{\alpha}\in\mathcal{A}}F(\boldsymbol{\alpha},\boldsymbol{\theta})$. Let us denote $G(\boldsymbol{\alpha},\tau)=\mathbf{1}'\boldsymbol{\alpha}-\tau$, and the optimal solution to (2) as $(\boldsymbol{\alpha}^*,\tau^*)$. We first give the following lemma which will be used in the proof of Theorem 2:

Lemma 1. *For the infinite kernel learning problem in (1) and (2), we have $F(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)=G(\boldsymbol{\alpha}^*,\tau^*)$, and $\frac{1}{2}(\boldsymbol{\alpha}^*\circ\mathbf{y})'\sum_{r=1}^{\infty}\theta_r^*\mathbf{K}_r(\boldsymbol{\alpha}^*\circ\mathbf{y})=\tau^*$.*

Proof: The lemma can be proved by using the KKT condition. For any $\theta_r^*>0$, we have $\frac{1}{2}(\boldsymbol{\alpha}^*\circ\mathbf{y})'\mathbf{K}_r(\boldsymbol{\alpha}^*\circ\mathbf{y})=\tau^*$. Since $\sum_{r=1}^{\infty}\theta_r=1$, then we can obtain $\frac{1}{2}(\boldsymbol{\alpha}^*\circ\mathbf{y})'\sum_{r=1}^{\infty}\theta_r^*\mathbf{K}_r(\boldsymbol{\alpha}^*\circ\mathbf{y})=\tau^*$. Therefore we have $F(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)=\mathbf{1}'\boldsymbol{\alpha}^*-\frac{1}{2}(\boldsymbol{\alpha}^*\circ\mathbf{y})'\sum_{r=1}^{\infty}\theta_r^*\mathbf{K}_r(\boldsymbol{\alpha}^*\circ\mathbf{y})=\mathbf{1}'\boldsymbol{\alpha}^*-\tau^*=G(\boldsymbol{\alpha}^*,\tau^*)$. ■

It can be observed that Lemma 1 is also satisfied for the multiple kernel learning problem with a finite number of kernels.

Recall that we have denoted the optimal solution of the MKL problem at the r-th iteration as $(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)$. Because there are at most r non-zero elements in $\boldsymbol{\theta}^r$, we assume these non-zero elements are the first r entries in $\boldsymbol{\theta}^r$ for ease of presentation. Then we can represent $(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)=\arg\min_{\boldsymbol{\theta}\in\mathcal{O}_r}\max_{\boldsymbol{\alpha}\in\mathcal{A}}F(\boldsymbol{\alpha},\boldsymbol{\theta})$, where $\mathcal{O}_r=\{\boldsymbol{\theta}|\boldsymbol{\theta}\in\mathcal{D}_{\boldsymbol{\theta}},\theta_i=0,\forall i>r\}$. We rewrite the Theorem 2 in the main text as follows:

Theorem 2. *With Algorithm 1 in the main text, $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)$ monotonically decreases as r increases, and the following inequality holds*

$$F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)\geq F(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)\geq F(\boldsymbol{\alpha}^r,\mathbf{e}_{r+1}),\quad(3)$$

where $\mathbf{e}_{r+1}\in\mathcal{D}_{\boldsymbol{\theta}}$ is the vector with all zeros except the $(r+1)$-th entry being 1. We also have $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)=F(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)=F(\boldsymbol{\alpha}^r,\mathbf{e}_{r+1})$ when Algorithm 1 converges at the r-th iteration.

Proof: We prove it in three steps. First, we prove that $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)$ monotonically decreases as r increases. Then, we show that $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)\geq F(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)\geq F(\boldsymbol{\alpha}^r,\mathbf{e}_{r+1})$ holds. Finally, we prove that $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)=F(\boldsymbol{\alpha}^*,\boldsymbol{\theta}^*)=F(\boldsymbol{\alpha}^r,\mathbf{e}_{r+1})$ when Algorithm 1 converges at the r-th iteration.

