Supplementary Material for "STEM: An approach to Multi-source Domain Adaptation with Guarantees"

This is the supplementary material for "*STEM: An approach to Multi-source Domain Adaptation with Guarantees*". In the first section of this supplementary material, we provide proof for our theory developed, while presenting implementation specification and additional experimental results in the second section.

1 Theoretical Developments

Theorem 1. *If ℓ is a convex function, the following statements hold true:*

i) $\mathcal{L}\left(h^S, \mathbb{D}^S_{\boldsymbol{\pi}}\right) \leq \max_{1\leq k\leq K} \mathcal{L}\left(h^S_k, \mathbb{D}^S_k\right)$.

ii) If each individual expert is an ϵ-qualified classifier (i.e., $\mathcal{L}\left(h^S_k, \mathbb{D}^S_k\right) \leq \epsilon$), the multi-source teacher expert h^S is also an ϵ-qualified classifier (i.e., $\mathcal{L}\left(h^S, \mathbb{D}^S_{\boldsymbol{\pi}}\right) \leq \epsilon$).

Proof. i) We have

$$\begin{aligned}
\mathcal{L}\left(h^S, \mathbb{D}^S_{\boldsymbol{\pi}}\right) &= \int_{\mathcal{X}\times\mathcal{Y}} \ell\left(h^S(\mathbf{x}, y)\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y)\, d\mathbf{x}dy \\
&\leq \int_{\mathcal{X}\times\mathcal{Y}} \ell\left(\sum_{k=1}^K \frac{\pi_k p^S_k(\mathbf{x}, y)}{\sum_{j=1}^K \pi_j p^S_j(\mathbf{x}, y)} h^S_k(\mathbf{x}, y)\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y)\, d\mathbf{x}dy \\
&\leq \int_{\mathcal{X}\times\mathcal{Y}} \sum_{k=1}^K \frac{\pi_k p^S_k(\mathbf{x}, y)}{\sum_{j=1}^K \pi_j p^S_j(\mathbf{x}, y)} \ell\left(h^S_k(\mathbf{x}), y\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y)\, d\mathbf{x}dy \\
&\leq \sum_{k=1}^K \pi_k \int_{\mathcal{X}\times\mathcal{Y}} \frac{p^S_k(\mathbf{x}, y)}{p^S_{\boldsymbol{\pi}}(\mathbf{x}, y)} \ell\left(h^S_k(\mathbf{x}), y\right) p^S_{\boldsymbol{\pi}}(\mathbf{x}, y)\, d\mathbf{x}dy \\
&= \sum_{k=1}^K \pi_k \int_{\mathcal{X}\times\mathcal{Y}} \ell\left(h^S_k(\mathbf{x}), y\right) p^S_k(\mathbf{x}, y)\, d\mathbf{x}dy \\
&= \sum_{k=1}^K \pi_k \mathcal{L}\left(h^S_k, \mathbb{D}^S_k\right) \leq \max_{1\leq k\leq K} \mathcal{L}\left(h^S_k, \mathbb{D}^S_k\right).
\end{aligned}$$

Note that $\ell\left(h^S_k(\mathbf{x}), y\right) := \ell\left(h^S_k(\mathbf{x}, y)\right)$ where $h^S_k(\mathbf{x}) = \left[h^S_k(\mathbf{x}, y)\right]_{y=1}^M$ and $\ell(t) = -\log(t)$ for the cross-entropy loss.

