# A Non-asymptotic Framework for Characterizing Dependency Structures in Multimodal Learning

Weida Wang*, Tao Shi*, Yaoyuan Liang*, Xinyi Tong*, Shao-Lun Huang*†
*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
†Corresponding author

***Abstract*—Dependency structures between modalities have been utilized explicitly and implicitly in multimodal learning to enhance classification performance, particularly when the training samples are insufficient. Recent efforts have concentrated on developing mathematical frameworks utilizing conditional dependency structures, but the non-asymptotic relations between the training sample size and various structures are not sufficiently addressed. To address this issue, we propose a mathematical framework that can be utilized to characterize conditional dependency structures in analytic ways. It provides an explicit description of the sample size in learning various structures in a non-asymptotic regime. Additionally, it demonstrates how task complexity and a fitness evaluation of conditional dependence structures affect the results. Furthermore, we develop an autonomously updated coefficient algorithm auto-CODES based on the theoretical framework and conduct experiments on multimodal emotion recognition tasks using the MELD dataset. The experimental results validate our theory and show the effectiveness of the proposed algorithm.**

## I. Introduction

Multimodal learning is an active research area in machine learning with wide applications in audio-visual speech recognition (AVSR) [1], emotion recognition [2], and information retrieval [3], etc. It aims at jointly extracting information and learning knowledge from different categories of data, such as texts, audio, and images [4], to gain better and more robust results. Treating each modality as a random variable, training the joint distribution of multiple modalities often demands a larger sample size, compared to the single modality cases. Thus, a critical issue is to exploit heterogeneous dependency structures across modalities, such that the label information can be effectively extracted for classification, especially when the number of training samples is insufficient. Addressing this issue, a mathematical framework has been proposed in [5] using the minimax principle. Especially, as illustrated in Fig. 1, a conditional dependency structure, in which the modalities and the label form a Markov chain, has been emphasized. Based on the framework, the authors in [5] show that the optimal estimator for the joint multimodal distribution can be approximated by the linear combination of two estimators considering the general structure and the conditional dependency structure, respectively. Note that the conclusion is given under two constraints: i) the sample size is under a certain regime, and ii) the true structure across multiple modalities is close to the conditional dependency structure. However, the actual scenes may not satisfy those constraints.

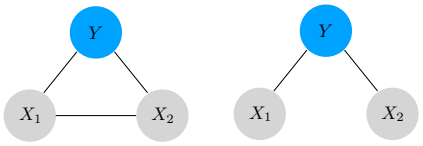


Fig. 1. Two dependency structures of modalities $X_1, X_2$ and labels $Y$, where the edges indicate dependencies. The left figure represents the general dependency structure, and the right one describes the conditional dependency $X_1 - Y - X_2$ where modalities are independent once their label information $Y$ is given.

In this work, we propose a non-asymptotic framework to characterize the dependency structures in multimodal learning, based on the framework in [5]. We first consider the multiple modalities in discrete random variable cases to give theoretical insights and interpretations. Specially, we adopt the joint distribution estimator as a linear combination of the mentioned estimators considering two different dependency structures with a combing coefficient to be determined through a testing loss. The proposed testing loss considers the average estimating performance of the linearly combined estimator under a given training sample size. By minimizing the testing loss, the optimal coefficient can be expressed analytically. Moreover, based on the expressions, we conclude that this similarity measurement is inversely proportional to the training sample size and the distance between two distributions representing two considered structures. In addition, it is proportional to the model complexity which characterizes the number of parameters used to represent the model. The results hold true for all numbers of sample sizes and the coefficient itself can be viewed as a similarity measurement to characterize how close the true structure is to the conditional dependency one.

On the other hand, we extend our analyses from discrete random variables to continuous ones by exploiting parametric models. Moreover, we develop a multimodal algorithm to implement our theoretical conclusions, based on the Soft-HGR multimodal algorithm [6]. The proposed algorithm can update the combining coefficient autonomously using the learned features from different modalities. In addition,

979-8-3503-4893-4/24/$31.00 ©2024 IEEE

we conduct experiments on multimodal emotion recognition tasks with a challenging dataset MELD, using the proposed autonomous updated coefficient on dependency structures (auto-CODES) algorithm. Experimental results demonstrate that the proposed auto-CODES algorithm outperforms existing approaches, and validates our theory that the optimal coefficient is inversely proportional to the training sample size.

