Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] We clearly state the assumptions we use in Section 2.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We have done a theoretic work that has few societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We think the error bars are not related to the core result of our experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We use few computation resources in our work.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [Yes] See Section 4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We don't use new assets.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 4.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Loss Function Leading to (2)

To obtain (2), first note that for each $i = 0, 1$, we can define the loss of task i as

$$L_i = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \hat{P}^{(i)}_{XY}(x,y) \log \frac{1}{Q_{XY}(x,y)}, \tag{19}$$

where Q_{XY} represents the distribution model and $\hat{P}^{(i)}_{XY}$ are the empirical distributions as defined in (1). As a result, training the linear combination of L_0 and L_1 can lead to the convex combination model.

$$\begin{aligned}
&\arg\min_{Q_{XY}\in\mathcal{P}} \ \alpha_0 L_0 + \alpha_1 L_1 \\
&= \arg\min_{Q_{XY}\in\mathcal{P}} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \left(\alpha_0 \hat{P}^{(0)}_{XY}(x,y) + \alpha_1 \hat{P}^{(1)}_{XY}(x,y)\right) \log \frac{1}{Q_{XY}(x,y)} \\
&= \arg\min_{Q_{XY}\in\mathcal{P}} \ D\left(\alpha_0 \hat{P}^{(0)}_{XY} + \alpha_1 \hat{P}^{(1)}_{XY} \middle\| Q_{XY}\right) \\
&= \alpha_0 \hat{P}^{(0)}_{XY} + \alpha_1 \hat{P}^{(1)}_{XY}.
\end{aligned} \tag{20}$$

B Proof of Theorem 2

To compute the testing loss, we first introduce the following lemma.

Lemma 6. *Suppose that random vector $(X_1, X_2, \cdots, X_m)$ follows the multinomial distribution with the corresponding event probabilities $(p_1, p_2, \cdots, p_m)$ and n independent trials, then the variance of random variable X_i is*

$$\mathrm{var}(X_i) = np_i(1-p_i), \tag{21}$$

and the covariance of X_i and X_j is

$$\mathrm{cov}(X_i, X_j) = -np_ip_j. \tag{22}$$

From Lemma 6, for $i = 0, \cdots, k$, we have

$$\mathbb{E}\left[\left(\hat{P}^{(i)}_{XY}(x,y) - P^{(i)}_{XY}(x,y)\right)^2\right] = \frac{1}{n_i} P^{(i)}_{XY}(x,y)(1 - P^{(i)}_{XY}(x,y)). \tag{23}$$

In turn, the testing loss as defined in (3) is

$$\begin{aligned}
L^{(\alpha_0,\alpha_1)}_{\text{test}} &= \mathbb{E}\left[\chi^2(P^{(0)}_{XY}, \alpha_0\hat{P}^{(0)}_{XY} + \alpha_1\hat{P}^{(1)}_{XY})\right] \\
&= \mathbb{E}\left[\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{1}{P^{(0)}_{XY}(x,y)} \left(P^{(0)}_{XY}(x,y) - \alpha_0\hat{P}^{(0)}_{XY}(x,y) - \alpha_1\hat{P}^{(1)}_{XY}(x,y)\right)^2\right] \\
&= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{\left(P^{(0)}_{XY}(x,y) - \alpha_0 P^{(0)}_{XY}(x,y) - \alpha_1 P^{(1)}_{XY}(x,y)\right)^2}{P^{(0)}_{XY}(x,y)} \\
&\quad + \alpha_0^2 \mathbb{E}\left[\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{\left(\hat{P}^{(0)}_{XY}(x,y) - P^{(0)}_{XY}(x,y)\right)^2}{P^{(0)}_{XY}(x,y)}\right] \\
&\quad + \alpha_1^2 \mathbb{E}\left[\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{\left(\hat{P}^{(1)}_{XY}(x,y) - P^{(1)}_{XY}(x,y)\right)^2}{P^{(0)}_{XY}(x,y)}\right]
\end{aligned} \tag{24}$$