To show that $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)$ monotonically decreases as r increases, we only need to prove $F(\boldsymbol{\alpha}^{r+1},\boldsymbol{\theta}^{r+1})\leq F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)$. Recall we have $F(\boldsymbol{\alpha}^r,\boldsymbol{\theta}^r)=\min_{\boldsymbol{\theta}\in\mathcal{O}_r}\max_{\boldsymbol{\alpha}\in\mathcal{A}}F(\boldsymbol{\alpha},\boldsymbol{\theta})=\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\theta}\in\mathcal{O}_r}F(\boldsymbol{\alpha},\boldsymbol{\theta})$, because $F(\boldsymbol{\alpha},\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$ and concave in $\boldsymbol{\alpha}$. Similarly, we have $F(\boldsymbol{\alpha}^{r+1},\boldsymbol{\theta}^{r+1})=\max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\theta}\in\mathcal{O}_{r+1}}F(\boldsymbol{\alpha},\boldsymbol{\theta})$. For each $\boldsymbol{\alpha}$, we have $\min_{\boldsymbol{\theta}\in\mathcal{O}_{r+1}}F(\boldsymbol{\alpha},\boldsymbol{\theta})\leq\min_{\boldsymbol{\theta}\in\mathcal{O}_r}F(\boldsymbol{\alpha},\boldsymbol{\theta})$, because the feasible set $\mathcal{O}_{r+1}\supset\mathcal{O}_r$. Therefore, we have

$F(\boldsymbol{\alpha}^{r+1},\boldsymbol{\theta}^{r+1}) = \max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\theta}\in\mathcal{O}_{r+1}} F(\boldsymbol{\alpha},\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}\in\mathcal{O}_{r+1}} F(\boldsymbol{\alpha}^{r+1},\boldsymbol{\theta}) \leq \min_{\boldsymbol{\theta}\in\mathcal{O}_{r}} F(\boldsymbol{\alpha}^{r+1},\boldsymbol{\theta}) \leq \max_{\boldsymbol{\alpha}\in\mathcal{A}}\min_{\boldsymbol{\theta}\in\mathcal{O}_{r}} F(\boldsymbol{\alpha},\boldsymbol{\theta}) = F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r})$. Thus, we have proved that $F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r})$ monotonically decreases as r increases.

Now we show $F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r}) \geq F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) \geq F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$ holds. The left part of (3) is obvious, since $F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r})$ monotonically decreases. We only need to prove the right part, $F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) \geq F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$. Let us denote $J(\boldsymbol{\alpha},\mathbf{K}) = \frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})'\mathbf{K}(\boldsymbol{\alpha}\circ\mathbf{y})$. Recall $\mathbf{K}_{r+1}$ is obtained by using the most violated constraint in (2), so we have $\mathbf{K}_{r+1} = \arg\max_{\mathbf{K}_\mathbf{M}} J(\boldsymbol{\alpha}^{r},\mathbf{K}_\mathbf{M})$ for any $\mathbf{M}\in\mathcal{M}$. Let us denote $\tilde{\tau} = J(\boldsymbol{\alpha}^{r},\mathbf{K}_{r+1})$, then we have $J(\boldsymbol{\alpha}^{r},\mathbf{K}_\mathbf{M}) \leq J(\boldsymbol{\alpha}^{r},\mathbf{K}_{r+1}) = \tilde{\tau}$ for all $\mathbf{M}\in\mathcal{M}$, which means $(\boldsymbol{\alpha}^{r},\tilde{\tau})$ is a feasible solution to (2). Because $(\boldsymbol{\alpha}^{*},\tau^{*})$ is the optimal solution to (2), we have $G(\boldsymbol{\alpha}^{*},\tau^{*}) \geq G(\boldsymbol{\alpha}^{r},\tilde{\tau}) = \mathbf{1}'\boldsymbol{\alpha}^{r} - \tilde{\tau} = \mathbf{1}'\boldsymbol{\alpha}^{r} - J(\boldsymbol{\alpha}^{r},\mathbf{K}_{r+1}) = \mathbf{1}'\boldsymbol{\alpha}^{r} - \frac{1}{2}(\boldsymbol{\alpha}^{r}\circ\mathbf{y})'\mathbf{K}_{r+1}(\boldsymbol{\alpha}^{r}\circ\mathbf{y}) = F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$. Since $F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) = G(\boldsymbol{\alpha}^{*},\tau^{*})$, we then have $F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) \geq F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$. Thus, we have proved that $F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r}) \geq F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) \geq F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$ holds.