## II. Problem Formulation and Analysis

### A. Linearly Combined Estimator

Let random variables $X_1$, $X_2$ and $Y$ denote two modalities and their corresponding label over finite alphabets $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{Y}$, respectively. Then, a multimodal dataset with $n$ sample tuples $\mathcal{D} \triangleq \{(x_1^{(i)}, x_2^{(i)}, y^{(i)})\}_{i=1}^n$ are generated in an independent, identically distributed (i.i.d.) manner from the true joint distribution $P_{X_1X_2Y}$, where $P_{X_1X_2Y}(x_1, x_2, y) > 0$ for all entries. Specifically, we consider two different estimators to approximate the joint distribution $P_{X_1X_2Y}$: (i) the empirical joint distribution $\hat{P}_{X_1X_2Y}$, and (ii) the empirical Markov-structured distribution $\hat{P}^{(\mathrm{M})}_{X_1X_2Y}$ characterizing the conditional dependency structure $X_1 - Y - X_2$, where

$$\hat{P}_{X_1X_2Y}(x_1, x_2, y) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_1^{(i)} = x_1\}\mathbb{1}\{x_2^{(i)} = x_2\}\mathbb{1}\{y^{(i)} = y\}, \tag{1a}$$

$$\hat{P}^{(\mathrm{M})}_{X_1X_2Y}(x_1, x_2, y) \triangleq \hat{P}_{X_1|Y}(x_1|y)\hat{P}_{X_2|Y}(x_2|y)\hat{P}_Y(y), \tag{1b}$$

and where $\mathbb{1}\{\cdot\}$ denotes the indicator function, $\hat{P}_Y$ denotes the marginal empirical distribution of labels $Y$ and $P^{(\mathrm{M})}$ denotes $P_{X_1|Y}P_{X_2|Y}P_Y$. For simplicity, we consider the case where the label distribution has been learned well, i.e., $\hat{P}_Y(y) = P_Y(y)$, for $y \in \mathcal{Y}$.

We aim to design $\tilde{P}_{X_1X_2Y}$ to estimate the true joint distribution $P_{X_1X_2Y}$. Specifically, the estimator is a linear combination of the two aforementioned estimators, i.e.,

$$\tilde{P}_{X_1X_2Y} \triangleq (1-\alpha)\cdot\hat{P}_{X_1X_2Y} + \alpha\cdot\hat{P}^{(\mathrm{M})}_{X_1X_2Y} \tag{2}$$

where the coefficient $\alpha \in [0, 1]$ is the parameter to be designed.

### B. Optimal Combination Coefficient

To measure the performance of estimator (2), we propose a testing loss based on the referenced $\chi^2$-divergence, which is defined as follows. [1]

**Definition 1.** *For discrete random variable $Z$ over finite alphabet $\mathcal{Z}$, and its distributions $P_Z$ and $Q_Z$, with reference distribution $R_Z$, the referenced $\chi^2$-divergence between $P_Z$ and $Q_Z$ is defined as:*

$$\chi^2_{R_Z}(P_Z, Q_Z) \triangleq \sum_{z\in\mathcal{Z}} \frac{(P_Z(z) - Q_Z(z))^2}{R_Z(z)}, \tag{3}$$

*where we denote $\chi^2(P_Z, Q_Z) \triangleq \chi^2_{P_Z}(P_Z, Q_Z)$, which corresponds to the Pearson $\chi^2$-divergence.*

Based on the referenced $\chi^2$-divergence, we define the testing loss as the average divergence between the dependency estimator (2) and the true joint distribution under the fixed training sample size $n$.

