Appendix

A The Loss Function Leads to the Linear Combined Estimator

In Section 3, we have introduced a linear combined estimator

$$\tilde{P}_{X_1X_2Y} \triangleq (1-\alpha)\cdot \hat{P}_{X_1X_2Y} + \alpha \cdot \hat{P}^{(\mathrm{M})}_{X_1X_2Y}. \tag{1}$$

We say that the estimator (1) is obtained from a widely-used loss function. First note that we can define the loss of different dependency structures as

$$L_0 \triangleq \sum_{x_1,x_2,y} \hat{P}_{X_1X_2Y}(x_1,x_2,y)\log\frac{1}{Q_{X_1X_2Y}(x_1,x_2,y)}, \tag{2}$$

$$L_1 \triangleq \sum_{x_1,x_2,y} \hat{P}^{(M)}_{X_1X_2Y}(x_1,x_2,y)\log\frac{1}{Q_{X_1X_2Y}(x_1,x_2,y)}, \tag{3}$$

where $Q_{X_1X_2Y}$ represents the distribution estimation, $\hat{P}_{X_1X_2Y}$ and $\hat{P}^{(M)}_{X_1X_2Y}$ are defined in Section 3. As a result, minimizing the linear combination of L_0 and L_1 can give us the optimal estimator:

$$\begin{aligned}
&\mathop{\arg\min}_{Q_{X_1X_2Y}\in\mathcal{P}} (1-\alpha)L_0 + \alpha L_1 \\
&= \mathop{\arg\min}_{Q_{X_1X_2Y}\in\mathcal{P}} \sum_{x_1,x_2,y}\Big((1-\alpha)\hat{P}_{X_1X_2Y}(x_1,x_2,y) + \alpha\hat{P}^{(M)}_{X_1X_2Y}(x_1,x_2,y)\Big)\log\frac{1}{Q_{X_1X_2Y}(x_1,x_2,y)} \\
&= \mathop{\arg\min}_{Q_{X_1X_2Y}\in\mathcal{P}} D_{\mathrm{KL}}\Big((1-\alpha)\hat{P}_{X_1X_2Y} + \alpha\hat{P}^{(M)}_{X_1X_2Y}\|Q_{X_1X_2Y}\Big) \\
&= (1-\alpha)\hat{P}_{X_1X_2Y} + \alpha\hat{P}^{(M)}_{X_1X_2Y},
\end{aligned} \tag{4}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the K-L divergence and the second equation is obtained from the fact that the empirical distribution $(1-\alpha)\hat{P}_{X_1X_2Y} + \alpha\hat{P}^{(M)}_{X_1X_2Y}$ is determined by training samples and does not change with $Q_{X_1X_2Y}$.

B Detailed Statement and Proof of Theorem 3

For the following sections, we abuse our notations by removing all the subscripts for simplicity. We first give the detailed version of Theorem 3.

The testing loss of the two modality case can be expressed as

$$\begin{aligned}
\tilde{\mathcal{L}}_{\mathrm{test}}(\alpha) = \left(\frac{1}{n}C + \frac{1}{n}V + \chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y})\right)\cdot\alpha^2 \\
- \frac{2}{n}C\cdot\alpha + \frac{1}{n}(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|-1),
\end{aligned} \tag{5}$$

and the optimal coefficient α^* to minimize the testing loss (5) can be given as

$$\alpha^* = \frac{\frac{1}{n}C}{\chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y}) + \frac{1}{n}C + \frac{1}{n}V}, \tag{6}$$

where

$$C \triangleq [|\mathcal{Y}|\cdot(|\mathcal{X}_1||\mathcal{X}_2| - |\mathcal{X}_1| - |\mathcal{X}_2|)+1] + \frac{1}{n}\left[\sum_y \frac{|\mathcal{X}_1||\mathcal{X}_2|}{P_Y(y)} - |\mathcal{Y}|\cdot(1+|\mathcal{X}_1|+|\mathcal{X}_2|)+2\right] \tag{7}$$

$$\begin{aligned}
V \triangleq & -\frac{6n^2-11n+6}{n^2}\chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y}) \\
& + \frac{2(n-1)(n-2)}{n^2}\left[\sum_{x_2,y}\chi^2(P_{X_1|X_2Y}, P_{X_1|Y}) + \sum_{x_1,y}\chi^2(P_{X_2|X_1Y}, P_{X_2|Y})\right] \\
& \quad + \frac{(n-1)}{n^2}\sum_y \frac{1}{P(y)}\sum_{x_1,x_2}\frac{P(x_1|y)P(x_2|y)-P(x_1,x_2|y)}{P(x_1,x_2|y)} \\
& \quad + \sum_y \frac{1}{P(y)}\left[\frac{2(n-1)}{n^2}(|\mathcal{X}_1|+|\mathcal{X}_2|+1) - \frac{(2n+1)}{n^2}|\mathcal{X}_1||\mathcal{X}_2|\right] + \frac{|\mathcal{X}_1||\mathcal{X}_2|}{n^2}\sum_y\frac{1}{(P(y))^2} \\
& \quad + \frac{n^2-3n+4}{n^2}|\mathcal{Y}|\,(|\mathcal{X}_1|+|\mathcal{X}_2|+1) - \frac{1}{n}|\mathcal{Y}| - \frac{(n-2)(n-3)}{n^2}.
\end{aligned} \tag{8}$$

B.1 Proof of Theorem 3

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

Lemma 1. *Suppose that random variables X_1, X_2, and Y follow a joint distribution $P_{X_1X_2Y}$. With n samples, we have the empirical distribution $\hat{P}_{X_1X_2Y}(x_1,x_2,y)$, and the Markov estimation $\hat{P}^{(\mathrm{M})}_{X_1X_2Y}(x_1,x_2,y) \triangleq \hat{P}_{X_1|Y}(x_1|y)\hat{P}_{X_2|Y}(x_2|y)\hat{P}_Y(y)$. Along with the assumption that the label distribution has been learned well ,i.e. $\hat{P}(y) = P(y)$, we have*

$$\mathbb{E}\left[\hat{P}^{(M)}(x_1,x_2,y)\right] = P^{(M)}(x_1,x_2,y) + \frac{1}{n}\left(P(x_1,x_2|y) - P^{(M)}(x_1,x_2,y)\right), \tag{9}$$

$$\mathbb{E}\left[\hat{P}^2(x_1,x_2,y)\right] = \frac{n-1}{n}P^2(x_1,x_2,y) + \frac{1}{n}P(x_1,x_2,y), \tag{10}$$

$$\begin{aligned}
&\mathbb{E}\left[\hat{P}(x_1,x_2,y)\hat{P}^{(M)}(x_1,x_2,y)\right] \\
&= \frac{(n-1)(n-2)}{n^2}P(x_1,x_2,y)P^{(M)}(x_1,x_2,y) \\
&\quad + \frac{n-1}{n^2}P(x_1,x_2|y)\Big(P(x_1,x_2,y)+P(x_1,y)+P(x_2,y)\Big) + \frac{1}{n^2}P(x_1,x_2|y),
\end{aligned} \tag{11}$$

$$\begin{aligned}
&\mathbb{E}\left[(\hat{P}^{(M)}(x_1,x_2,y))^2\right] \\
&= \frac{(n-1)(n-2)(n-3)}{n^3}(P^{(M)}(x_1,x_2,y))^2 \\
&\quad + \frac{2(n-1)(n-2)}{n^3}P(x_1|y)P(x_2|y)\Big(P(x_1,x_2,y)+P(x_1,y)+P(x_2,y)\Big) \\
&\quad + \frac{(n-1)}{n^3}\Big[P(x_1|y)P(x_2|y) + 2P(x_1,x_2|y)\Big(P(x_1,x_2|y)+P(x_1|y)+P(x_2|y)\Big)\Big] \\
&\quad + \frac{1}{n^3}\frac{1}{P(y)}P(x_1,x_2|y).
\end{aligned} \tag{12}$$