$$= \alpha_1^2 \chi^2\left(P^{(0)}_{XY}, P^{(1)}_{XY}\right) + \alpha_0^2 \frac{1}{n_0} V^{(0)} + \alpha_1^2 \frac{1}{n_1} V^{(1)}, \tag{25}$$

where to obtain (24) we have used the facts that $\hat{P}_{XY}^{(0)}$ and $\hat{P}_{XY}^{(1)}$ are independent, $\mathbb{E}[\hat{P}_{XY}^{(0)}] = P_{XY}^{(0)}$, and $\mathbb{E}[\hat{P}_{XY}^{(1)}] = P_{XY}^{(1)}$, and where to obtain (25) we have used (23) and the fact that $n_0 \cdot \hat{P}_{XY}^{(0)}$ and $n_1 \cdot \hat{P}_{XY}^{(1)}$ follows the multinomial distribution.

Similarly, we can obtain Theorem 3.

C Proof of Theorem 4

C.1 Expression of $\hat{\boldsymbol{g}}_i$ as defined in (13)

From

$$\begin{aligned}
&\chi^2_{R_{XY}}\left(\hat{P}_{XY}^{(i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}\right) \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{1}{P_X^{(0)}(x)P_Y^{(0)}(y)} \cdot \left(\hat{P}_{XY}^{(i)}(x,y) - P_X^{(0)}(x)P_Y^{(0)}(y) - P_X^{(0)}(x)P_Y^{(0)}(y)\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y)\right)^2,
\end{aligned}$$

we have, for all $y' \in \mathcal{Y}$,

$$\begin{aligned}
&\frac{\partial \chi^2_{R_{XY}}\left(\hat{P}_{XY}^{(i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}\right)}{\partial \boldsymbol{g}(y')} \qquad (26)\\
&= -2\sum_{x\in\mathcal{X}}\left[\hat{P}_{XY}^{(i)}(x,y') - P_X^{(0)}(x)P_Y^{(0)}(y')\right]\boldsymbol{f}(x) + 2\sum_{x\in\mathcal{X}} P_X^{(0)}(x)P_Y^{(0)}(y')\left(\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y')\right)\boldsymbol{f}(x)\\
&= -2\sum_{x\in\mathcal{X}}\hat{P}_{XY}^{(i)}(x,y')\boldsymbol{f}(x) + 2P_Y^{(0)}(y')\boldsymbol{\Lambda}_{\boldsymbol{f}}\boldsymbol{g}(y') \qquad (27)
\end{aligned}$$

where to obtain the last equality, we have used the assumption that $\mathbb{E}_{P_X^{(0)}}[\boldsymbol{f}(X)] = \mathbf{0}$ and the notation that $\boldsymbol{\Lambda}_{\boldsymbol{f}} \triangleq \mathbb{E}_{P_X^{(0)}}[\boldsymbol{f}(X)\boldsymbol{f}^{\mathrm{T}}(X)]$.

Set the gradient (27) to zero, and we obtain

$$\hat{\boldsymbol{g}}_i(y) = \frac{1}{P_Y^{(0)}(y)}\boldsymbol{\Lambda}_{\boldsymbol{f}}^{-1}\left(\sum_{x\in\mathcal{X}}\hat{P}_{XY}^{(i)}(x,y)\boldsymbol{f}(x)\right). \qquad (28)$$

C.2 Proof of (16)

We first express the testing error (15) as

$$\begin{aligned}
L_{\mathrm{test}}^{(\boldsymbol{\alpha})} &= \mathbb{E}\left[\chi^2_{R_{XY}}\left(P_{XY}^{(0)}, P_X^{(0)}Q_{Y|X}^{(\boldsymbol{\alpha})}\right)\right]\\
&= \chi^2_{R_{XY}}\left(P_{XY}^{(0)}, \sum_{i=0}^{k}\alpha_i P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}\right) + \sum_{i=0}^{k}\alpha_i^2 \underbrace{\mathbb{E}\left[\chi^2_{R_{XY}}\left(P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\hat{\boldsymbol{g}}_i)}\right)\right]}_{\frac{1}{n_i}\tilde{V}^{(i)}},
\end{aligned} \qquad (29)$$

where to obtain the last equality we have used the fact that the empirical distributions $\hat{P}_{XY}^{(i)}$ $(i = 0, \cdots, k)$ are independent and

$$\mathbb{E}\left[\tilde{P}_{Y|X}^{(\boldsymbol{f},\hat{\boldsymbol{g}}_i)}\right] = \mathbb{E}\left[\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}\right].$$