When Algorithm 1 converges at the r-th iteration, it means that we cannot find a feasible $\mathbf{M}$ which violates the constraint in (2). In other words, we have $(\boldsymbol{\alpha}^{r}\circ\mathbf{y})'\mathbf{K}_\mathbf{M}(\boldsymbol{\alpha}^{r}\circ\mathbf{y}) \leq \tau^{r}$ for any $\mathbf{M}\in\mathcal{M}$. Moreover, because $(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r})$ is the optimal solution to the MKL problem at the r-th iteration, we have $\frac{1}{2}(\boldsymbol{\alpha}^{r}\circ\mathbf{y})'\sum_{r}^{\infty}\theta_r\mathbf{K}_r(\boldsymbol{\alpha}^{r}\circ\mathbf{y}) = \tau^{r}$. Therefore, we have $F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1}) = \mathbf{1}'\boldsymbol{\alpha}^{r} - (\boldsymbol{\alpha}^{r}\circ\mathbf{y})'\mathbf{K}_{r+1}(\boldsymbol{\alpha}^{r}\circ\mathbf{y}) \geq \mathbf{1}\boldsymbol{\alpha}^{r} - \tau^{r} = \mathbf{1}\boldsymbol{\alpha}^{r} - \frac{1}{2}(\boldsymbol{\alpha}^{r}\circ\mathbf{y})'\sum_{r}^{\infty}\theta_r\mathbf{K}_r(\boldsymbol{\alpha}^{r}\circ\mathbf{y}) = F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r})$. Recall we have proved that $F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r}) \geq F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) \geq F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$, so we conclude that $F(\boldsymbol{\alpha}^{r},\boldsymbol{\theta}^{r}) = F(\boldsymbol{\alpha}^{*},\boldsymbol{\theta}^{*}) = F(\boldsymbol{\alpha}^{r},\mathbf{e}_{r+1})$ when Algorithm 1 converges. This completes the proof. ■

II. Proof of Proposition 1

In the main text, we have shown that the dual form of our SHFA can be written as follows:

$$\min_{\mathbf{y}\in\mathcal{Y},\mathbf{H}\succeq\mathbf{0}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad -\frac{1}{2}\boldsymbol{\alpha}'(\mathbf{Q}_{\mathbf{H},\mathbf{y}}+\mathbf{D})\boldsymbol{\alpha} \qquad (4)$$
$$\text{s.t.} \quad \text{trace}(\mathbf{H}) \leq \lambda,$$