**Definition 2.** *For estimator $\tilde{P}_{X_1X_2Y}$ with coefficient $\alpha$ and the corresponding true distribution $P_{X_1X_2Y}$, the testing loss and the optimal coefficient $\alpha^*$ are defined as:*

$$\tilde{\mathcal{L}}_{\mathrm{dep}}(\alpha) \triangleq \mathbb{E}\Big[\chi^2(P_{X_1X_2Y}, \tilde{P}_{X_1X_2Y})\Big], \tag{4}$$

$$\alpha^* \triangleq \arg\min_{\alpha\in[0,1]} \tilde{\mathcal{L}}_{\mathrm{dep}}(\alpha), \tag{5}$$

*where the expectation is taken over all $n$ i.i.d. samples generated from the true distribution.*

Then, we have the following characterization of our proposed testing loss (4) over the linearly combined dependency estimator (2) and the optimal combining coefficient $\alpha^*$ (5).

**Theorem 3.** *The testing loss (4) can be expressed as:*

$$\tilde{\mathcal{L}}_{\mathrm{dep}}(\alpha) = \left(\frac{1}{n}C + \frac{1}{n}V + \chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y})\right)\cdot\alpha^2 - \frac{2}{n}C\cdot\alpha + \frac{1}{n}(|\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}| - 1), \tag{6}$$

*and the optimal coefficient $\alpha^*$ to minimize the $\chi^2$-divergence dependency loss (4) can be given as:*

$$\alpha^* = \frac{\frac{1}{n}C}{\chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y}) + \frac{1}{n}C + \frac{1}{n}V}, \tag{7}$$

*where*

$$C \triangleq |\mathcal{Y}|\cdot\Big[|\mathcal{X}_1||\mathcal{X}_2| - (|\mathcal{X}_1| + |\mathcal{X}_2|)\Big] + 1 + a_n, \tag{8}$$

$$V \triangleq -6\cdot\chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y}) + |\mathcal{Y}|(|\mathcal{X}_1| + |\mathcal{X}_2|) - 2 + 2\sum_{x_2,y}\chi^2(P_{X_1|X_2Y}, P_{X_1|Y}) + 2\sum_{x_1,y}\chi^2(P_{X_2|X_1Y}, P_{X_2|Y}) + b_n, \tag{9}$$

*where $a_n$ and $b_n$ are of the order $O(\frac{1}{n})$, which will go to constants when $n$ goes to infinity.*

By considering the conditional dependency structure and tuning the coefficient $\alpha$, the improvement from the unbiased estimator (1a) to the optimal dependency estimator can be calculated through the differences in their corresponding testing losses.

[1] Conventionally, such performance is conventionally computed by logarithm loss. However, in our setting, it will be ill-defined when some $(x_1, x_2, y)$ tuple is missing in training samples. By that time, we have $\hat{P}_{X_1X_2Y}(x_1, x_2, y) = 0$ while $P_{X_1X_2Y}(x_1, x_2, y) > 0$, which would bring the logarithm loss to infinite.

**Corollary 4.** *The improvement of considering the optimal coefficient $\alpha^*$ can be given as:*

$$\tilde{\mathcal{L}}_{dep}(0) - \tilde{\mathcal{L}}_{dep}(\alpha^*) = \frac{1}{n} \cdot \frac{C^2}{C + V + n \cdot \chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y})}, \tag{10}$$

*where parameters $C$ and $V$ are defined in Theorem 3.*

From (7), we can notice that the optimal combining coefficient $\alpha^*$ is determined by three major factors: (i) the training sample size $n$, (ii) the fitness of the conditional dependency structure to describe the true underlying distribution, measured by $\chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y})$ and terms in the parameter $V$, and (iii) the task complexity $C$, characterized by the number of parameters needed to estimate the joint distribution. The last characterization comes from the fact that when the task is to learn all the entries of the true distribution, the number of parameters required corresponds to the cardinality of the sample space.

Most importantly, based on (7), we show that the optimal coefficient $\alpha^*$ is inversely proportional to the number of training samples and the fitness measure of the conditional dependency structure to estimate the true one. In addition, it is proportional to the task complexity measured by the number of model parameters. With the optimal coefficient $\alpha^*$, we obtain the optimal dependency estimator $\tilde{P}^*_{X_1X_2Y}$, which represents the most appropriate dependency structure to approximate the true joint distribution. Our work generalizes the conclusions in [5] and show the explicit relation between training sample size and different dependency structures in a non-asymptotic regime.