Proof. For (9), we have

$$\begin{aligned}
\mathbb{E}\left[\hat{P}^{(M)}(x_1,x_2,y)\right] &= \frac{1}{P(y)}\mathbb{E}\left[\hat{P}(x_1,y)\hat{P}(x_2,y)\right] \\
&= \frac{1}{P(y)}\Big[\frac{n-1}{n}P(x_1,y)P(x_2,y) + \frac{1}{n}P(x_1,x_2,y)\Big] \\
&= P^{(M)}(x_1,x_2,y) + \frac{1}{n}\left(P(x_1,x_2|y) - P^{(M)}(x_1,x_2,y)\right),
\end{aligned} \tag{13}$$

where to obtain (13), we have used a fact that for random variables Z_1 and Z_2, with joint distribution $P_{Z_1 Z_2}$, along with their empirical distribution $\hat{P}_{Z_1 Z_2}$ after n samples pairs $(z_1^{(i)}, z_2^{(i)}), i = 1, \ldots, n$, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{P}(Z_1)\hat{P}(Z_2)\right] &= \mathbb{E}\left[\frac{1}{n^2}\mathbb{1}\{z_1^{(i)} = Z_1\} \cdot \mathbb{1}\{z_2^{(i)} = Z_2\}\right] \\
&= \frac{1}{n^2}\Big(n(n-1)P(z_1)P(z_2) + nP(z_1, z_2)\Big) \\
&= \frac{n-1}{n}P(z_1)P(z_2) + \frac{1}{n}P(z_1, z_2)
\end{aligned}$$

Expressions (10)- (12) can be obtained similarly. □

Then, the testing loss (5) can be expressed as

$$\begin{aligned}
\tilde{\mathcal{L}}_{\text{test}}(\alpha) &= \mathbb{E}\left[\chi^2(P_{X_1X_2Y}, (1-\alpha)\cdot \hat{P}_{X_1X_2Y} + \alpha \cdot \hat{P}^{(\mathrm{M})}_{X_1X_2Y})\right] \\
&= \sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\mathbb{E}\left[\Big((1-\alpha)\cdot \hat{P}(x_1,x_2,y) + \alpha\cdot \hat{P}^{(\mathrm{M})}(x_1,x_2,y) - P(x_1,x_2,y)\Big)^2\right] \\
&= \sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\mathbb{E}\Bigg[\bigg[\Big(\hat{P}(x_1,x_2,y) - P(x_1,x_2,y)\Big) \\
&\qquad\qquad + \alpha\Big(\hat{P}^{(\mathrm{M})}(x_1,x_2,y) - \hat{P}(x_1,x_2,y)\Big)\bigg]^2\Bigg] \\
&= \alpha^2 \sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\mathbb{E}\left[\Big(\hat{P}^{(\mathrm{M})}(x_1,x_2,y) - \hat{P}(x_1,x_2,y)\Big)^2\right] \qquad (14) \\
&\quad + 2\alpha \sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\mathbb{E}\bigg[\Big(\hat{P}^{(\mathrm{M})}(x_1,x_2,y) - \hat{P}(x_1,x_2,y)\Big) \\
&\qquad\qquad \cdot \Big(\hat{P}(x_1,x_2,y) - P(x_1,x_2,y)\Big)\bigg] \qquad (15) \\
&\quad + \sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\mathbb{E}\left[\Big(\hat{P}(x_1,x_2,y) - P(x_1,x_2,y)\Big)^2\right]. \qquad (16)
\end{aligned}$$

To obtain the testing loss $\tilde{\mathcal{L}}_{\text{test}}$, we need to derive the expressions after the coefficient α in terms (14)-(16).

As for (16), we have that

$$\begin{aligned}
&\sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\mathbb{E}\left[\Big(\hat{P}(x_1,x_2,y) - P(x_1,x_2,y)\Big)^2\right] \\
&= \sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)}\left[\frac{n-1}{n}P^2(x_1,x_2,y) + \frac{1}{n}P(x_1,x_2,y) - 2P^2(x_1,x_2,y) + P^2(x_1,x_2,y)\right] \qquad (17) \\
&= \frac{1}{n}\sum_{x_1,x_2,y}(1 - P(x_1,x_2,y)) \\
&= \frac{1}{n}(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}| - 1), \qquad (18)
\end{aligned}$$

where to obtain (17), we have used (10).

As for (15), we have that

$$\begin{aligned}
&\sum_{x_1,x_2,y} \frac{1}{P(x_1,x_2,y)} \mathbb{E}\left[\left(\hat{P}^{(\mathrm{M})}(x_1,x_2,y)-\hat{P}(x_1,x_2,y)\right)\left(\hat{P}(x_1,x_2,y)-P(x_1,x_2,y)\right)\right] \\
&= \sum_{x_1,x_2,y}\left[\frac{1}{P(x_1,x_2,y)}\mathbb{E}\left[\hat{P}^{(\mathrm{M})}(x_1,x_2,y)\hat{P}(x_1,x_2,y)\right]-\mathbb{E}\left[\hat{P}^{(\mathrm{M})}(x_1,x_2,y)\right]\right. \\
&\qquad\qquad \left.-\frac{1}{n}(1-P(x_1,x_2,y))\right] \\
&= \sum_{x_1,x_2,y}\frac{1}{P(x_1,x_2,y)}\left[\frac{(n-1)(n-2)}{n^2}P(x_1,x_2,y)P^{(M)}(x_1,x_2,y)+\frac{1}{n^2}P(x_1,x_2|y)\right. \\
&\qquad \left.+\frac{n-1}{n^2}P(x_1,x_2|y)\left(P(x_1,x_2,y)+P(x_1,y)+P(x_2,y)\right)\right] \\
&\qquad\quad -\sum_{x_1,x_2,y}\left[P^{(M)}(x_1,x_2,y)+\frac{1}{n}\left(P_{X_1X_2|Y}(x_1,x_2|y)-P^{(M)}(x_1,x_2,y)\right)\right] \\
&\qquad\qquad -\frac{1}{n}\sum_{x_1,x_2,y}(1-P(x_1,x_2,y)) \qquad\qquad\qquad\qquad\qquad\qquad\qquad (19) \\
&= \frac{(n-1)(n-2)}{n^2}+\frac{n-1}{n^2}|\mathcal{Y}|\left(1+|\mathcal{X}_1|+|\mathcal{X}_2|\right)+\frac{|\mathcal{X}_1||\mathcal{X}_2|}{n^2}\sum_y\frac{1}{P_Y(y)}-1 \\
&\qquad\qquad -\frac{1}{n}(|\mathcal{Y}|-1)-\frac{1}{n}(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|-1) \\
&= -\frac{1}{n}\left[|\mathcal{Y}|\cdot(|\mathcal{X}_1||\mathcal{X}_2|-|\mathcal{X}_1|-|\mathcal{X}_2|)+1\right]+\frac{1}{n^2}\left[\sum_y\frac{|\mathcal{X}_1||\mathcal{X}_2|}{P_Y(y)}-|\mathcal{Y}|\cdot(1+|\mathcal{X}_1|+|\mathcal{X}_2|)+2\right],
\end{aligned} \tag{20}$$

where to obtain (19), we have used (9) and (11).

