Supplemental Material to the paper: A Two-Stage Weighting Framework for Multi-Source Domain Adaptation

A: Proof of Lemma 1

Proof. Define $\Phi(S) = \sup_{h\in\mathbb{H}} E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)$. Changing the i-th point in the s-th source affects $\Phi(S)$ by at most $\gamma^s_i = \mu\beta^s\alpha^s_i$, while changing a point in the target affects $\Phi(S)$ by at most $\gamma^s_i = 1/n$ ($s = 0$). Applying McDiarmid's inequality [29] to $\Phi(S)$, the following holds with probability at least $1 - \delta/2$:

$$\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Next, using standard techniques used in [17], we bound the expectation as follows:

$$\begin{aligned}
E_S[\Phi(S)] &= E_S\left[\sup_{h\in\mathbb{H}} E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right] \\
&= E_S\left[\sup_{h\in\mathbb{H}} E_{\bar{S}}[\hat{E}^{\bar{S}}_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)]\right] \\
&\leq E_{S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \hat{E}^{\bar{S}}_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right] \\
&= E_{S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma^s_i\left(\mathcal{L}(h(\bar{x}^s_i), \bar{f}_s(\bar{x}^s_i)) - \mathcal{L}(h(x^s_i), f_s(x^s_i))\right)\right] \\
&= E_{\sigma,S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\sigma^s_i\gamma^s_i\left(\mathcal{L}(h(\bar{x}^s_i), \bar{f}_s(\bar{x}^s_i)) - \mathcal{L}(h(x^s_i), f_s(x^s_i))\right)\right] \\
&\leq 2E_{\sigma,S}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\sigma^s_i\gamma^s_i\mathcal{L}(h(x^s_i), f_s(x^s_i))\right] \leq 2\Re_S(G) = \Re_S(H),
\end{aligned}$$

where the last step follows from the standard techniques for relating the Rademacher complexities [30], and $\mathbb{G}$ is a class of functions given by:

$$\mathbb{G} = \{x \mapsto \mathcal{L}(h'(x), h(x)) : h, h' \in \mathbb{H}\}.$$

Thus, for any $h \in \mathbb{H}$, the following holds with probability at least $1 - \delta/2$:

$$E^S_{\alpha,\beta}(h) \leq \hat{E}^S_{\alpha,\beta}(h) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Similarly, by defining $\Phi'(S) = \sup_{h\in\mathbb{H}} \hat{E}^S_{\alpha,\beta}(h) - E^S_{\alpha,\beta}(h)$ and bounding the expectation of $\Phi'(S)$, we can show that for any $h \in \mathbb{H}$, the following holds with probability at least $1 - \delta/2$:

$$\hat{E}^S_{\alpha,\beta}(h) \leq E^S_{\alpha,\beta}(h) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Thus, with probability at least $1 - \delta$:

$$\left|\hat{E}^S_{\alpha,\beta}(h) - E^S_{\alpha,\beta}(h)\right| \leq \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Next, we bound $\Re_S(H)$ as follows [30]:

$$\begin{aligned}
\Re_S(H) &= E_{S,\sigma}\left[\sup_{h\in\mathbb{H}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s h(x_i^s)| \middle| S=(x_i^s)\right] \\
&= E_{S,\sigma}\left[\sup_{u\in\mathbb{H}_{|S}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s u_i^s| \middle| S=(x_i^s)\right] \\
&= E_{S,\sigma}\left[\sup_{u\in\mathbb{H}_{|S}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s u_i^s| \middle| S=(x_i^s)\right] \\
&\leq E_S\left[\max_{u\in\mathbb{H}_{|S}} ||u||\sqrt{2\log|\mathbb{H}_{|S}|}\right] \text{ (Massart's Lemma [31])} \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} E_S\left[\sqrt{2\log|\mathbb{H}_{|S}|}\right] \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} \sqrt{2\log|\prod_{\mathbb{H}}(m)|} \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} \sqrt{2d\log\frac{em}{d}},
\end{aligned}$$

where $\mathbb{H}_{|S}$ is the restriction of $\mathbb{H}$ on S, $\prod_{\mathbb{H}}(m)$ is the growth function for $\mathbb{H}$ given by the maximum number of ways m points can be classified by $\mathbb{H}$, and e is the natural number. □

B: Proof of Theorem 1

Proof. Let $h^* = \arg\min_{h\in\mathbb{H}}\{\epsilon_T(h)+\epsilon_{\alpha,\beta}(h)\}$. By the triangle inequality, we have

$$\begin{aligned}
|\epsilon_{\alpha,\beta}(h)-\epsilon_T(h)| &\leq |\epsilon_{\alpha,\beta}(h)-\epsilon_{\alpha,\beta}(h,h^*)| + |\epsilon_{\alpha,\beta}(h,h^*)-\epsilon_T(h,h^*)| + |\epsilon_T(h,h^*)-\epsilon_T(h)| \\
&\leq \epsilon_{\alpha,\beta}(h^*) + |\epsilon_{\alpha,\beta}(h,h^*)-\epsilon_T(h,h^*)| + \epsilon_T(h^*) \\
&\leq \lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right).
\end{aligned}$$