ii) It is trivial from (i). □

As indicated by Theorem 1, the multi-source teacher expert h^Spredicts well data examples sampled from $\mathcal{D}^S_{\boldsymbol{\pi}}$ which is a mixture of $\mathcal{D}^S_{1:K}$. It is natural to ask the question of the factors that influence the performance of h^S when *predicting on the target joint distribution* $\mathbb{D}^T$. To facilitate the following theorem, we define a *hybrid joint distribution* $\mathbb{D}^h$ with the density function $p^h(\mathbf{x}, y)$ consisting of the pairs $(\mathbf{x}, y)$ in which the data example $\mathbf{x} \sim \mathbb{P}^S_{\boldsymbol{\pi}} = \sum_{k=1}^K \pi_k \mathbb{P}^S_k$ and the label y is sampled using $p^T(y \mid \mathbf{x})$:

$$p^h(\mathbf{x}, y) = p^T(y \mid \mathbf{x}) \sum_{k=1}^K \pi_k p^S_k(\mathbf{x}) = p^T(y \mid \mathbf{x})\, p^S_{\boldsymbol{\pi}}(\mathbf{x}).$$

Theorem 3. *If ℓ is a convex function and upper-bounded by a positive constant L, the general loss $\mathcal{L}\left(h^S, \mathbb{D}^T\right)$ is upper-bounded by:*

i) $A\left[\max_k \mathcal{L}\left(h_k^S, \mathbb{D}_k^S\right) + L \max_k \mathbb{E}_{\mathbb{P}_k^S}\left[\|\Delta p_k\left(y \mid \mathbf{x}\right)\|_1\right]\right]^{\frac{\alpha-1}{\alpha}}$ *where* $A = \exp\left\{R^\alpha\left(\mathbb{P}^T \| \mathbb{P}_{\boldsymbol{\pi}}^S\right)\right\}^{\frac{\alpha-1}{\alpha}} L^{\frac{1}{\alpha}}$ *in which* $R^\alpha\left(\mathbb{P}^T \| \mathbb{P}_{\boldsymbol{\pi}}^S\right)$ *represents the Rényi divergence between those distributions and* $\Delta p_k\left(y \mid \mathbf{x}\right) := \left[\left|p_k^S\left(y = m \mid \mathbf{x}\right) - p^T\left(y = m \mid \mathbf{x}\right)\right|\right]_{m=1}^M$ *represents the label shift between the labeling assignment mechanisms of an individual source domain and target domain.*

ii) $A\left[\epsilon + L \max_k \mathbb{E}_{\mathbb{P}_k^S}\left[\|\Delta p_k\left(y \mid \mathbf{x}\right)\|_1\right]\right]^{\frac{\alpha-1}{\alpha}}$ *provided that* $\mathcal{L}\left(h_k^S, \mathbb{D}_k^S\right) \leq \epsilon, \forall k = 1, ..., K$.

Proof. i) We have

$$\begin{aligned}
\mathcal{L}\left(h^S, \mathbb{D}^h\right) &= \int \ell\left(h^S\left(\mathbf{x}\right), y\right) p^h\left(\mathbf{x}, y\right) d\mathbf{x}dy = \mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) + \int \ell\left(h^S\left(\mathbf{x}\right), y\right)\left[p^h\left(\mathbf{x}, y\right) - p_{\boldsymbol{\pi}}^S\left(\mathbf{x}, y\right)\right] d\mathbf{x}dy \\
&\leq \mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) + \int \ell\left(h^S\left(\mathbf{x}\right), y\right)\left|p^h\left(\mathbf{x}, y\right) - p_{\boldsymbol{\pi}}^S\left(\mathbf{x}, y\right)\right| d\mathbf{x}dy \\
&\leq \mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) + L \int \left|p^h\left(\mathbf{x}, y\right) - p_{\boldsymbol{\pi}}^S\left(\mathbf{x}, y\right)\right| d\mathbf{x}dy \\
&= \mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) + L \int \sum_{k=1}^K \pi_k p_k^S\left(\mathbf{x}\right)\left|p^T\left(y \mid \mathbf{x}\right) - p_k^S\left(y \mid \mathbf{x}\right)\right| d\mathbf{x}dy \\
&= \mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) + L \sum_{k=1}^K \pi_k \int p_k^S\left(\mathbf{x}\right) \sum_{y=1}^M \left|p^T\left(y \mid \mathbf{x}\right) - p_k^S\left(y \mid \mathbf{x}\right)\right| d\mathbf{x} \\
&= \mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) + L \sum_{k=1}^K \pi_k \mathbb{E}_{\mathbb{P}_k^S}\left[\|\Delta p_k\left(y \mid \mathbf{x}\right)\|_1\right] \\
&\leq \max_k \mathcal{L}\left(h_k^S, \mathbb{D}_k^S\right) + L \max_k \mathbb{E}_{\mathbb{P}_k^S}\left[\|\Delta p_k\left(y \mid \mathbf{x}\right)\|_1\right].
\end{aligned}$$