As for (14), we have that

$$\begin{aligned}
&\sum_{x_1,x_2,y}\frac{1}{P(x_1,x_2,y)}\mathbb{E}\left[\left(\hat{P}^{(\mathrm{M})}(x_1,x_2,y)-\hat{P}(x_1,x_2,y)\right)^2\right] \\
&= \sum_{x_1,x_2,y}\frac{1}{P(x_1,x_2,y)}\left[\mathbb{E}\left[\left(\hat{P}^{(\mathrm{M})}(x_1,x_2,y)\right)^2\right]-2\mathbb{E}\left[\hat{P}^{(\mathrm{M})}(x_1,x_2,y)\hat{P}(x_1,x_2,y)\right]\right. \\
&\qquad\qquad \left.+\mathbb{E}\left[\left(\hat{P}(x_1,x_2,y)\right)^2\right]\right] \\
&= \sum_{x_1,x_2,y}\frac{1}{P(x_1,x_2,y)}\left[\frac{(n-1)(n-2)(n-3)}{n^3}\left(P^{(M)}(x_1,x_2,y)\right)^2\right. \\
&\qquad\qquad +\frac{2(n-1)(n-2)}{n^3}P(x_1|y)P(x_2|y)\Big(P(x_1,x_2,y)+P(x_1,y)+P(x_2,y)\Big) \\
&\qquad\qquad\quad +\frac{(n-1)}{n^3}\Big[2P(x_1,x_2|y)\Big(P(x_1,x_2|y)+P(x_1|y)+P(x_2|y)\Big)+P(x_1|y)P(x_2|y)\Big] \\
&\qquad\qquad\qquad \left.+\frac{1}{n^3}\frac{1}{P(y)}P(x_1,x_2|y)\right]
\end{aligned}$$

$$
\begin{aligned}
&-2\sum_{x_1,x_2,y}\frac{1}{P(x_1,x_2,y)}\Bigg[\frac{(n-1)(n-2)}{n^2}P(x_1,x_2,y)P^{(M)}(x_1,x_2,y)+\frac{1}{n^2}P(x_1,x_2|y)\\
&\qquad+\frac{n-1}{n^2}P(x_1,x_2|y)\Big(P(x_1,x_2,y)+P(x_1,y)+P(x_2,y)\Big)\Bigg]\\
&+\sum_{x_1,x_2,y}\frac{1}{P(x_1,x_2,y)}\left(\frac{n-1}{n}P^2(x_1,x_2,y)+\frac{1}{n}P(x_1,x_2,y)\right) \qquad (21)\\
=&\ \frac{(n-1)(n-2)(n-3)}{n^3}\left(\chi^2(P_{X_1X_2Y},P^{(M)}_{X_1X_2Y})+1\right)\\
&+\frac{2(n-1)(n-2)}{n^3}\Bigg[\sum_{x_2,y}\chi^2(P_{X_1|X_2Y},P_{X_1|Y})+\sum_{x_1,y}\chi^2(P_{X_2|X_1Y},P_{X_2|Y})\\
&\qquad+|\mathcal{Y}|\left(|\mathcal{X}_1|+|\mathcal{X}_2|+1\right)\Bigg]\\
&\qquad+\frac{(n-1)}{n^3}\sum_{y}\frac{1}{P(y)}\left(\sum_{x_1,x_2}\frac{P(x_1|y)P(x_2|y)}{P(x_1,x_2|y)}+2\left(|\mathcal{X}_1|+|\mathcal{X}_2|+1\right)\right)\\
&\qquad+\frac{|\mathcal{X}_1||\mathcal{X}_2|}{n^3}\sum_{y}\frac{1}{(P(y))^2}-\frac{2(n-1)(n-2)}{n^2}-\frac{2|\mathcal{X}_1||\mathcal{X}_2|}{n^2}\sum_{y}\frac{1}{P(y)}\\
&\qquad-\frac{2(n-1)}{n^2}|\mathcal{Y}|\left(|\mathcal{X}_1|+|\mathcal{X}_2|+1\right)+1+\frac{1}{n}\left(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|-1\right) \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
=&\ \frac{(n-1)(n-2)(n-3)}{n^3}\chi^2(P_{X_1X_2Y},P^{(M)}_{X_1X_2Y})\\
&+\frac{2(n-1)(n-2)}{n^3}\left[\sum_{x_2,y}\chi^2(P_{X_1|X_2Y},P_{X_1|Y})+\sum_{x_1,y}\chi^2(P_{X_2|X_1Y},P_{X_2|Y})\right]\\
&+\frac{(n-1)}{n^3}\sum_{y}\frac{1}{P(y)}\sum_{x_1,x_2}\frac{P(x_1|y)P(x_2|y)-P(x_1,x_2|y)}{P(x_1,x_2|y)}\\
&+\sum_{y}\frac{1}{P(y)}\left[\frac{2(n-1)}{n^3}\left(|\mathcal{X}_1|+|\mathcal{X}_2|+1\right)-\frac{(n+1)}{n^3}|\mathcal{X}_1||\mathcal{X}_2|\right]+\frac{|\mathcal{X}_1||\mathcal{X}_2|}{n^3}\sum_{y}\frac{1}{(P(y))^2}\\
&-\frac{4(n-1)}{n^3}|\mathcal{Y}|\left(|\mathcal{X}_1|+|\mathcal{X}_2|+1\right)+\frac{1}{n}|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}|-\frac{(n-1)(n-6)}{n^3}, \qquad (23)
\end{aligned}
$$

where to obtain (21), we have used (10)-(12); to obtain (22), we have used the fact that for a random variable Z and its distributions P_Z, Q_Z, we have that $\sum_z \frac{(Q_Z(z)-P_Z(z))^2}{P_Z(z)} = \chi^2(P_Z, Q_Z)+1$.

From (18), (20), and (23), we obtain the expressions of the Theorem 3.

C Detailed Statement and Proof of Theorem 5

C.1 Detailed Statement of Theorem 5

We first give the statement and proof of Theorem 5 when $k=2$. In Theorem 5, we adopt the factorization form to give the calculation of the optimal coefficient α^* using features. Now, we give the detailed form of the optimal α^*. The numerator of it consists of 5 different terms, while the denominator consists of 13 different terms. All these terms can be calculated by different empirical means of training features $\boldsymbol{f}$ and $\boldsymbol{g}$.