where $\mathbf{Q}_{\mathbf{H},\mathbf{y}} = \left(\mathbf{K}^{\frac{1}{2}}(\mathbf{H}+\mathbf{I})\mathbf{K}^{\frac{1}{2}}+\mathbf{1}\mathbf{1}'\right)\circ(\mathbf{y}\mathbf{y}') \in \mathbb{R}^{n\times n}$, $\mathbf{y} = [\mathbf{y}_s',\mathbf{y}_t',\mathbf{y}_u']'$ is the label vector in which $\mathbf{y}_s$ and $\mathbf{y}_t$ are given and $\mathbf{y}_u$ is unknown, $\mathcal{Y} = \{\mathbf{y}\in\{-1,+1\}^n | \mathbf{y} = [\mathbf{y}_s',\mathbf{y}_t',\mathbf{y}_u']', \mathbf{1}'\mathbf{y}_u = \delta\}$ is the domain of $\mathbf{y}$, $\boldsymbol{\alpha} = [\alpha_1^s,\ldots,\alpha_{n_s}^s,\alpha_1^t,\ldots,\alpha_{n_t}^t,\alpha_1^u,\ldots,\alpha_{n_u}^u]' \in \mathbb{R}^n$ with α_i^s's, α_i^t's and α_i^u's are the dual variables corresponding to the constraints for source samples, labeled target samples and unlabeled target samples, respectively, $\mathcal{A} = \{\boldsymbol{\alpha}|\boldsymbol{\alpha}\geq\mathbf{0}, \mathbf{1}'\boldsymbol{\alpha} = 1\}$ is the domain of $\boldsymbol{\alpha}$ and $\mathbf{D}\in\mathbb{R}^{n\times n}$ is a diagonal matrix with the diagonal elements as $\frac{1}{C}$ for the labeled data from both domains and $\frac{1}{C_u}$ for the unlabeled target data.

Now we rewrite and prove the Proposition 1 as follows:

Proposition 1. *The objective of (4) is lower-bounded by the optimum of the following optimization problem:*

$$\min_{\boldsymbol{\gamma}\in\mathcal{D}_\gamma,\mathbf{H}\succeq\mathbf{0}}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_l \gamma_l\mathbf{Q}_{\mathbf{H},\mathbf{y}_l}+\mathbf{D}\right)\boldsymbol{\alpha} \qquad (5)$$
$$\textit{s.t.} \quad \textit{trace}(\mathbf{H}) \leq \lambda,$$

where $\mathbf{y}_l$ is the l-th feasible labeling candidates, $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_{|\mathcal{Y}|}]'$, is the coefficient vector for the linear combination of all feasible labeling candidates and $\mathcal{D}_\gamma = \{\boldsymbol{\gamma}|\boldsymbol{\gamma}\geq 0, \mathbf{1}'\boldsymbol{\gamma}\leq 1\}$ is the domain of $\boldsymbol{\gamma}$.

Proof: For ease of presentation, let us define $F(\mathbf{y}) = \min_{\mathbf{H}\succeq 0}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'(\mathbf{Q}_{\mathbf{H},\mathbf{y}}+\mathbf{D})\boldsymbol{\alpha}$ subject to $\text{trace}(\mathbf{H}) \leq \lambda$ and also define $G(\boldsymbol{\gamma}) = \min_{\mathbf{H}\succeq 0}\max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_l \gamma_l\mathbf{Q}_{\mathbf{H},\mathbf{y}_l}+\mathbf{D}\right)\boldsymbol{\alpha}$ subject to $\text{trace}(\mathbf{H}) \leq \lambda$, then the optimal solutions to (4) and (5) can be represented as $\mathbf{y}^{*} = \arg\min_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{y})$ and $\boldsymbol{\gamma}^{*} = \arg\min_{\boldsymbol{\gamma}\in\mathcal{D}_\gamma} G(\boldsymbol{\gamma})$, respectively. Intuitively, we can construct a feasible solution $\boldsymbol{\gamma}_{\mathbf{y}^{*}} = [0,\ldots,0,1,0,\ldots,0]$ where the only non-zero entry corresponds to the optimal feasible labeling $\mathbf{y}^{*}$, and we have $G(\boldsymbol{\gamma}_{\mathbf{y}^{*}}) = F(\mathbf{y}^{*})$. Since $\boldsymbol{\gamma}^{*}$ is the optimal solution to (5), we also have $G(\boldsymbol{\gamma}^{*}) \leq G(\boldsymbol{\gamma}_{\mathbf{y}^{*}}) = F(\mathbf{y}^{*})$, which means the optimal objective of (5) is a always lower than that of (4). ■