Next, we bound $(1+\mu)\epsilon_T(\hat{h})$ as follows:

$$\begin{aligned}
& (1+\mu)\epsilon_T(\hat{h}) \\
\leq\ & \mu\epsilon_{\alpha,\beta}(\hat{h}) + \epsilon_T(\hat{h}) + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & \mu\hat{\epsilon}_{\alpha,\beta}(\hat{h}) + \hat{\epsilon}_T(\hat{h}) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & \mu\hat{\epsilon}_{\alpha,\beta}(h_T^*) + \hat{\epsilon}_T(h_T^*) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & \mu\epsilon_{\alpha,\beta}(h_T^*) + \epsilon_T(h_T^*) + 2\Re_S(H) + 2\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(\lambda_{\alpha,\beta} + \frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & (\mu+1)\epsilon_T(h_T^*) + 2\Re_S(H) + 2\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \mu\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right)
\end{aligned}$$

Thus,

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + \frac{2\Re_S(H)}{1+\mu} + \frac{2}{1+\mu}\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \tag{8}$$

□

Note that our proof follows a similar procedure in [19]. The main differences include (1) we employ the weighted Rademacher complexity, which provides a tighter bound than the one in [19] based on the VC dimension; (2) the empirical minimizer $\hat{h}$ of our joint error function includes two terms involving both source and target domain data with a differential weight μ, while the one in [19] involves one term only. For the special case when $\mu = 1$ and α_i^s's are given a uniform weight, i.e., $\alpha_i^s = 1/n_s$, our bound in (7) is strictly tighter than the one in [19] (due to the $1/2$ factor in the last term). In the general case with different choices of μ and α_i^s's, our bound can be further improved.

C: More details on the datasets and parameters used for the implementation of different methods

The statistics of the test datasets used is summarized in Table 2.

Dataset	Number of domains	Dimension	Number of classes
20 Newsgroups	13	100	2
Sentiment Analysis	4	200	2
Surface Electromyogram (SEMG)	8	12	4

Table 2: Statistics of the test datasets

The categories of 20 Newsgroups dataset that were used in the experiments as source and target domains are as listed in Table 3.

20 Newsgroups	
Categories	inst
comp.os.ms-windows.misc	100
comp.sys.ibm.pc.hardware	100
comp.sys.mac.hardware	98
comp.windows.x	100
rec.motorcycles	100
rec.sport.baseball	100
rec.sport.hockey	100
sci.electronics	100
sci.med	100
sci.space	100
talk.politics.miseast	94
talk.politics.misc	78
talk.religion.misc	64

Table 3: Summary of categories (domains)

A Gaussian kernel with $\sigma = 10$ was used to compute the α values for each source. The weighted hypothesis for each source was learned using Support Vector Machines implemented in the LibSVM package, with a linear kernel and a regularization penalty $C = 10$. The β weights were computed based on a binary similarity matrix, i.e., $W_{ij} = 0$ if the i-th data point is among the N nearest neighbors of the j-th data point or the j-th data point is among the N nearest neighbors of the i-th data point; we set $N = 10$. We implemented TCA with a linear kernel and KMM with a Gaussian kernel as they gave the best results. All parameters were tuned using 10-fold cross-validation.

D: Additional empirical results

Figure 3 shows the α-weighted data samples in both source domain D1 and source domain D2 of the toy data shown in Figure 1. We observe that data samples having similar marginal probabilities in both the domains get higher weight, shown by the size of the points. The size of the points are proportional to their weights. We also observe that since at this stage the source data is re-weighted based only on marginal probability distribution difference, hence some of the data samples from source domain D1 having conflicting conditional probabilities with target domain data also get higher weight as they share similar marginal probability distributions.

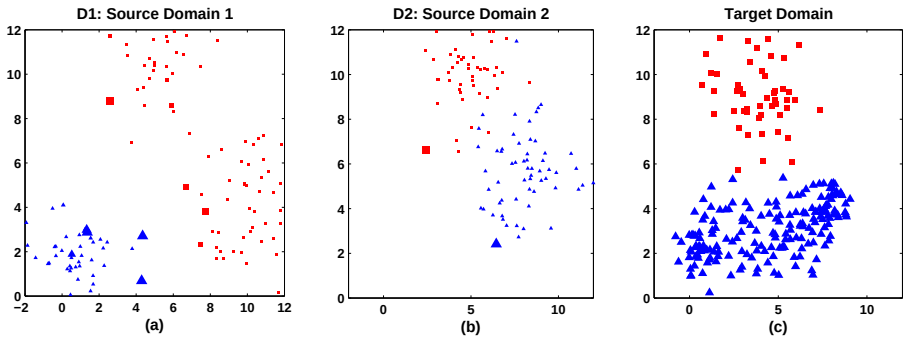

Figure 3: Data samples in source domains D1 and D2 re-weighted by α_i^s. We can observe that points from source domain D1 also get large weights due to the similarity in marginal probabilities (the size of a point is proportional to its weight).