$$\alpha^* = \frac{a}{b},$$

where

$$
\begin{aligned}
a \triangleq{} & \sum_{y} \left(\frac{1}{n^2} - \frac{1}{n}P(y)\right) \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2'|y)P(x_1'',x_2''|y) \\
& + \sum_{y} \left(\frac{1}{n} - \frac{1}{n^2P(y)}\right) \sum_{x_1',x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2'|y) \\
& + \frac{2(n-1)}{n^2} \sum_{y} P(y) \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2'|y)P(x_1''|y)P(x_2''|y) \\
& - \frac{(n-1)}{n^2} \sum_{y} \sum_{x_1'} \sum_{x_2',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2'')P(x_1',x_2'|y)P(x_2''|y) \\
& - \frac{(n-1)}{n^2} \sum_{y} \sum_{x_2'} \sum_{x_1',x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2')P(x_1',x_2'|y)P(x_1''|y) \\
b \triangleq{} & \frac{(n-1)(n-2)(n-3)}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1'|y)P(x_2'|y)P(x_1''|y)P(x_2''|y) \\
& + \sum_{y} \left[\frac{2(n-1)(n-2)}{n^3} - \frac{2(n-1)(n-2)}{n^2}P(y)\right] \\
& \qquad \cdot \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2'|y)P(x_1''|y)P(x_2''|y) \\
& + \frac{(n-1)(n-2)}{n^3} \sum_{y} \sum_{x_1'} \sum_{x_2',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2'')P(x_1'|y)P(x_2'|y)P(x_2''|y) \\
& + \frac{2(n-1)(n-2)}{n^3} \sum_{y} \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2''|y)P(x_1''|y)P(x_2'|y) \\
& + \frac{(n-1)(n-2)}{n^3} \sum_{y} \sum_{x_2'} \sum_{x_1',x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2')P(x_1'|y)P(x_2'|y)P(x_1''|y) \\
& + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_2'} \sum_{x_1',x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2')P(x_1'|y)P(x_1'',x_2'|y) \\
& + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1'} \sum_{x_2',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2'')P(x_2'|y)P(x_1',x_2''|y) \\
& + \sum_{y} \left[\frac{n-1}{n^3}\frac{1}{P(y)} - \frac{2(n-1)}{n^2}\right] \sum_{x_2'} \sum_{x_1',x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2')P(x_1''|y)P(x_1',x_2'|y) \\
& + \sum_{y} \left[\frac{n-1}{n^3}\frac{1}{P(y)} - \frac{2(n-1)}{n^2}\right] \sum_{x_1'} \sum_{x_2',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2'')P(x_2''|y)P(x_1',x_2'|y) \\
& + \sum_{y} \left[\frac{n-1}{n^3}\frac{1}{P(y)} - \frac{2(n-1)}{n^2} + \frac{n-1}{n}P(y)\right] \\
& \qquad \cdot \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2'|y)P(x_1'',x_2''|y) \\
& + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1',x_2'} \sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2''|y)P(x_1'',x_2'|y) \\
& + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1',x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1'|y)P(x_2'|y) \\
& + \sum_{y} \left[\frac{1}{n^3}\frac{1}{P^2(y)} - \frac{2}{n^2P(y)} + 1\right] \sum_{x_1',x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2'|y).
\end{aligned}
$$

C.2 Proof of Theorem 5

We first give the expressions of the optimal weight $\boldsymbol{g}^*$. We rewrite the expression of the training loss function as the following.

$$\begin{aligned}
\tilde{\mathcal{L}}_{\text{train}}(\boldsymbol{f},\boldsymbol{g}) &\triangleq (1-\alpha)\chi^2_R\left(\hat{P}_{X_1X_2Y}, \hat{P}_{X_1X_2}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}\right) + \alpha\chi^2_R\left(\hat{P}^{(\mathrm{M})}_{X_1X_2Y}, \hat{P}_{X_1X_2}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}\right)\\
&= (1-\alpha)\sum_{x_1,x_2,y}\frac{(Q(x_1,x_2,y)-\hat{P}(x_1,x_2,y))^2}{R(x_1,x_2,y)}\\
&\quad + \alpha\sum_{x_1,x_2,y}\frac{(Q(x_1,x_2,y)-\hat{P}^{(M)}(x_1,x_2,y))^2}{R(x_1,x_2,y)}\\
&= \sum_{x_1,x_2,y}\frac{P^2(x_1,x_2)P^2(y)}{R(x_1,x_2,y)}\Big(1+2\boldsymbol{f}^T(x_1,x_2)\boldsymbol{g}(y)+\boldsymbol{g}^T(y)\boldsymbol{f}(x_1,x_2)\boldsymbol{f}^T(x_1,x_2)\boldsymbol{g}(y)\Big)\\
&\quad - \frac{2P(x_1,x_2)P(y)}{R(x_1,x_2,y)}\Big[(1-\alpha)\hat{P}(x_1,x_2,y)+\alpha\hat{P}^{(M)}(x_1,x_2,y)\Big](1+\boldsymbol{f}^T(x_1,x_2)\boldsymbol{g}(y))\\
&\quad + \frac{(1-\alpha)\hat{P}^2(x_1,x_2,y)+\alpha(\hat{P}^{(M)}(x_1,x_2,y))^2}{R(x_1,x_2,y)},
\end{aligned}$$

where we define $Q(x_1,x_2,y) \triangleq [\hat{P}_{X_1X_2}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}](x_1,x_2,y)$.
Thus, the differentiation of the training loss function over $\boldsymbol{g}(y')$ is as follows.

$$\begin{aligned}
\frac{\partial\tilde{\mathcal{L}}_{\text{train}}(\boldsymbol{f},\boldsymbol{g})}{\partial\boldsymbol{g}(y')} &= \sum_{x_1,x_2}\frac{P^2(x_1,x_2)P^2(y')}{R(x_1,x_2,y')}\Big(2\boldsymbol{f}(x_1,x_2)+2\boldsymbol{f}(x_1,x_2)\boldsymbol{f}^T(x_1,x_2)\boldsymbol{g}(y')\Big)\\
&\qquad - \frac{2P(x_1,x_2)P(y')}{R(x_1,x_2,y')}\Big[(1-\alpha)\hat{P}(x_1,x_2,y')+\alpha\hat{P}^{(M)}(x_1,x_2,y')\Big]\boldsymbol{f}(x_1,x_2)\\
&= \sum_{x_1,x_2}P(x_1,x_2)P(y')\Big(2\boldsymbol{f}(x_1,x_2)+2\boldsymbol{f}(x_1,x_2)\boldsymbol{f}^T(x_1,x_2)\boldsymbol{g}(y')\Big)\\
&\qquad - 2\Big[(1-\alpha)\hat{P}(x_1,x_2,y')+\alpha\hat{P}^{(M)}(x_1,x_2,y')\Big]\boldsymbol{f}(x_1,x_2) \qquad (24)\\
&= 2P(y')\Lambda_{\boldsymbol{f}}\boldsymbol{g}(y') - 2\sum_{x_1,x_2}\Big[(1-\alpha)\hat{P}(x_1,x_2,y')+\alpha\hat{P}^{(M)}(x_1,x_2,y')\Big]\boldsymbol{f}(x_1,x_2), \qquad (25)
\end{aligned}$$

where we obtain (24) by using the definition of $R(x_1,x_2,y) \triangleq P(x_1,x_2)P(y)$; and we obtain (25) from the assumption that $\sum_{x_1,x_2}P(x_1,x_2)\boldsymbol{f}(x_1,x_2)=0$, and we have also used the notation that $\Lambda_{\boldsymbol{f}} \triangleq \sum_{x_1,x_2}P(x_1,x_2)\boldsymbol{f}(x_1,x_2)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)$.

Set the gradient to zero, we obtain the optimal weights $\boldsymbol{g}^*(y')$.

$$\boldsymbol{g}^*(y') = \frac{1}{P(y')}\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1,x_2}\Big[(1-\alpha)\hat{P}(x_1,x_2,y')+\alpha\hat{P}^{(M)}(x_1,x_2,y')\Big]\boldsymbol{f}(x_1,x_2). \tag{26}$$

Then, we give the proof. At first, we have the following three lemmas.