Note that we have used $\mathcal{L}\left(h^S, \mathbb{D}_{\boldsymbol{\pi}}^S\right) \leq \max_k \mathcal{L}\left(h_k^S, \mathbb{D}_k^S\right)$ (referred to Theorem 1).
Finally, we manipulate $\mathcal{L}\left(h^S, \mathbb{D}^T\right)$ as

$$\begin{aligned}
\mathcal{L}\left(h^S, \mathbb{D}^T\right) &= \int_{\mathcal{X}\times\mathcal{Y}} \ell\left(h^S\left(\mathbf{x}\right), y\right) p^T\left(\mathbf{x}, y\right) d\mathbf{x}dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p^T\left(\mathbf{x}, y\right)}{p^h\left(\mathbf{x}, y\right)^{\frac{\alpha-1}{\alpha}}} p^h\left(\mathbf{x}, y\right)^{\frac{\alpha-1}{\alpha}} \ell\left(h^S\left(\mathbf{x}\right), y\right) d\mathbf{x}dy.
\end{aligned}$$

The Holder inequality gives us

$$\mathcal{L}\left(h^S, \mathbb{D}^T\right) \leq \left[\int_{\mathcal{X}\times\mathcal{Y}} \frac{p^T\left(\mathbf{x}, y\right)^\alpha}{p^h\left(\mathbf{x}, y\right)^{\alpha-1}} d\mathbf{x}dy\right]^{\frac{1}{\alpha}} \left[\int_{\mathcal{X}\times\mathcal{Y}} p^h\left(\mathbf{x}, y\right) \ell\left(h^S\left(\mathbf{x}\right), y\right)^{\frac{\alpha}{\alpha-1}} d\mathbf{x}dy\right]^{\frac{\alpha-1}{\alpha}}.$$

Referring to the definition of the Rényi divergence and note that $\ell\left(h^S\left(\mathbf{x}\right), y\right) \leq L$, we obtain

$$\mathcal{L}\left(h^S, \mathbb{D}^T\right) \leq \left[\exp\left\{R^\alpha\left(\mathbb{D}^T \| \mathbb{D}^h\right)\right\} \mathcal{L}\left(h^S, \mathbb{D}^h\right)\right]^{\frac{\alpha-1}{\alpha}} L^{\frac{1}{\alpha}}.$$

We further derive

$$\begin{aligned}
R^\alpha\left(\mathbb{D}^T \| \mathbb{D}^h\right) &= \frac{1}{\alpha-1} \log \int \left[\frac{p^T\left(\mathbf{x}, y\right)}{p^h\left(\mathbf{x}, y\right)}\right]^{\alpha-1} p^T\left(\mathbf{x}, y\right) d\mathbf{x}dy \\
&= \frac{1}{\alpha-1} \log \left(\int \left[\frac{p^T\left(\mathbf{x}\right)}{p_{\boldsymbol{\pi}}^S\left(\mathbf{x}\right)}\right]^{\alpha-1} \sum_{y=1}^M p^T\left(\mathbf{x}, y\right) d\mathbf{x}\right) \\
&= \frac{1}{\alpha-1} \log \left(\int \left[\frac{p^T\left(\mathbf{x}\right)}{p_{\boldsymbol{\pi}}^S\left(\mathbf{x}\right)}\right]^{\alpha-1} p^T\left(\mathbf{x}\right) d\mathbf{x}\right) = R^\alpha\left(\mathbb{P}^T \| \mathbb{P}_{\boldsymbol{\pi}}^S\right).
\end{aligned}$$