In addition, to better understand Theorem 3 and Corollary 4, we discuss two special cases as follows.

*Case 1*: When the true dependency structure is Markovian, i.e. $X_1 - Y - X_2$, the optimal coefficient will becomes $1 - V(C+V)^{-1}$, which is nearly $1 - |\mathcal{X}_1|^{-1} - |\mathcal{X}_2|^{-1}$. Since the cardinality terms $|\mathcal{X}_1|$ and $|\mathcal{X}_2|$ are usually large, the optimal coefficient $\alpha^*$ is quite close to 1, representing that the true joint distribution should be close to a Markov-structured conditional distribution [2], the improvement from the unbiased estimator to the optimal dependency estimator is also relatively large.

*Case 2*: When the training sample size is small, the optimal coefficient $\alpha^*$ approaches 1, indicating a "near Markov" model where performance improves by accounting for conditional dependencies. Unlike existing methods, our approach optimally adjusts the combining coefficient based on sample size, dependency structure fit, and task complexity.

Last but not least, the $\chi^2$-divergence dependency loss (4) can be interpreted as a bias-variance trade-off tuned by the coefficient $\alpha$, and the optimal bias-variance trade-off can be achieved when the $\chi^2$-divergence dependency loss is minimized. This unique interpretation is defined as:

$$\tilde{\mathcal{L}}_{\text{dep}}(\alpha) = \underbrace{\frac{1}{n}\left(C\alpha^2 + V\alpha^2 + 2C\alpha + |\mathcal{X}_1||\mathcal{X}_2||\mathcal{Y}| - 1\right)}_{\text{variance term(s)}} + \underbrace{\alpha^2\chi^2(P_{X_1X_2Y}, P^{(M)}_{X_1X_2Y})}_{\text{bias term}} \tag{11}$$

The variance terms will vanish as the number of training samples increases. Besides, the bias term characterizes the cost of utilizing the conditional dependency structure to approximate the true distribution.

## III. auto-CODES: Autonomous Updated Coefficient on Dependency Structure

In this section, we propose an algorithm named autonomous updated coefficient on dependency structures (auto-CODES) as a realization of our aforementioned theoretical framework. Specifically, we first extend our theory from discrete to the continuous domain using representations in factorization form. Then, the expression of the optimal coefficient $\alpha^*$ is given by multimodal features, and the objective loss function is designed as a linear combination of general and conditional dependency structures. Finally, we give the proposed auto-CODES algorithm and the discrimination rule using maximum a posterior (MAP).

### A. Multimodal Representations in Factorization Form

To apply our mathematical framework, we introduce a parameterized model for continuous data density. The model has two parts: first, an early fusion model concatenates modality-specific representations, which are then passed through a neural network to learn a joint representation, $\boldsymbol{f}$. Second, an embedding layer transforms one-hot encoded labels into dense representations, $\boldsymbol{g}$. The joint training optimizes both $\boldsymbol{f}$ and $\boldsymbol{g}$ simultaneously.

Our framework considers an inference model $\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}$ in the following factorization form[3], which is widely used in natural language processing [7] and image recognition [8]:

$$\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{Y|X_1X_2}(y|x_1,x_2) \triangleq P_Y(y)(1 + \langle \boldsymbol{f}(x_1,x_2), \boldsymbol{g}(y)\rangle). \tag{12}$$

Analogous to the linearly combined dependency estimator (2), we consider the linear combination of two types of inference models, representing the general dependency structure and conditional dependency structure respectively, which is defined as:

$$Q^{(\alpha)}_{Y|X_1X_2} \triangleq (1-\alpha)\tilde{P}^{(\boldsymbol{f}^*_0,\boldsymbol{g}^*_0)}_{Y|X_1X_2} + \alpha\tilde{P}^{(\boldsymbol{f}^*_1,\boldsymbol{g}^*_1)}_{Y|X_1X_2} \tag{13}$$

$$(\boldsymbol{f}^*_0, \boldsymbol{g}^*_0) \triangleq \arg\min_{\boldsymbol{f_0},\boldsymbol{g_0}} \chi^2_R\left(\hat{P}_{X_1X_2Y}, P_{X_1X_2}\tilde{P}^{(\boldsymbol{f}_0,\boldsymbol{g}_0)}_{Y|X_1X_2}\right) \tag{14}$$