Lemma 2. *We have*

$$\begin{aligned}\mathbb{E}[\hat{P}(x_1',x_2',y)\hat{P}(x_1'',x_2'',y)] &= \frac{n-1}{n}P(x_1',x_2',y)P(x_1'',x_2'',y)\\ &\quad + \frac{1}{n}P(x_1',x_2',y)\mathbb{1}\{(x_1',x_2')=(x_1'',x_2'')\}. \end{aligned}\tag{27}$$

$$\begin{aligned}\mathbb{E}[\hat{P}(x_1',x_2',y)\hat{P}^{(M)}(x_1'',x_2'',y)] &= \frac{(n-1)(n-2)}{n^2}P(x_1',x_2',y)P^{(M)}(x_1'',x_2'',y)\\ &\quad + \frac{n-1}{n^2}P(x_1',x_2'|y)P(x_2'',y)\mathbb{1}\{x_1'=x_1''\}\\ &\quad + \frac{n-1}{n^2}P(x_1',x_2'|y)P(x_1'',y)\mathbb{1}\{x_2'=x_2''\}\\ &\quad + \frac{n-1}{n^2}P(x_1',x_2'|y)P(x_1'',x_2'',y)\\ &\quad + \frac{1}{n^2}P(x_1',x_2'|y)\mathbb{1}\{(x_1',x_2')=(x_1'',x_2'')\}. \end{aligned}\tag{28}$$

Lemma 3. *We have*

$$\begin{aligned}&\mathbb{E}[\hat{P}^{(M)}(x_1',x_2',y)\hat{P}^{(M)}(x_1'',x_2'',y)]\\ &= \frac{(n-1)(n-2)(n-3)}{n^3}P^{(M)}(x_1',x_2',y)P^{(M)}(x_1'',x_2'',y)\\ &\quad + \frac{(n-1)(n-2)}{n^3P^2(y)}\Big(P(x_1',x_2',y)P(x_1'',y)P(x_2'',y) + P(x_1',y)P(x_2',y)P(x_2'',y)\mathbb{1}\{x_1'=x_1''\}\\ &\qquad + P(x_1',x_2'',y)P(x_1'',y)P(x_2',y) + P(x_1'',x_2',y)P(x_1',y)P(x_2'',y)\\ &\qquad + P(x_1',y)P(x_2',y)P(x_1'',y)\mathbb{1}\{x_2'=x_2''\} + P(x_1',y)P(x_2',y)P(x_1'',x_2'',y)\Big)\\ &\quad + \frac{(n-1)}{n^3P^2(y)}\Big(P(x_1',y)P(x_1'',x_2'',y)\mathbb{1}\{x_2'=x_2''\} + P(x_2',y)P(x_1'',x_2'',y)\mathbb{1}\{x_1'=x_1''\}\\ &\qquad + P(x_1'',y)P(x_1',x_2',y)\mathbb{1}\{x_2'=x_2''\} + P(x_2'',y)P(x_1',x_2',y)\mathbb{1}\{x_1'=x_1''\}\\ &\qquad + P(x_1',x_2',y)P(x_1'',x_2'',y) + P(x_1',x_2'',y)P(x_1'',x_2',y)\\ &\qquad + P(x_1',y)P(x_2',y)\mathbb{1}\{(x_1',x_2')=(x_1'',x_2'')\}\Big)\\ &\quad + \frac{1}{n^2P^2(y)}P(x_1',x_2',y)\mathbb{1}\{(x_1',x_2')=(x_1'',x_2'')\}. \end{aligned}\tag{29}$$

Lemma 4. *We have*

$$\begin{aligned}&\sum_{x_1',x_2'}\sum_{x_1'',x_2''}\sum_{x_1,x_2}P(x_1,x_2)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')\boldsymbol{f}^{\mathrm{T}}(x_1'',x_2'')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1,x_2)\\ &\quad = \sum_{x_1',x_2'}\sum_{x_1'',x_2''}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'') \end{aligned}\tag{30}$$

Proof. Notice that for vectors v_1, v_2, their inner product $v_1^{\mathrm{T}}v_2 = \mathrm{tr}(v_2\cdot v_1^{\mathrm{T}})$. Thus, we have

$$\begin{aligned}&\sum_{x_1',x_2'}\sum_{x_1'',x_2''}\sum_{x_1,x_2}P(x_1,x_2)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')\boldsymbol{f}^{\mathrm{T}}(x_1'',x_2'')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1,x_2)\\ &= \sum_{x_1',x_2'}\sum_{x_1'',x_2''}\sum_{x_1,x_2}P(x_1,x_2)\,\mathrm{tr}\left(\boldsymbol{f}(x_1,x_2)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')\boldsymbol{f}^{\mathrm{T}}(x_1'',x_2'')\Lambda_{\boldsymbol{f}}^{-1}\right)\\ &= \sum_{x_1',x_2'}\sum_{x_1'',x_2''}\mathrm{tr}\left(\sum_{x_1,x_2}P(x_1,x_2)\boldsymbol{f}(x_1,x_2)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')\boldsymbol{f}^{\mathrm{T}}(x_1'',x_2'')\Lambda_{\boldsymbol{f}}^{-1}\right)\\ &= \sum_{x_1',x_2'}\sum_{x_1'',x_2''}\mathrm{tr}\left(\Lambda_{\boldsymbol{f}}\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')\boldsymbol{f}^{\mathrm{T}}(x_1'',x_2'')\Lambda_{\boldsymbol{f}}^{-1}\right)\\ &= \sum_{x_1',x_2'}\sum_{x_1'',x_2''}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'') \end{aligned}$$

$\square$

Then, we give

$$\begin{aligned}\langle \boldsymbol{f}(x_1,x_2), \boldsymbol{g}^*(y)\rangle &= \frac{\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)}{P(y)}\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1,x_2}\Big[(1-\alpha)\hat{P}(x_1,x_2,y)+\alpha\hat{P}^{(M)}(x_1,x_2,y)\Big]\boldsymbol{f}(x_1,x_2)\\ &= \frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\hat{P}(x_1',x_2',y)\boldsymbol{f}(x_1',x_2')\\ &\quad -\alpha\cdot\frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\Big(\hat{P}(x_1',x_2',y)-\hat{P}^{(M)}(x_1',x_2',y)\Big)\boldsymbol{f}(x_1',x_2')\\ &= \hat{a}_1(x_1,x_2,y)-\alpha\hat{a}_2(x_1,x_2,y),\end{aligned}$$

where

$$\begin{aligned}\hat{a}_1(x_1,x_2,y) &\triangleq \frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\hat{P}(x_1',x_2',y)\boldsymbol{f}(x_1',x_2'),\\ \hat{a}_2(x_1,x_2,y) &\triangleq \frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\Big(\hat{P}(x_1',x_2',y)-\hat{P}^{(M)}(x_1',x_2',y)\Big)\boldsymbol{f}(x_1',x_2').\end{aligned}$$

Now, we give the testing loss as

$$\tilde{\mathcal{L}}_{\text{test}}(\alpha)=\mathbb{E}[\chi^2_R(P_{X_1X_2Y},P_{X_1X_2}Q^{\alpha}_{Y|X_1X_2})] \tag{31}$$

$$\begin{aligned}&=\sum_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\langle\boldsymbol{f}(x_1,x_2),\boldsymbol{g}^*(y)\rangle^2]\\ &\quad+2\sum_{x_1,x_2,y}(R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\langle\boldsymbol{f}(x_1,x_2),\boldsymbol{g}^*(y)\rangle]+\chi^2_R(R_{X_1X_2Y},P_{X_1X_2Y})\\ &=\alpha^2\sum_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\hat{a}_2^2(x_1,x_2,y)] \qquad (32)\\ &\quad-2\alpha\sum_{x_1,x_2,y}(R(x_1,x_2,y)\mathbb{E}[\hat{a}_1(x_1,x_2,y)\cdot\hat{a}_2(x_1,x_2,y)]\\ &\qquad+(R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\hat{a}_2(x_1,x_2,y)]) \qquad (33)\\ &\quad+\sum_{x_1,x_2,y}(R(x_1,x_2,y)\mathbb{E}[\hat{a}_1^2(x_1,x_2,y)]+2(R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\hat{a}_1(x_1,x_2,y)])\\ &\quad+\chi^2_R(R_{X_1X_2Y},P_{X_1X_2Y})\end{aligned}$$

Thus, the optimal α^* is

$$\begin{aligned}\alpha^*=&\frac{\sum\limits_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\hat{a}_1(x_1,x_2,y)\cdot\hat{a}_2(x_1,x_2,y)]}{\sum\limits_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\hat{a}_2^2(x_1,x_2,y)]}\\ &+\frac{\sum\limits_{x_1,x_2,y}(R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\hat{a}_2(x_1,x_2,y)]}{\sum\limits_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\hat{a}_2^2(x_1,x_2,y)]}.\end{aligned} \tag{34}$$

Next, we calculate the three summations in (34).