Finally, we reach the following inequality:

$$\mathcal{L}\left(h^S, \mathbb{D}^T\right) \leq \left[\exp\left\{R^\alpha\left(\mathbb{P}^T \| \mathbb{P}^S_{\boldsymbol{\pi}}\right)\right\} \mathcal{L}\left(h^S, \mathbb{D}^h\right)\right]^{\frac{\alpha-1}{\alpha}} L^{\frac{1}{\alpha}}.$$

ii) It is trivial from (i) and $\mathcal{L}\left(h^S_k, \mathbb{D}^S_k\right) \leq \epsilon, \forall k = 1, ..., K$. □

In what follows, we present how to train the multi-source teacher expert h^S. Our workaround to train h^S comes from the following theoretical observation. Assume that we have K distributions $\mathbb{R}_{1:K}$ with density functions $r_{1:K}(\mathbf{z})$. We form a joint distribution $\mathcal{D}$ of a data instance $\mathbf{z}$ and label $t \in \{1, ..., K\}$ by sampling an index $t \sim \text{Cat}(\boldsymbol{\pi})$, sampling $\mathbf{x} \sim \mathbb{R}_t$, and collecting $(\mathbf{z}, t)$ as a sample from $\mathcal{D}$. With this setting, we have the following corollary.

Corollary 2. *If we train a source domain discriminator $\mathcal{C}$ to classify samples from the joint distribution $\mathcal{D}$ using the cross-entropy loss (i.e., $CE(\cdot, \cdot)$), the optimal source domain discriminator $\mathcal{C}^*$ defined as*

$$\mathcal{C}^* = \textit{argmin}_{\mathcal{C}} \mathbb{E}_{(\mathbf{z},t)\sim\mathcal{D}}\left[CE\left(\mathcal{C}\left(\mathbf{z}\right), t\right)\right]$$

satisfies $\mathcal{C}^*(\mathbf{z}) = \left[\frac{\pi_i r_i(\mathbf{z})}{\sum_j \pi_j r_j(\mathbf{z})}\right]_{i=1}^K$.

Proof. We have

$$\begin{aligned}\mathbb{E}_{(\mathbf{z},t)\sim\mathcal{D}}\left[CE\left(\mathcal{C}\left(\mathbf{z}\right), t\right)\right] &= \sum_{t=1}^K \pi_t \int CE\left(\mathcal{C}(\mathbf{z}), t\right) r_t\left(\mathbf{z}\right) d\mathbf{z} \\ &= -\int \sum_{t=1}^K \log \mathcal{C}\left(\mathbf{z}, t\right) \pi_t r_t\left(\mathbf{z}\right) d\mathbf{z}.\end{aligned}$$

Given $\mathbf{z}$, we now find $\mathcal{C}^* = [\mathcal{C}^*_t]_{t=1}^K$ subjected to $\|\mathcal{C}^*\|_1 = 1$ and $\mathcal{C}^* \geq \mathbf{0}$ to maximize

$$\max_{\mathcal{C}:\|\mathcal{C}\|_1=1} \sum_{t=1}^K \log \mathcal{C}_t \pi_t r_t\left(\mathbf{z}\right).$$

The Lagrange function is as follows:

$$\mathcal{L} = \sum_{t=1}^K \log \mathcal{C}_t \pi_t r_t\left(\mathbf{z}\right) - \lambda\left(\sum_{t=1}^K \mathcal{C}_t - 1\right).$$

Setting the derivatives to 0, we obtain

$$\frac{\partial \mathcal{L}}{\partial \mathcal{C}_t} = \frac{\pi_t r_t\left(\mathbf{z}\right)}{C_t} - \lambda = 0,\ t = 1, ..., K.$$