$$(\boldsymbol{f}^*_1, \boldsymbol{g}^*_1) \triangleq \arg\min_{\boldsymbol{f_1},\boldsymbol{g_1}} \chi^2_R\left(\hat{P}^{(M)}_{X_1X_2Y}, P_{X_1X_2}\tilde{P}^{(\boldsymbol{f}_1,\boldsymbol{g}_1)}_{Y|X_1X_2}\right), \tag{15}$$

[2]Due to the consideration of a limited number of training samples and the assumption on the distribution of the label $Y$, it will not be strictly 1.

[3]Note that it can be negative in real applications. But we can also use it to make discriminative decisions through the maximum a posterior (MAP) rule.

where the reference distribution $R \triangleq P_{X_1X_2}P_Y$.

Furthermore, we define the testing loss and the corresponding optimal coefficient $\alpha^*$ as

$$\tilde{\mathcal{L}}_{\text{test}}^{(\boldsymbol{f},\boldsymbol{g})}(\alpha) \triangleq \mathbb{E}\Big[\chi_R^2\Big(P_{X_1X_2Y}, P_{X_1X_2}Q_{Y|X_1X_2}^{(\alpha)}\Big)\Big], \quad (16)$$

$$\alpha^* \triangleq \arg\min_{\alpha\in[0,1]} \tilde{\mathcal{L}}_{\text{dep}}^{(\boldsymbol{f},\boldsymbol{g})}(\alpha). \quad (17)$$

### B. Our Proposed Algorithm

Based on the linear estimator (13), the objective function that we aim to optimize during the training phase can be chosen as a linear combination of two referenced $\chi^2$-divergences, which measures the distances between the learned distribution and distributions corresponding to two different dependency structures:

$$\tilde{\mathcal{L}}_{\text{train}}^{(\alpha)}(\boldsymbol{f},\boldsymbol{g}) = (1-\alpha)\chi_R^2\Big(\hat{P}_{X_1X_2Y}, \hat{P}_{X_1X_2}\hat{P}_{Y|X_1X_2}^{(\boldsymbol{f},\boldsymbol{g})}\Big) + \alpha\chi_R^2\Big(\hat{P}_{X_1X_2Y}^{(M)}, \hat{P}_{X_1X_2}\hat{P}_{Y|X_1X_2}^{(\boldsymbol{f},\boldsymbol{g})}\Big) \quad (18)$$

According to the Soft-HGR multimodal algorithm established in [6], [9], the objective (18) can be transformed into the following loss function that can be computed by the multimodal and label features $(\boldsymbol{f},\boldsymbol{g})$:

$$\tilde{\mathcal{L}}_{\text{train}}^{(\alpha)}(\boldsymbol{f},\boldsymbol{g}) = (1-\alpha)\mathcal{L}(\boldsymbol{f},\boldsymbol{g}) + \alpha\mathcal{L}^{(M)}(\boldsymbol{f},\boldsymbol{g}), \quad (19)$$

$$\mathcal{L}(\boldsymbol{f},\boldsymbol{g}) = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{f}^{\mathrm{T}}(x_1^{(i)},\ x_2^{(i)})\boldsymbol{g}(y^{(i)}) - \frac{1}{2}\operatorname{tr}(\operatorname{cov}(\boldsymbol{f})\operatorname{cov}(\boldsymbol{g})) \quad (20)$$

$$\mathcal{L}^{(M)}(\boldsymbol{f},\boldsymbol{g}) = \sum_{j=1}^{m}\hat{P}_Y(j)\Big(\frac{1}{n_j-1}\sum_{i=1}^{n_j}\boldsymbol{f}^{\mathrm{T}}(\underline{x}_1^{(i,j)},\ \underline{x}_2^{(i,j)})\boldsymbol{g}(j) - \frac{1}{2}\operatorname{tr}(\operatorname{cov}(\boldsymbol{f}_j)\operatorname{cov}(\boldsymbol{g}))\Big), \quad (21)$$

where $\hat{P}_Y(j) = \sum_{i=1}^{n}\mathbb{1}\{y^{(i)} = j\}, j = 1,\dots,m$, $\boldsymbol{f}(x_1^{(i)},x_2^{(i)})$ is the extracted feature representation of the $i$-th sample $(x_1^{(i)},x_2^{(i)},y^{(i)})$, $\boldsymbol{g}(i)$ is the embedding for label $i$.

