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),
\[
\min \max_{\theta \in D_\theta} \alpha - \frac{1}{2}(\alpha \circ y)' \sum_{r=1}^\infty \theta_r K_r(\alpha \circ y),
\]
where \(K_r = K_\frac{1}{r}(\lambda M_{r+1} + I)K_\frac{1}{r}, A = \{\alpha | y'\alpha = 0, 0 \leq \alpha \leq C1\} \) and \(D_\theta = \{\theta | y'\theta \leq 1, \theta \geq 0\}\). As shown in the main text, the dual form of (1) can be written as follows by introducing a dual variable \(\tau\) for \(\theta\):
\[
\max_{\tau, \alpha \in A} \quad 1'\alpha - \tau,
\]
\[
\text{s.t.} \
\frac{1}{2}(\alpha \circ y)'K_r(\alpha \circ y) \leq \tau, \quad \forall r.
\]
Actually, the problem in (1) can also be deemed as the dual form of (2) by considering \(\beta r\) as the dual variable for the \(r\)-th constraint in (2).

In the main text, we have defined
\[
F(\alpha, \theta) = 1'\alpha - \frac{1}{2}(\alpha \circ y)' \sum_{r=1}^\infty \theta_r K_r(\alpha \circ y),
\]
and denoted the optimal solution to (1) as \((\alpha^*, \theta^*) = \arg \min_{\theta \in D_\theta} \max_{\alpha \in A} F(\alpha, \theta)\). Let us denote \(G(\alpha, \tau) = 1'\alpha - \tau\), and the optimal solution to (2) as \((\alpha^*, \tau^*)\). 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 \(F(\alpha^*, \theta^*) = G(\alpha^*, \tau^*)\), and \(\frac{1}{2}(\alpha^* \circ y)' \sum_{r=1}^\infty \theta_r^* K_r(\alpha^* \circ y) = \tau^*\).

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

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 \((\alpha^r, \theta^r)\). Because there are at most \(r\) non-zero elements in \(\theta^r\), we assume these non-zero elements are the first \(r\) entries in \(\theta^r\) for ease of presentation. Then we can represent \((\alpha^r, \theta^r) = \arg \min_{\theta \in O_r} \max_{\alpha \in A} F(\alpha, \theta)\), where \(O_r = \{\theta | \theta \in D_\theta, \theta_i = 0, \forall i > r\}\). We rewrite Theorem 2 in the main text as follows:

**Theorem 2.** With Algorithm 1 in the main text, \(F(\alpha^r, \theta^r)\) monotonically decreases as \(r\) increases, and the following inequality holds
\[
F(\alpha^r, \theta^r) \geq F(\alpha^r, \theta^r) \geq F(\alpha^r, e_{r+1}),
\]
where \(e_{r+1} \in D_\theta\) is the vector with all zeros except the \((r+1)\)-th entry being \(1\).

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

To show that \(F(\alpha^r, \theta^r)\) monotonically decreases as \(r\) increases, we only need to prove \(F(\alpha^{r+1}, \theta^{r+1}) \leq F(\alpha^r, \theta^r)\). Recall we have \(F(\alpha^r, \theta^r) = \min_{\theta \in O_r} \max_{\alpha \in A} F(\alpha, \theta) = \max_{\alpha \in A} \min_{\theta \in O_r} F(\alpha, \theta)\), because \(F(\alpha, \theta)\) is convex in \(\theta\) and concave in \(\alpha\). Similarly, we have \(F(\alpha^{r+1}, \theta^{r+1}) = \max_{\alpha \in A} \min_{\theta \in O_{r+1}} F(\alpha, \theta)\). For each \(\alpha\), we have \(\min_{\theta \in O_{r+1}} F(\alpha, \theta) \leq \min_{\theta \in O_r} F(\alpha, \theta)\), because the feasible set \(O_{r+1} \supset O_r\). Therefore, we have
\( F(\alpha^{r+1}, \theta^{r+1}) = \max_{\alpha \in A} \min_{\theta \in \Theta_{r+1}} F(\alpha, \theta) = \min_{\theta \in \Theta_{r+1}} \max_{\alpha \in A} F(\alpha^{r+1}, \theta) \leq \min_{\theta \in \Theta_{r+1}} \max_{\alpha \in A} F(\alpha, \theta) = F(\alpha^*, \theta^*) \). Thus, we have proved that \( F(\alpha^*, \theta^*) \) monotonically decreases as \( r \) increases.