Note that $\sum_{t=1}^K \mathcal{C}_t = 1$, we arrive at

$$\mathcal{C}^*_t = \frac{\pi_t r_t\left(\mathbf{z}\right)}{\sum_j \pi_j r_j\left(\mathbf{z}\right)},\ t = 1, ..., K.$$

Finally, we reach the conclusion. □

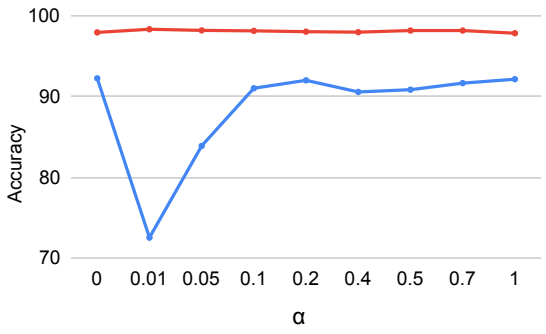
2 Implementation Specification and Additional Experimental Results

2.1 Data preparation and preprocessing

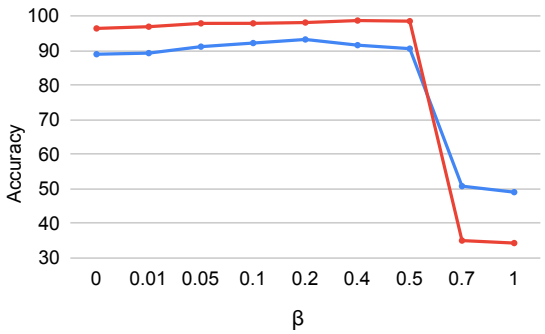
Digits-five. The resolution of digit images is resized to 32×32, and we normalize pixel values to the range of $[-1, 1]$.
Office-Caltech10 and DomainNet. We resize the resolution of images to 224×224, and then use those preprocessing images for ResNet-101 [1]. Additionally, for DomainNet, we find that adaptation tasks are much challenging due to the dissimilarity across domains. To reduce the domain gap, we apply the horizontal flip transformation to increase data diversity during the training process.

Architecture	Digit-five	Office-Caltech10	DomainNet
Input size	$32 \times 32 \times 3$	$224 \times 224 \times 3$	$224 \times 224 \times 3$
Generator G	instance normalization 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 2×2 max-pool, stride 2 dropout, $p = 0.5$ Gaussian noise, σ = 1 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 2×2 max-pool, stride 2 dropout, $p = 0.5$ Gaussian noise, σ = 1	ResNet101 256 dense, ReLU dropout, $p = 0.5$ Gaussian noise, σ = 1	transformation ResNet101
Classifier $h^S_{1:K}, h^T$	3×3 conv. 64 lReLU 3×3 conv. 64 lReLU 3×3 conv. 64 lReLU global average pool M dense, softmax	M dense, softmax	M dense, softmax
Source domain discriminator C	100 dense, ReLU K dense, ReLU	K dense, ReLU	K dense, ReLU

Table 1: Network architecture of STEM. The Leaky ReLU (lReLU) parameter a is set to 0.1. K and M denote the number of source domains and the number of classes respectively.

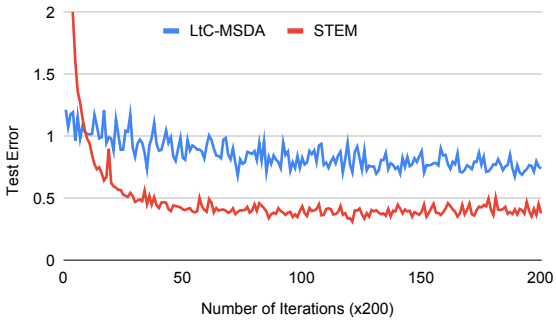


(a) α ($\beta = 0.1$)