At first, we derive $\sum_{x_1,x_2,y}(R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\hat{a}_2(x_1,x_2,y)]$.

$$\begin{aligned}
&\sum_{x_1,x_2,y}(R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\hat{a}_2(x_1,x_2,y)]\\
&=\sum_{x_1,x_2,y}\mathbb{E}[\frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\Big(\hat{P}(x_1',x_2',y)-\hat{P}^{(M)}(x_1',x_2',y)\Big)\boldsymbol{f}(x_1',x_2')]\\
&\qquad\qquad\cdot(P(x_1,x_2)P(y)-P(x_1,x_2,y))\\
&=\sum_{y}\Big(\sum_{x_1,x_2}(P(x_1,x_2)-P(x_1,x_2|y))\Big)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\\
&\qquad\cdot\sum_{x_1',x_2'}\Big[P(x_1',x_2',y)-\Big(P^{(M)}(x_1',x_2',y)+\frac{1}{n}(P(x_1',x_2'|y)-P^{(M)}(x_1',x_2',y))\Big)\Big]\boldsymbol{f}(x_1',x_2')
\end{aligned}\tag{35}$$

$$\begin{aligned}
&=-\sum_{y}\sum_{x_1,x_2}P(x_1,x_2|y)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\\
&\qquad\cdot\sum_{x_1',x_2'}\Big[P(x_1',x_2',y)-\Big(P^{(M)}(x_1',x_2',y)+\frac{1}{n}(P(x_1',x_2'|y)-P^{(M)}(x_1',x_2',y))\Big)\Big]\boldsymbol{f}(x_1',x_2')
\end{aligned}\tag{36}$$

$$\begin{aligned}
&=-\sum_{y}P(y)\Big[\sum_{x_1,x_2}P(x_1,x_2|y)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Big]\Lambda_{\boldsymbol{f}}^{-1}\Big[\sum_{x_1',x_2'}(P(x_1',x_2'|y)-P(x_1'|y)P(x_2'|y))\,\boldsymbol{f}(x_1',x_2')\Big]\\
&\quad+\frac{1}{n}\sum_{y}\Big[\sum_{x_1,x_2}P(x_1,x_2|y)\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Big]\Lambda_{\boldsymbol{f}}^{-1}\Big[\sum_{x_1',x_2'}\Big(P(x_1',x_2'|y)-P^{(M)}(x_1',x_2',y)\Big)\,\boldsymbol{f}(x_1',x_2')\Big],
\end{aligned}\tag{37}$$

where we obtain (36) from the fact that $\sum_{x_1,x_2}P(x_1,x_2)\boldsymbol{f}(x_1,x_2)=0$.

Then, we derive $\sum_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\hat{a}_1(x_1,x_2,y)\cdot\hat{a}_2(x_1,x_2,y)]$.

$$\begin{aligned}
&\sum_{x_1,x_2,y}R(x_1,x_2,y)\mathbb{E}[\hat{a}_1(x_1,x_2,y)\cdot\hat{a}_2(x_1,x_2,y)]\\
&=\sum_{x_1,x_2,y}P(x_1,x_2)P(y)\mathbb{E}\Big[\frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\hat{P}(x_1',x_2',y)\boldsymbol{f}(x_1',x_2')\\
&\qquad\qquad\cdot\frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1'',x_2''}\Big(\hat{P}(x_1'',x_2'',y)-\hat{P}^{(M)}(x_1'',x_2'',y)\Big)\boldsymbol{f}(x_1'',x_2'')\Big]\\
&=\sum_{y}\frac{1}{P(y)}\sum_{x_1',x_2'}\sum_{x_1'',x_2''}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')\mathbb{E}\Big[\hat{P}(x_1',x_2',y)\hat{P}(x_1'',x_2'',y)\Big]\\
&\quad-\sum_{y}\frac{1}{P(y)}\sum_{x_1',x_2'}\sum_{x_1'',x_2''}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')\mathbb{E}\Big[\hat{P}(x_1',x_2',y)\hat{P}^{(M)}(x_1'',x_2'',y)\Big]
\end{aligned}\tag{38}$$

$$\begin{aligned}
&=\frac{n-1}{n}\sum_{y}\frac{1}{P(y)}\sum_{x_1',x_2'}\sum_{x_1'',x_2''}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2',y)P(x_1'',x_2'',y)\\
&\quad+\frac{1}{n}\sum_{y}\frac{1}{P(y)}\sum_{x_1',x_2'}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2',y)\\
&\quad-\frac{(n-1)(n-2)}{n^2}\sum_{y}\frac{1}{P(y)}\sum_{x_1',x_2'}\sum_{x_1'',x_2''}\boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2',y)P^{(M)}(x_1'',x_2'',y)
\end{aligned}$$

$$
\begin{aligned}
&- \frac{(n-1)}{n^2}\sum_{y} \frac{1}{P(y)} \sum_{x_1',x_2'}\sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2',y)P(x_1'',x_2''|y)\\
&- \frac{(n-1)}{n^2}\sum_{y} \frac{1}{P(y)} \sum_{x_1'}\sum_{x_2',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2'')P(x_1',x_2'|y)P(x_2'',y)\\
&- \frac{(n-1)}{n^2}\sum_{y} \frac{1}{P(y)} \sum_{x_2'}\sum_{x_1',x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2')P(x_1',x_2'|y)P(x_1'',y)\\
&- \frac{1}{n^2}\sum_{y} \frac{1}{P(y)} \sum_{x_1',x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2'|y), \qquad (39)
\end{aligned}
$$

where we obtain (38) from (30), and we obtain (39) from (27) and (28).

Thus, we have the numerator of the optimal coefficient α^*.

$$
\begin{aligned}
&\sum_{x_1,x_2,y} R(x_1,x_2,y)\mathbb{E}[\hat{a}_1(x_1,x_2,y)\cdot \hat{a}_2(x_1,x_2,y)]\\
&\qquad + \sum_{x_1,x_2,y} (R(x_1,x_2,y)-P(x_1,x_2,y))\mathbb{E}[\hat{a}_2(x_1,x_2,y)]\\
&= \sum_{y}\Big(\frac{1}{n^2}-\frac{1}{n}P(y)\Big)\sum_{x_1',x_2'}\sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')P(x_1',x_2'|y)P(x_1'',x_2''|y)\\
&\quad + \sum_{y}\Big(\frac{1}{n}-\frac{1}{n^2P(y)}\Big)\sum_{x_1',x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2'|y)\\
&\quad + \frac{2(n-1)}{n^2}\sum_{y} P(y)\sum_{x_1',x_2'}\sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2')P(x_1',x_2'|y)P(x_1''|y)P(x_2''|y)\\
&\quad - \frac{(n-1)}{n^2}\sum_{y}\sum_{x_1'}\sum_{x_2',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1',x_2'')P(x_1',x_2'|y)P(x_2''|y)\\
&\quad - \frac{(n-1)}{n^2}\sum_{y}\sum_{x_2'}\sum_{x_1',x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2')P(x_1',x_2'|y)P(x_1''|y) \qquad (40)
\end{aligned}
$$