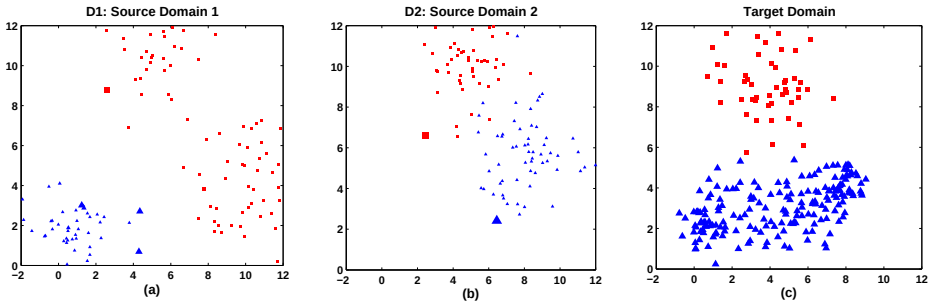

Figure 4: Data samples in the source domains D1 and D2 re-weighted by both α_i^s and β^s. We observe that the points with conflicting conditional probabilities get moderated by β^s (the size of a point is proportional to its weight).

Figure 4 shows the results of applying β-weights to the data samples in both source domain D1 and source domain D2 of the toy data. We observe that the data samples in source domain D1 with conflicting conditional probabilities get reduced when moderated with β weights, as source domain D2 is more similar to target data in conditional probability distribution than the source domain D1.

Figure 5 shows the performance of 2SW-MDA on toy dataset shown in Figure 1 with varying μ. The result is consistent with the theoretical result established in this paper.

Figure 6 shows the results of applying the proposed 2SW-MDA method on another set of toy dataset consisting of two source domains and a target domain with different marginal and conditional probability differences. We observe that the distribution D1 which has conflicting conditional probabilities with target domain data gets under-weighted by the proposed weighting scheme and hence transfer happens mostly from the source distribution D2, which shares similar marginal and conditional probability differences with the target domain. We get β value of 0.17 for D1 and 0.83 for D2.

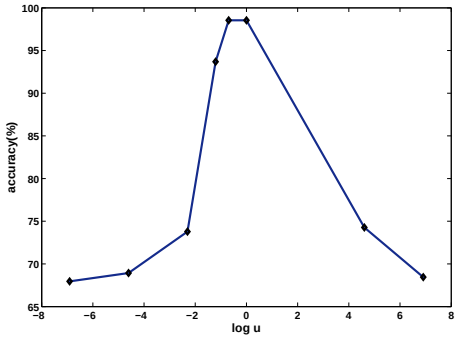

Figure 5: Performance of proposed 2SW-MDA method on the toy dataset shown in Figure 1 with varying μ - Accuracy (%).

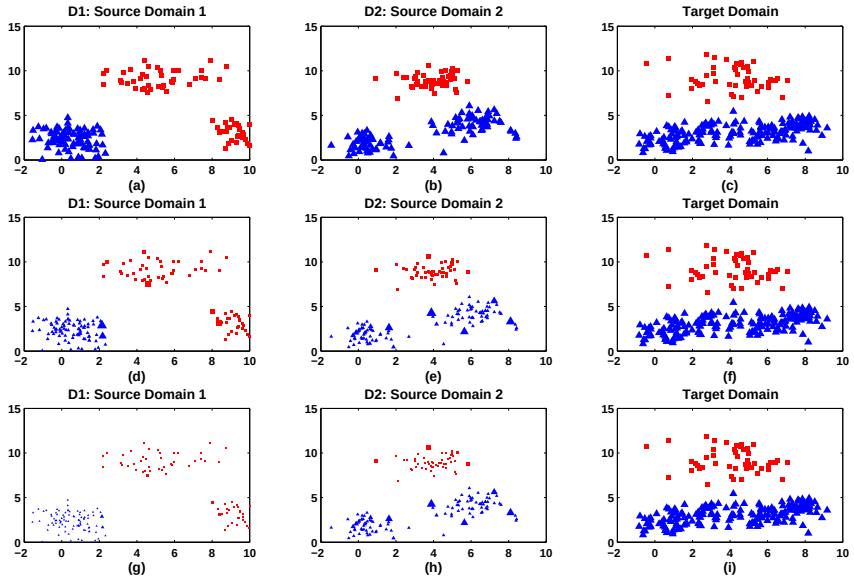

Figure 6: Results on another toy dataset: First row shows the original distribution of two source domains D1 and D2 and a target domain. The second and third rows show the results of applying α and β weights, respectively. We observe that source domain data samples with similar marginal and conditional probabilities get higher weight. The β values for D1 and D2 are 0.17 and 0.83 respectively, individual accuracies being 61.65% and 89.51% and proposed method gives 98.51%.