Next, for the terms in (29), we have

$$\begin{aligned}
&\chi^2_{R_{XY}}\left(P_{XY}^{(0)}, \sum_{i=0}^{k}\alpha_i P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}\right)\\
&= \chi^2_{R_{XY}}\left(P_{XY}^{(0)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}\right) + \chi^2_{R_{XY}}\left(P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}, \sum_{i=0}^{k}\alpha_i P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}\right),
\end{aligned} \qquad (30)$$

which comes from the fact that

$$
\begin{aligned}
&\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \frac{\left(P_{XY}^{(0)}(x,y)-[P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}](x,y)\right)[P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}](x,y)}{P_X^{(0)}(x)P_Y^{(0)}(y)} \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \left(P_{XY}^{(0)}(x,y)-[P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}](x,y)\right)\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y) \\
&= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_{XY}^{(0)}(x,y)\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y) \\
&\quad - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_X^{(0)}(x)P_Y^{(0)}(y)\boldsymbol{f}^{\mathrm{T}}(x)\frac{1}{P_Y^{(0)}(y)}\boldsymbol{\Lambda}_{\boldsymbol{f}}^{-1}\left(\sum_{x\in\mathcal{X}} P_{XY}^{(0)}(x,y)\boldsymbol{f}(x)\right)\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y) \\
&= \mathbb{E}_{P_{XY}^{(0)}}[\boldsymbol{f}^{\mathrm{T}}(X)\boldsymbol{g}(Y)] - \mathbb{E}_{P_{XY}^{(0)}}[\boldsymbol{f}^{\mathrm{T}}(X)\boldsymbol{g}(Y)] \\
&= 0.
\end{aligned} \tag{31}
$$

Moreover, we have

$$
\begin{aligned}
&\tilde{V}^{(i)} \\
&= n_i\mathbb{E}\left[\chi^2_{R_{XY}}\left(P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\hat{\boldsymbol{g}}_i)}\right)\right] \\
&= n_i\mathbb{E}\left[\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P_X^{(0)}(x)P_Y^{(0)}(y)\left(\boldsymbol{f}^{\mathrm{T}}(x)\frac{1}{P_Y^{0}(y)}\boldsymbol{\Lambda}_{\boldsymbol{f}}^{-1}\left(\sum_{x\in\mathcal{X}}\left(P_{XY}^{(i)}(x,y)-\hat{P}_{XY}^{(i)}(x,y)\right)\boldsymbol{f}(x)\right)\right)^2\right] \\
&= n_i\sum_{y=1}^{|\mathcal{Y}|}\frac{1}{P_Y^{(0)}(y)}\operatorname{tr}\left(\boldsymbol{\Lambda}_{\boldsymbol{f}}^{-1}\right. \\
&\quad \left.\cdot\,\mathbb{E}\left[\left(\sum_{x\in\mathcal{X}}\left(P_{XY}^{(i)}(x,y)-\hat{P}_{XY}^{(i)}(x,y)\right)\boldsymbol{f}(x)\right)\left(\sum_{x\in\mathcal{X}}\left(P_{XY}^{(i)}(x,y)-\hat{P}_{XY}^{(i)}(x,y)\right)\boldsymbol{f}(x)\right)^{\mathrm{T}}\right]\right) \\
&= \sum_{y=1}^{|\mathcal{Y}|}\frac{P_Y^{(i)}(y)}{P_Y^{(0)}(y)}\operatorname{tr}\left(\boldsymbol{\Lambda}_{\boldsymbol{f}}^{-1}\mathbb{E}_{P_{X|Y=y}^{(i)}}[\boldsymbol{f}(X)\boldsymbol{f}^{\mathrm{T}}(X)]\right) - \sum_{y=1}^{|\mathcal{Y}|}\frac{[P_Y^{(i)}(y)]^2}{P_Y^{(0)}(y)}\left\|\boldsymbol{\Lambda}_{\boldsymbol{f}}^{-\frac{1}{2}}\mathbb{E}_{P_{X|Y=y}^{(i)}}[\boldsymbol{f}(X)]\right\|^2.
\end{aligned} \tag{32}
$$

Using (29), (30) and (32), we obtain (16) as desired.