In (19), $\operatorname{cov}(\boldsymbol{f})$, $\operatorname{cov}(\boldsymbol{g})$ and $\hat{P}_Y(j)$ can be computed from the data.

$$\operatorname{cov}(\boldsymbol{f}) \leftarrow \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{f}(x_1^{(i)},\dots,x_k^{(i)})\boldsymbol{f}^{\mathrm{T}}(x_1^{(i)},\dots,x_k^{(i)}),$$

$$\operatorname{cov}(\boldsymbol{g}) \leftarrow \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{g}(y^{(i)})\boldsymbol{g}^{\mathrm{T}}(y^{(i)}).$$

In (21), when calculating $\mathcal{L}^{(M)}(\boldsymbol{f},\boldsymbol{g})$ which represents the loss for the conditional dependency structure, it needs a permutation on training samples' all modalities within the subset of the same label. We denote the subset of training samples with label $j \in \{1,\dots,m\}$ as $\mathcal{D}_j = \{(x_1^{(i,j)},x_2^{(i,j)})\}_{i=1}^{d_j}$, where $d_j$ is the number of samples whose label is $j$ in the original dataset $\mathcal{D}$. Entry $\underline{x}_t^{(i,j)}$ is randomly chosen from $D_t, t = 1,\dots,m$, and $n_j = \prod_{t=1}^{m} d_t$. Then $\operatorname{cov}(\boldsymbol{f}_j)$ can be approximated through: $\operatorname{cov}(\boldsymbol{f}_j) \leftarrow \frac{1}{n_j-1}\sum_{t=1}^{n_j}\boldsymbol{f}(x_1^{(t,j)},x_2^{(t,j)})\boldsymbol{f}^{\mathrm{T}}(x_1^{(t,j)},x_2^{(t,j)})$.

Based on the above analysis, our auto-CODES algorithm can be implemented as an iteration of two main optimizations: (i) the optimization of coefficient $\alpha$ for given $(\boldsymbol{f},\boldsymbol{g})$ by minimizing the $\chi^2$-divergence dependency loss $\tilde{\mathcal{L}}_{\text{dep}}(\alpha)$ (16); (ii) the optimization of feature pairs $(\boldsymbol{f},\boldsymbol{g})$ for given $\alpha$ to minimize the training loss (18) through the deep neural network. The proposed auto-CODES algorithm is summarized in Algorithm 1.

Finally, after training the auto-CODES algorithm for multiple epochs and obtaining the optimal extracted features $\boldsymbol{f}^*$ and $\boldsymbol{g}^*$, the classification of a newly observed sample $(x_1^{\text{new}},x_2^{\text{new}})$ can be performed by the maximum a posterior (MAP) decision rule:

$$\begin{aligned}&\tilde{y}(x_1,x_2)\\ &= \arg\max_{y\in\mathcal{Y}} P_{Y|X_1X_2}(y|x_1^{\text{new}},\ x_2^{\text{new}})\\ &= \arg\max_{y\in\mathcal{Y}} P_Y(y)(1+\langle\boldsymbol{f}^*(x_1^{\text{new}},\ x_2^{\text{new}}),\boldsymbol{g}^*(y)\rangle).\end{aligned}$$