Now we show \( F(\alpha^*, \theta^*) \geq F(\alpha^*, \theta^r) \) holds. The left part of (3) is obvious, since \( F(\alpha^*, \theta^r) \) monotonically decreases. We only need to prove the right part, \( F(\alpha^*, \theta^r) \geq F(\alpha^*, \theta_{r+1}) \). Let us denote \( J(\alpha, K) = \frac{1}{2} (\alpha \circ y)^T K(\alpha \circ y) \). Recall \( K_{r+1} \) is obtained by using the most violated constraint in (2), so we have \( K_{r+1} = \arg \max_{K \in \mathbb{R}^{n \times n}} J(\alpha^*, K) \) for any \( M \in \mathcal{M} \). Let us denote \( \tilde{\tau} = J(\alpha^*, K_{r+1}) \), then we have \( J(\alpha^*, K_{r+1}) \leq J(\alpha^*, K_{r+1}) = \tilde{\tau} \). For all \( M \in \mathcal{M} \), which means \( (\alpha^*, \tilde{\tau}) \) is a feasible solution to (2). (\( \alpha^*, \tilde{\tau} \)) is the optimal solution to (2), we have \( G(\alpha^*, \tilde{\tau}) \geq G(\alpha^*, \tilde{\tau}) = 1' \alpha^r - \tilde{\tau} = 1' \alpha^r - J(\alpha^*, \tilde{\tau}) = 1' \alpha^r - \frac{1}{2} (\alpha \circ y)^T K_{r+1}(\alpha \circ y) \). Since \( F(\alpha^*, \theta^r) = G(\alpha^*, \theta^r) \), we then have \( F(\alpha^*, \theta^r) \geq F(\alpha^*, \theta^r) \). Thus, we have proved that \( F(\alpha^*, \theta^r) \geq F(\alpha^*, \theta^r) \geq F(\alpha^*, \theta_{r+1}) \) holds.

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

\section{II. PROOF OF PROPOSITION 1}

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

\[
\min_{y \in \gamma, H \succeq 0} \max_{\alpha \in A} \frac{1}{2} \alpha^T (Q_H y + D) \alpha \quad (4)
\]

\text{s.t.} \quad \text{trace}(H) \leq \lambda,

where \( Q_{H, y} = \left( K_{1/2} (H + I) K_{1/2} + \mathbf{1}_1^T \right) \circ (y y^T) \in \mathbb{R}^{n \times n} \), \( y = [y_s, y_t, y_u] \) is the label vector in which \( y_s \) and \( y_t \) are given and \( y_u \) is unknown, \( Y = \{ y \in \{-1, 1\}^n | y = [y_s, y_t, y_u] \} \) is the domain of \( y \), \( \alpha = [\alpha^1_s, \cdots, \alpha^1_t, \cdots, \alpha^u_s, \alpha^u_t, \cdots, \alpha^u_u] \in \mathbb{R}^n \) with \( \alpha^i_j \)'s and \( \alpha^u_j \)'s are the dual variables corresponding to the constraints for source samples, labeled target samples and unlabeled target samples, respectively, \( A = \{ \alpha | \alpha \geq 0, 1'\alpha = 1 \} \) is the domain of \( \alpha \) and \( D \in \mathbb{R}^{n \times n} \) is a diagonal matrix with the diagonal elements as \( \frac{1}{2} \) for the labeled data from both domains and \( \frac{1}{C} \) for the unlabeled target data.

Now we rewrite and prove the Proposition 1 as follows:

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

\[
\min_{\gamma \in \mathbb{D}, \gamma \geq 0} \max_{\alpha \in A} \frac{1}{2} \alpha^T \left( \sum_l \gamma_l Q_{H, y_l} + D \right) \alpha \quad (5)
\]

\text{s.t.} \quad \text{trace}(H) \leq \lambda,

where \( \gamma_l \) is the \( l \)-th feasible labeling candidates, \( \gamma = [\gamma_1, \cdots, \gamma|y|] \), is the coefficient vector for the linear combination of all feasible labeling candidates and \( \mathbb{D}_\gamma = \{ \gamma | \gamma \geq 0, 1'\gamma = 1 \} \) is the domain of \( \gamma \).

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