D Proof of Proposition 5

Since $\boldsymbol{g}^*$ as defined in (14) satisfies $\boldsymbol{g}^* = \sum_{i=0}^k \alpha_i\hat{\boldsymbol{g}}_i$, the weight $\boldsymbol{g}^*$ in the convex combination model is the solution of the following optimization problem

$$
\boldsymbol{g}^* = \arg\min_{\boldsymbol{g}}\ \chi^2_{R_{XY}}\left(\sum_{i=0}^k \alpha_i\hat{P}_{XY}^{(i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}\right). \tag{33}
$$

Then, we prove that the following two optimization problems have the same solution $\boldsymbol{g}^*$

$$
\begin{aligned}
\boldsymbol{g}^* &= \arg\min_{\boldsymbol{g}}\ \chi^2_{R_{XY}}\left(\sum_{i=0}^k \alpha_i\hat{P}_{XY}^{(i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}\right) \\
&= \arg\min_{\boldsymbol{g}}\ \sum_{i=0}^k \alpha_i\chi^2_{R_{XY}}\left(\hat{P}_{XY}^{(i)}, P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g})}\right).
\end{aligned} \tag{34}
$$

To obtain (34), we use the fact that the difference

$$\chi^2_{R_{XY}}\left(\sum_{i=0}^{k}\alpha_i\hat{P}^{(i)}_{XY}, P^{(0)}_X\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X}\right) - \sum_{i=0}^{k}\alpha_i\chi^2_{R_{XY}}\left(\hat{P}^{(i)}_{XY}, P^{(0)}_X\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X}\right)$$
$$= \sum_{x\in\mathcal{X},y\in\mathcal{Y}}\frac{\left(\sum_{i=0}^{k}\alpha_i\hat{P}^{(i)}_{XY}(x,y)\right)^2}{P^{(0)}_X(x)P^{(0)}_Y(y)} - \sum_{x\in\mathcal{X},y\in\mathcal{Y}}\frac{\sum_{i=0}^{k}\alpha_i\left[\hat{P}^{(i)}_{XY}(x,y)\right]^2}{P^{(0)}_X(x)P^{(0)}_Y(y)} \tag{35}$$

is irrelevant to $\boldsymbol{g}$.

Using (34), we obtain Proposition 5 as desired.

E Details for experiments

E.1 Training Loss

To compute the training loss $L^{(\boldsymbol{\alpha},\boldsymbol{f},\boldsymbol{g})}$ as defined in (17), we first introduce the following lemma.

Lemma 7 ([48], Proposition 2)**.** *Let $(\boldsymbol{f}^*,\boldsymbol{g}^*)$ be the features that minimize the χ^2-distance loss $\chi^2_{R_{XY}}(\hat{P}^{(0)}_{XY},\hat{P}^{(0)}_X\hat{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X})$, where $\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X}(x,y)\triangleq\hat{P}_Y(y)(1+\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y))$, for all x,y, and the reference distribution being $\hat{P}^{(0)}_X\hat{P}^{(0)}_Y$. Then, we have*

$$\mathbb{E}_{\hat{P}^{(0)}_X}[\boldsymbol{f}^*(X)] = \mathbb{E}_{\hat{P}^{(0)}_Y}[\boldsymbol{g}^*(Y)] = \mathbf{0}, \tag{36}$$

and $(\boldsymbol{f}^*,\boldsymbol{g}^*)$ are also the optimal features that maximize the H-score of target samples:

$$H^{(0)}(\boldsymbol{f},\boldsymbol{g})\triangleq\mathbb{E}_{\hat{P}^{(0)}_{XY}}[\tilde{\boldsymbol{f}}^{\mathrm{T}}(X)\tilde{\boldsymbol{g}}(Y)] - \frac{1}{2}\operatorname{tr}(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}), \tag{37}$$

where $\tilde{\boldsymbol{f}}(X)\triangleq\boldsymbol{f}(X)-\mathbb{E}_{\hat{P}^{(0)}_X}[\boldsymbol{f}(X)]$, $\tilde{\boldsymbol{g}}(Y)\triangleq\boldsymbol{g}(Y)-\mathbb{E}_{\hat{P}^{(0)}_Y}[\boldsymbol{g}(Y)]$, $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}}$ and $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}$ are the covariance matrices of features on target samples, defined as:

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}}\triangleq\mathbb{E}_{\hat{P}^{(0)}_X}[\tilde{\boldsymbol{f}}(X)\tilde{\boldsymbol{f}}^{\mathrm{T}}(X)], \tag{38}$$

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}\triangleq\mathbb{E}_{\hat{P}^{(0)}_Y}[\tilde{\boldsymbol{g}}(Y)\tilde{\boldsymbol{g}}^{\mathrm{T}}(Y)]. \tag{39}$$

Similarly we can define the H-score for sources. For the task $i=1,\cdots,k$,

$$H^{(i)}(\boldsymbol{f},\boldsymbol{g})\triangleq\mathbb{E}_{\hat{P}^{(i)}_{XY}}[\tilde{\boldsymbol{f}}^{\mathrm{T}}(X)\tilde{\boldsymbol{g}}(Y)] - \frac{1}{2}\operatorname{tr}(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}). \tag{40}$$

Then, line 4 in Algorithm 1 can be implemented by

$$(\boldsymbol{f}^*,\boldsymbol{g}^*)\leftarrow\arg\max_{\boldsymbol{f},\boldsymbol{g}}\sum_{i=0}^{k}\alpha_i H^{(i)}(\boldsymbol{f},\boldsymbol{g}). \tag{41}$$

E.2 Computation of Testing Loss

In computing the testing loss, after obtaining $\boldsymbol{f}^*$ from (41), we use the normalization $\tilde{\boldsymbol{f}}^*(X)\triangleq\boldsymbol{f}^*(X)-\mathbb{E}_{\hat{P}^{(0)}_X}[\boldsymbol{f}^*(X)]$ to subtract the sample mean and obtain zero-mean features. Then, the covariance matrix $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}^*}$ of $\boldsymbol{f}^*$ is computed as

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}^*}\triangleq\mathbb{E}_{\hat{P}^{(0)}_X}[\tilde{\boldsymbol{f}}^*(X)\tilde{\boldsymbol{f}}^{*\mathrm{T}}(X)]. \tag{42}$$

Moreover, for $i=0,\cdots,k$, the $\tilde{V}^{(i)}$ [cf. (32)] in (16) can be estimated as

$$\tilde{V}^{(i)}$$
$$= \sum_{y=1}^{|\mathcal{Y}|}\frac{\hat{P}^{(i)}_Y(y)}{\hat{P}^{(0)}_Y(y)}\operatorname{tr}\left(\hat{\boldsymbol{\Lambda}}^{-1}_{\boldsymbol{f}^*}\mathbb{E}_{\hat{P}^{(i)}_{X|Y=y}}[\tilde{\boldsymbol{f}}^*(X)\tilde{\boldsymbol{f}}^{*\mathrm{T}}(X)]\right) - \sum_{y=1}^{|\mathcal{Y}|}\frac{[\hat{P}^{(i)}_Y(y)]^2}{\hat{P}^{(0)}_Y(y)}\left\|\hat{\boldsymbol{\Lambda}}^{-\frac{1}{2}}_{\boldsymbol{f}^*}\mathbb{E}_{\hat{P}^{(i)}_{X|Y=y}}[\tilde{\boldsymbol{f}}^*(X)]\right\|^2. \tag{43}$$