Next, we derive the denominator

$$
\begin{aligned}
&\sum_{x_1,x_2,y} R(x_1,x_2,y)\mathbb{E}[\hat{a}_2^2(x_1,x_2,y)]\\
&= \sum_{x_1,x_2,y} P(x_1,x_2)P(y)\mathbb{E}\Big[\frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1',x_2'}\Big(\hat{P}(x_1',x_2',y)-\hat{P}^{(M)}(x_1',x_2',y)\Big)\boldsymbol{f}(x_1',x_2')\\
&\qquad\qquad \cdot \frac{1}{P(y)}\boldsymbol{f}^{\mathrm{T}}(x_1,x_2)\Lambda_{\boldsymbol{f}}^{-1}\sum_{x_1'',x_2''}\Big(\hat{P}(x_1'',x_2'',y)-\hat{P}^{(M)}(x_1'',x_2'',y)\Big)\boldsymbol{f}(x_1'',x_2'')\Big]\\
&= \sum_{y}\frac{1}{P(y)}\sum_{x_1',x_2'}\sum_{x_1'',x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1',x_2')\Lambda_{\boldsymbol{f}}^{-1}\boldsymbol{f}(x_1'',x_2'')\\
&\qquad\qquad \cdot \mathbb{E}\Big[\Big(\hat{P}(x_1',x_2',y)-\hat{P}^{(M)}(x_1',x_2',y)\Big)\Big(\hat{P}(x_1'',x_2'',y)-\hat{P}^{(M)}(x_1'',x_2'',y)\Big)\Big] \qquad (41)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{y} \frac{1}{P(y)} \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') \mathbb{E}\Big[\hat{P}(x_1', x_2', y) \hat{P}(x_1'', x_2'', y)\Big] \\
&\quad - 2 \sum_{y} \frac{1}{P(y)} \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') \mathbb{E}\Big[\hat{P}(x_1', x_2', y) \hat{P}^{(M)}(x_1'', x_2'', y)\Big] \\
&\qquad + \sum_{y} \frac{1}{P(y)} \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') \mathbb{E}\Big[\hat{P}^{(M)}(x_1', x_2', y) \hat{P}^{(M)}(x_1'', x_2'', y)\Big] \\
&= \frac{(n-1)(n-2)(n-3)}{n^3} \sum_{y} \frac{1}{P(y)} \\
&\qquad \cdot \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1'|y) P(x_2'|y) P(x_1''|y) P(x_2''|y) \\
&\quad + \sum_{y} \Big[\frac{2(n-1)(n-2)}{n^3} - \frac{2(n-1)(n-2)}{n^2} P(y)\Big] \\
&\qquad \cdot \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1', x_2'|y) P(x_1''|y) P(x_2''|y) \\
&\quad + \frac{(n-1)(n-2)}{n^3} \sum_{y} \sum_{x_1'} \sum_{x_2', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1', x_2'') P(x_1'|y) P(x_2'|y) P(x_2''|y) \\
&\quad + \frac{2(n-1)(n-2)}{n^3} \sum_{y} \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1', x_2''|y) P(x_1''|y) P(x_2'|y) \\
&\quad + \frac{(n-1)(n-2)}{n^3} \sum_{y} \sum_{x_2'} \sum_{x_1', x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2') P(x_1'|y) P(x_2'|y) P(x_1''|y) \\
&\quad + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_2'} \sum_{x_1', x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2') P(x_1'|y) P(x_1'', x_2'|y) \\
&\quad + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1'} \sum_{x_2', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1', x_2'') P(x_2'|y) P(x_1', x_2''|y) \\
&\quad + \sum_{y} \Big[\frac{n-1}{n^3} \frac{1}{P(y)} - \frac{2(n-1)}{n^2}\Big] \sum_{x_2'} \sum_{x_1', x_1''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2') P(x_1''|y) P(x_1', x_2'|y) \\
&\quad + \sum_{y} \Big[\frac{n-1}{n^3} \frac{1}{P(y)} - \frac{2(n-1)}{n^2}\Big] \sum_{x_1'} \sum_{x_2', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1', x_2'') P(x_2''|y) P(x_1', x_2'|y) \\
&\quad + \sum_{y} \Big[\frac{n-1}{n^3} \frac{1}{P(y)} - \frac{2(n-1)}{n^2} + \frac{n-1}{n} P(y)\Big] \\
&\qquad \cdot \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1', x_2'|y) P(x_1'', x_2''|y) \\
&\quad + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1', x_2''|y) P(x_1'', x_2'|y) \\
&\quad + \frac{n-1}{n^3} \sum_{y} \frac{1}{P(y)} \sum_{x_1', x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1', x_2') P(x_1'|y) P(x_2'|y) \\
&\quad + \sum_{y} \Big[\frac{1}{n^3} \frac{1}{P^2(y)} - \frac{2}{n^2 P(y)} + 1\Big] \sum_{x_1', x_2'} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1', x_2') P(x_1', x_2'|y)
\end{aligned}
\tag{42}
$$

where we obtain (42) from (29).

The cases of $k \geq 3$ can be given likewise.

D Details of Experiments

D.1 Training Loss

In this section, we demonstrate the approach to computing the training loss from sample features.

To compute the training loss $\tilde{\mathcal{L}}_{\text{train}}(\boldsymbol{f}, \boldsymbol{g})$, we first introduce the following lemma, which links the χ^2 divergence to H-Score which can then be expressed through the loss that we have given.

Lemma 5. *Let* $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ *be the features that minimize the* χ^2*-divergence loss* $\chi^2_R\left(\hat{P}_{X_1X_2Y}, \hat{P}_{X_1X_2}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}\right)$, *where* $\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}(x_1, x_2, y) \triangleq \hat{P}_Y(y)(1 + \boldsymbol{f}^T(x_1, x_2)\boldsymbol{g}(y))$, *for all* (x_1, x_2, y), *and the reference distribution is* $\hat{P}_{X_1X_2}\hat{P}_Y$. *Then, we have*

$$\mathbb{E}_{\hat{P}_{X_1X_2}}[\boldsymbol{f}^*(X_1, X_2)] = \mathbb{E}_{\hat{P}_Y}[\boldsymbol{g}^*(Y)] = \mathbf{0}, \tag{43}$$

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

$$H(\boldsymbol{f}, \boldsymbol{g}) \triangleq \mathbb{E}_{\hat{P}_{X_1X_2}}[\tilde{\boldsymbol{f}}^T(X_1, X_2)\tilde{\boldsymbol{g}}(Y)] - \frac{1}{2}\operatorname{tr}(\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}}\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}), \tag{44}$$

where $\tilde{\boldsymbol{f}}(X_1, X_2) \triangleq \boldsymbol{f}(X) - \mathbb{E}_{\hat{P}_{X_1X_2}}[\boldsymbol{f}(X_1X_2)]$, $\tilde{\boldsymbol{g}}(Y) \triangleq \boldsymbol{g}(Y) - \mathbb{E}_{\hat{P}_Y}[\boldsymbol{g}(Y)]$, $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}}$ *and* $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}}$ *are the covariance matrices of features on the training samples, defined as:*

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{f}} \triangleq \mathbb{E}_{\hat{P}_{X_1X_2}}[\tilde{\boldsymbol{f}}(X_1, X_2)\tilde{\boldsymbol{f}}^T(X_1, X_2)], \tag{45}$$

$$\hat{\boldsymbol{\Lambda}}_{\boldsymbol{g}} \triangleq \mathbb{E}_{\hat{P}_Y}[\tilde{\boldsymbol{g}}(Y)\tilde{\boldsymbol{g}}^T(Y)]. \tag{46}$$