**Algorithm 1** An Auto-updated Coefficient on Dependency Structures (auto-CODES) Algorithm

**Input:** multimodal data samples $\{(x_1^{(i)},x_2^{(i)},y^{(i)})\}_{i=1}^n$
Initialize $\alpha^* = 0$
**repeat**
  $(\boldsymbol{f}^*,\boldsymbol{g}^*) \leftarrow \arg\min_{\boldsymbol{f},\boldsymbol{g}} \tilde{\mathcal{L}}_{\text{train}}^{(\alpha^*)}(\boldsymbol{f},\boldsymbol{g})$
  $\alpha^* \leftarrow \arg\min_{\alpha\in[0,1]} \tilde{\mathcal{L}}_{\text{test}}^{(\boldsymbol{f}^*,\boldsymbol{g}^*)}(\alpha)$
**until** $\alpha^*$ converges
$(\boldsymbol{f}^*,\boldsymbol{g}^*) \leftarrow \tilde{\mathcal{L}}_{\text{train}}^{(\alpha^*)}(\boldsymbol{f},\boldsymbol{g})$
**return** $\boldsymbol{f}^*,\boldsymbol{g}^*,\alpha^*$

## IV. Experiments

To validate our framework, we apply the proposed auto-CODES algorithm to multimodal emotion recognition using the Multimodal EmotionLines Dataset (MELD) [10]. This challenging task involves predicting emotion labels for each utterance in a conversation through multiple modalities, including text and audio. MELD, derived from the TV series Friends, contains 13,708 utterances across 1,433 dialogues, featuring multi-speaker interactions. Each utterance is labeled with both coarse sentiment (positive, neutral, negative) and fine-grained emotions (anger, disgust, sadness, joy, neutral, surprise, fear). We evaluate performance using accuracy and weighted-average F1-score, with results averaged over 20 runs. To verify the effectiveness of auto-CODES, we compare several loss functions.

To extract multimodal features $\boldsymbol{f}$, we utilize DialogueRNN [11] as our backbone network, which models contextual information and speaker information in conversations by employing three different Gated Recurrent Units (GRUs) [12]: global GRU, speaker GRU, and emotion GRU. The speaker-modeling nature of DialogueRNN makes it

TABLE 1
THE COMPARISON BETWEEN AUTO-CODES AND OTHER LOSS FUNCTIONS ON MELD DATASET IN DIFFERENT TRAINING SAMPLE SIZE SETTINGS.

| Sample Size (Dialogue Size) | Metric | Method | | | | |
|---|---|---|---|---|---|---|
| | | CE | MaskedNLL | Soft-HGR | Focal | auto-CODES |
| 107 (10) | accuracy | 50.222±2.709 | 50.544±1.380 | 50.758±2.005 | 50.784±2.143 | **53.640±1.345** |
| | F1-score | 40.099±2.296 | 42.147±2.269 | 43.649±1.999 | 43.998±2.227 | **47.206±1.179** |
| 302 (30) | accuracy | 51.272±1.467 | 52.084±0.990 | 52.655±1.421 | 53.036±1.576 | **55.943±0.925** |
| | F1-score | 44.091±1.275 | 45.951±1.527 | 46.589±1.986 | 47.358±1.170 | **50.988±0.712** |
| 516 (50) | accuracy | 54.176±1.098 | 53.586±1.042 | 55.613±1.065 | 55.901±0.937 | **58.467±0.783** |
| | F1-score | 49.397±1.100 | 48.537±0.934 | 50.509±1.700 | 50.107±0.776 | **52.138±0.872** |
| 1389 (150) | accuracy | 62.644±0.702 | 62.875±0.622 | 63.278±0.755 | 63.157±0.626 | **63.831±0.581** |
| | F1-score | 61.657±0.783 | 61.649±0.589 | 62.616±0.801 | 62.457±0.578 | **62.884±0.699** |
| 11098 (1152) (Full Training Set) | accuracy | 64.995±0.436 | 65.364±0.517 | 65.900±0.386 | 66.482±0.485 | **67.203±0.443** |
| | F1-score | 65.027±0.518 | 65.349±0.450 | 65.687±0.602 | 65.595±0.415 | **66.191±0.506** |

TABLE 2
THE RELATION BETWEEN THE OPTIMAL COEFFICIENT $\alpha^*$ DERIVED BY AUTO-CODES AND DIFFERENT SAMPLE SIZE $n$ ON MELD.

| MELD | | |
|---|---|---|
| **Sample Size** $n$ | $\alpha^*$ | $n \cdot \alpha^*$ |
| 107 | 0.0496 ± 0.0011 | 5.31 |
| 302 | 0.0172 ± 0.0009 | 5.19 |
| 516 | 0.0101 ± 0.0008 | 5.21 |
| 1389 | 0.0035 ± 0.0006 | 4.86 |
| 11098 | 0.0005 ± 0.0001 | 5.55 |

suitable for multi-party dialogue scenarios in MELD, as it is more effective in identifying the emotions of different speakers.