The bias term $\chi^2_{R_{XY}}\left(P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}, \sum_{i=0}^k \alpha_i P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}\right)$ in (16) can be estimated as

$$\chi^2_{R_{XY}}\left(P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_0)}, \sum_{i=0}^k \alpha_i P_X^{(0)}\tilde{P}_{Y|X}^{(\boldsymbol{f},\boldsymbol{g}_i)}\right) = \sum_{i=1}^k\sum_{j=1}^k \alpha_i\alpha_j D_{ij}, \tag{44}$$

where

$$D_{ij} \triangleq \operatorname{tr}\left(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}^*}^{-1}\left(\sum_{y\in\mathcal{Y}}\frac{(h(0,y)-h(i,y))(h(0,y)-h(j,y))^{\mathrm{T}}}{\hat{P}_Y^{(0)}(y)}\right)\right) \tag{45}$$

and

$$h(l,y) \triangleq \hat{P}_Y^{(l)}(y)\mathbb{E}_{\hat{P}_{X|Y=y}^{(l)}}[\tilde{\boldsymbol{f}}^*(X)],\ l=0,\cdots,k. \tag{46}$$

Finally, the testing loss (16) can be expressed as

$$L_{\text{test}}^{(\boldsymbol{\alpha})} = \sum_{i=1}^k\sum_{j=1}^k \alpha_i\alpha_j D_{ij} + \sum_{i=0}^k \frac{\alpha_i^2}{n_i}\tilde{V}^{(i)}. \tag{47}$$

Then, $\boldsymbol{\alpha}^*$ can be computed by solving a non-negative quadratic programming problem.

E.3 Details of Implementations

E.3.1 Multi-Source Transfer Learning

In this experiment, the corresponding labels for the 5 binary classification tasks are as follows: task 0 (airplane and automobile), task 1 (bird and cat), task 2 (deer and dog), task 3 (frog and horse), and task 4 (ship and truck). After training the loss (40) for each source task, if the test accuracy on target samples is less than 50%, we would flip the binary label for this source. Accordingly, in task 3, frog matches automobile and horse matches airplane. In other tasks, the labels of the source match the target in alphabetical order.

Moreover, we normalize all the images for 3 channels under the mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. When the sample size of the target training set is 6, to make the training process stable, we use data augmentation to generate 50 samples by random horizontal flips and random crops.

The feature $\boldsymbol{f}$ is generated of 10 dimensions by GoogLeNet, followed by a fully connection layer $(1024 \to 32)$ with ReLU activation, and then a fully connection layer $(32 \to 10)$. Throughout the training process, we use the Adam optimizer with a learning rate of 0.001 and the batch size for each source task is 50. We train the networks in 20 epochs and before each epoch we reshuffle the training samples.

E.3.2 Few-shot Transfer Learning Tasks on Office-31

In this experiment, the feature $\boldsymbol{f}$ is generated by the pretained and fixed VGG-16 network, followed by a fully connection layer $(4096 \to 1024)$ with ReLU activation, and a fully connection layer $(1024 \to 64)$. Throughout the training process, we use the Adam optimizer with learning rate of 0.0002 and in 100 epochs.

E.3.3 Few-shot Transfer Learning Tasks on Office-Caltech

In this experiment, the feature $\boldsymbol{f}$ is generated by a fully connection layer $(4096 \to 1024)$ with ReLU activation followed by a fully connection layer $(1024 \to 10)$. The inputs of $\boldsymbol{f}$ are the pretrained DeCAF features. Throughout the training process, we use the Adam optimizer with learning rate of 0.01 and in 100 epochs.

E.4 Instruction for codes

We provide code examples in "code.zip". In the folder "./code/data", we provide the features we used. Folder "o31_feature" contains the features of Office-31 dataset, and folder "oc_feature"

contains the features of Office-Caltech dataset. "cifar10_alpha.py" is an example for CIFAR-10 dataset for the case of 6 target samples. "o31_atod_renew.py" is an example for Office-31 dataset for the case of task A→D, with the details of computing renewed $\boldsymbol{\alpha}$. "oc_atoc.py" is an example for Office-Caltech dataset for the case of task A→C.