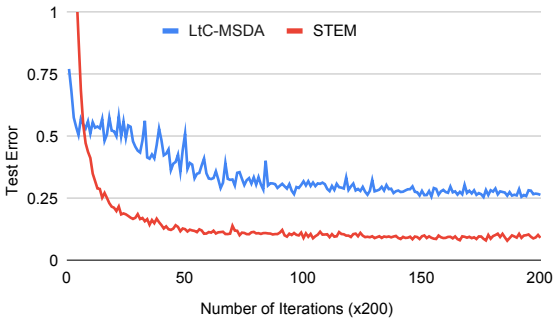


(b) β ($\alpha = 1.0$)

Figure 1: Test accuracy (%) when tweaking α and β on "→**sv**" (*blue*) and "→**sy**" (*red*) tasks.



(a) "→**sv**" task



(b) "→**sy**" task

Figure 2: The comparison of test error of our STEM (*red*) with the current state-of-the-art LtC-MSDA (*blue*)

2.2 Architecture

The network architecture in detail for each dataset is shown in Table 1. We use a small convolutional neural network for Digit-five since this dataset is easy to be converged. For the two other datasets, we both use Resnet-101 pre-trained on ImageNet as the backbone, which is frozen when training on Office-Caltech10 and fine-tuned on DomainNet. All experiments are run on a computer with an NVIDIA Tesla V100 SXM2 with 16 GB memory.

2.3 Additional ablation study

In section 3.6.5 of the main paper, we update our network by minimizing loss funtion:

$$\sum_{k=1}^{K} \mathcal{L}_k^{ie} + \alpha \mathcal{L}^{\mathcal{C}} + \mathcal{L}^m + \beta \mathcal{L}^{clus} - \gamma \mathcal{L}^d. \tag{1}$$

These two parameters α and β help to discriminate source domains in latent space and ensure the clustering assumption [2], respectively. Figure 1 depicts the model performance with a diverse range of α and β on "→**sv**" and "→**sy**" tasks. Following the result, β significantly affects both tasks in $[0.5, 1]$ and less sensitive in $[0, 0.5]$, while adjusting α gives a stable performance in most cases. After searching their values in the range of $[0, 1]$, we observe that our network achieves good performances with $\alpha = 1.0$ and $\beta = 0.1$.

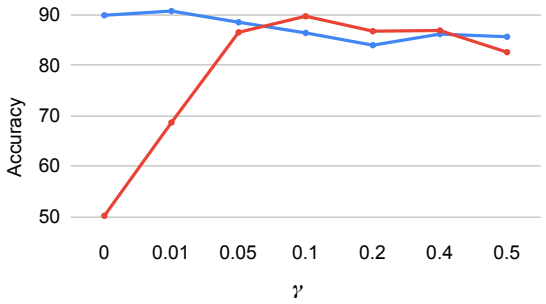


Figure 3: Effect of GAN relevant trade-off parameter γ to the performance on the →**sv** (*blue*) and →**mm** (*red*) tasks.

GAN components (i.e., generator and discriminator) play a vital role in our framework to diminish the data distribution discrepancy of the target domain and mixture of source domains in the latent space. We conduct experiments to study the effect of the parameter γ on the transferring performance. We experiment on the →**sv** and →**mm** task by varying γ in $[0, 0.5]$. As shown in Figure 3, on the →**mm** task, γ quite significantly influences the performance, while for the case of →**sv** task, it stably affects the performance. This totally depends on how distant the target domain and source domains in the latent space. We empirically find that $\gamma = 0.1$ works satisfactorily for most of the cases.

2.4 Convergence

We testify the convergence of our STEM with the test error on "→**sv**" and "→**sy**" tasks and compare our proposal with the state-of-the-art method named LtC-MSDA [3]. The comparison is presented in Figure 2. Following the result, our STEM enjoys faster and stable convergence in which the test error is significantly lower than LtC-MSDA method.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[2] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *AISTATS*, volume 2005, pages 57–64. Citeseer, 2005.

[3] H. Wang, M. Xu, B. Ni, and W. Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *Computer Vision – ECCV*, 2020.