We can derive the H-score for Markov structure $H^{(M)}(\boldsymbol{f}, \boldsymbol{g})$ *similarly. Then, the training loss can be implemented by*

$$(\boldsymbol{f}^*, \boldsymbol{g}^*) \leftarrow \arg\max_{\boldsymbol{f}, \boldsymbol{g}} (1 - \alpha)H(\boldsymbol{f}, \boldsymbol{g}) + \alpha H^{(M)}(\boldsymbol{f}, \boldsymbol{g}). \tag{47}$$

Unfold the loss functions in (47), we have our training loss function,

$$\begin{aligned}
&\tilde{\mathcal{L}}^{(\alpha)}_{\text{train}}(\boldsymbol{f}, \boldsymbol{g}) = (1 - \alpha)\mathcal{L}_{dep}(\boldsymbol{f}, \boldsymbol{g}) + \alpha\mathcal{L}^{(M)}_{dep}(\boldsymbol{f}, \boldsymbol{g}), \\
&\mathcal{L}_{dep}(\boldsymbol{f}, \boldsymbol{g}) \triangleq \frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{f}^{\mathrm{T}}(x_1^{(i)}, \ldots, x_k^{(i)})\boldsymbol{g}(y^{(i)}) - \frac{1}{2}\operatorname{tr}(\operatorname{cov}(\boldsymbol{f})\operatorname{cov}(\boldsymbol{g})), \\
&\mathcal{L}^{(M)}_{dep}(\boldsymbol{f}, \boldsymbol{g}) \triangleq \sum_{j=1}^{m} \hat{P}_Y(j)\left[\frac{1}{n_j - 1}\sum_{i=1}^{n_j} \boldsymbol{f}^{\mathrm{T}}(\underline{x}_1^{(i,j)}, \ldots, \underline{x}_k^{(i,j)})\boldsymbol{g}(j) - \frac{1}{2}\operatorname{tr}(\operatorname{cov}(\boldsymbol{f}_j)\operatorname{cov}(\boldsymbol{g}))\right].
\end{aligned}$$

As for loss $\mathcal{L}_{dep}(\boldsymbol{f}, \boldsymbol{g})$, the calculation is straightforward. With n training samples $(x_1^{(i)}, \ldots, x_2^{(i)}, y^{(i)}), i = 1, \ldots, n$, and two branches of parameterized neural networks with output units: $\boldsymbol{f}$ and $\boldsymbol{g}$, we can compute the loss in the following

$$\begin{aligned}
\boldsymbol{f}(x_1^{(i)}, \ldots, x_k^{(i)}) &\leftarrow \boldsymbol{f}(x_1^{(i)}, \ldots, x_k^{(i)}) - \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{f}(x_1^{(t)}, \ldots, x_k^{(t)}), i = 1, \ldots, n \\
\boldsymbol{g}(y^{(i)}) &\leftarrow \boldsymbol{g}(y^{(i)}) - \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{g}(y^{(t)}), i = 1, \ldots, n \\
\operatorname{cov}(\boldsymbol{f}) &\leftarrow \frac{1}{n-1}\sum_{t=1}^{n} \boldsymbol{f}(x_1^{(t)}, \ldots, x_k^{(t)})\boldsymbol{f}^{\mathrm{T}}(x_1^{(t)}, \ldots, x_k^{(t)}) \\
\operatorname{cov}(\boldsymbol{g}) &\leftarrow \frac{1}{n-1}\sum_{t=1}^{n} \boldsymbol{g}(y^{(t)})\boldsymbol{g}^{\mathrm{T}}(y^{(t)}).
\end{aligned}$$

Table 1: The optimal coefficients α^* derived by auto-CODES and grid search method on different training sample sizes on the IEMOCAP dataset.

Sample size n (dialogue size)	**grid search (gs)**		**auto-CODES (auto)**		
	α_{gs}	accuracy	α_{auto}	accuracy	$n \cdot \alpha_{auto}$
446 (10)	0.01	42.784 ± 1.465	0.0151 ± 0.0010	43.023 ± 1.706	6.73
565 (12)	0.01	44.754 ± 1.220	0.0129 ± 0.0009	45.478 ± 1.236	7.29
647 (14)	0.01	45.730 ± 1.420	0.0111 ± 0.0005	46.239 ± 1.128	7.18

As for loss $\mathcal{L}_{dep}^{(M)}(\boldsymbol{f}, \boldsymbol{g})$, it needs a permutation on samples' modalities within the subset of the same label. We denote the subset of training samples with label $j \in \{1, \dots, k\}$ as $\mathcal{D}_j = \{(x_1^{(i,j)}, \dots, x_k^{(i,j}))\}_{i=1}^{d_j}$, where d_j is the number of samples whose label is j in the overall dataset $\mathcal{D}$. $\underline{x}_t^{(i,j)}$ is chosen from $\{x_t^{(i,j)}\}_{i=1}^{d_j}, t = 1, \dots, k$, and $n_j = \prod_{t=1}^{k} d_t$. $\operatorname{cov}(\boldsymbol{f}_j) \leftarrow \frac{1}{n_j - 1} \sum_{t=1}^{n_j} \boldsymbol{f}(x_1^{(t,j)}, \dots, x_k^{(t,j)}) \boldsymbol{f}^{\mathrm{T}}(x_1^{(t,j)}, \dots, x_k^{(t,j)})$.

D.2 Optimal Coefficient in Testing Loss

There are 18 terms in the expression of the optimal α^*. Here we illustrate an example that they can be computed by the mean of features, and the other terms can be computed similarly.

As for term $\sum_y \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1', x_2'|y) P(x_1'', x_2''|y) \left(\dfrac{1}{n^2} - \dfrac{1}{n} P(y) \right)$, it can be computed by three parts.

First, we compute the covariance matrix $\Lambda_{\boldsymbol{f}}$ using features of data, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{f}(x_1^{(i)}, x_2^{(i)}) \boldsymbol{f}^{\mathrm{T}}(x_1^{(i)}, x_2^{(i)}).$$

Then, we computed $\sum_{x_1', x_2'} P(x_1', x_2'|y) \boldsymbol{f}(x_1', x_2')$ by $\dfrac{\sum_{i=1}^{n} \boldsymbol{f}(x_1^{(i)}, x_2^{(i)}) \mathbb{1}\{y^{(i)} = y\}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = y\}}$.

At last, we sum them up over $y \in \mathcal{Y}$.
In summary,

$$\begin{aligned}
&\sum_y \left(\frac{1}{n^2} - \frac{P(y)}{n} \right) \sum_{x_1', x_2'} \sum_{x_1'', x_2''} \boldsymbol{f}^{\mathrm{T}}(x_1', x_2') \Lambda_{\boldsymbol{f}}^{-1} \boldsymbol{f}(x_1'', x_2'') P(x_1', x_2'|y) P(x_1'', x_2''|y) \\
&= \sum_y \left(\frac{1}{n^2} - \frac{P(y)}{n} \right) \frac{\sum_{i=1}^{n} \boldsymbol{f}(x_1^{(i)}, x_2^{(i)}) \mathbb{1}\{y^{(i)} = y\}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = y\}} \cdot \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{f}(x_1^{(j)}, x_2^{(j)}) \boldsymbol{f}^{\mathrm{T}}(x_1^{(j)}, x_2^{(j)}) \\
&\qquad \cdot \frac{\sum_{\ell=1}^{n} \boldsymbol{f}(x_1^{(\ell)}, x_2^{(\ell)}) \mathbb{1}\{y^{(\ell)} = y\}}{\sum_{\ell=1}^{n} \mathbb{1}\{y^{(\ell)} = y\}}.
\end{aligned}$$

D.3 Results on IEMOCAP

The further results on IEMOCAP are shown in Table 1.