As mentioned in Section III-A, we employ early fusion as the fusion mechanism, in which individual modalities are concatenated after being extracted from the dataset and then sent into a deep neural network to learn a joint representation $\boldsymbol{f}$.

**Hyperparameter Setting:** We use a batch size of 16 and Adam optimizer [13] with $\beta_1 = 0.9$, $\beta_2 = 0.99$. Training runs for 100 epochs with an initial learning rate of $10^{-4}$, decaying by 0.95 every 10 epochs. Dropout [14] with a rate of 0.1 is applied to all layers of DialogueRNN to prevent overfitting. The dimensions of multimodal feature $\boldsymbol{f}$ and label feature $\boldsymbol{g}$ are both 32. The validation set is 10% of the training data.

**(1) Cross-Entropy Loss (CE)**: Standard classification loss. **(2) MaskedNLL**: Handles zero-padding by masking padded samples [11]. **(3) Soft-HGR Loss**: Captures shared information across modalities without hard whitening [6]. **(4) Focal Loss**: Tackles class imbalance by focusing on hard-to-classify samples [15].

First, the comparisons between our proposed auto-CODES and baseline models on the MELD dataset under different settings of training sample size are presented in Table 1. The results demonstrate that auto-CODES consistently outperforms existing methods in all training sample size settings, which proves the effectiveness and superiority of our proposed method at both small-scale and large-scale regimes. In Table 1, the observed improvements of auto-CODES over other baselines at small-scale sample size settings are more significant than large-scale sample size scenarios. From the dialogue size of 10 to 50, auto-CODES gains average relative improvements of $5.23\%$ on accuracy and $6.06\%$ on the F1-score against the second-best baseline model. However, for larger dialogue sizes ranging from 150 to 1152, the improvements of auto-CODES over baseline methods on average are $0.88\%$ and $0.43\%$ on accuracy and F1-score metrics, respectively. This phenomenon may be attributed to the fact that when the training dataset size is relatively large, the improvement brought by our method will become less noticeable compared with small-scale scenarios, as illustrated in Corollary 4.

Moreover, we investigate the interplay between the optimal coefficient $\alpha^*$ and different numbers of training samples $n$. As shown in the table Table 2, $n \cdot \alpha^*$ is approximately a constant value at all training sample size settings, with averages of 5.18. In other words, the optimal coefficient $\alpha^*$ is nearly inversely proportional to $n$ which validates our theoretical result.

## V. CONCLUSION

We propose a theoretical framework to characterize the exact relationship between sample size and conditional dependency structures in multimodal learning. Based on this, we introduce a weighted training algorithm, auto-CODES, which updates coefficients iteratively across dependency structures. Experiments on the MELD dataset demonstrate its effectiveness in multimodal emotion recognition.

**Acknowledgements.** The research of Shao-Lun Huang is supported in part by National Key R&D Program of China under Grant 2021YFA0715202, Shenzhen Ubiquitous Data Enabling Key Lab under Grant ZDSYS20220527171406015, the Shenzhen Science and Technology Program under Grant KQTD20170810150821146 and Meituan. The authors thank Xiangxiang Xu and Lizhong Zheng for their discussions and suggestions.

## REFERENCES


[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang, "Learning better representations for audio-visual emotion recognition with common information," *Applied Sciences*, vol. 10, no. 20, p. 7239, 2020.

[3] M. Li, S.-L. Huang, and L. Zhang, "A general framework for incomplete cross-modal retrieval with missing labels and missing modalities," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4763–4767.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[5] T. Peng, W. Wang, and S.-L. Huang, "A mathematical framework to characterize the dependency structures in multimodal learning with minimax principle," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1489–1494.

[6] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5281–5288.

[7] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," *Advances in neural information processing systems*, vol. 27, 2014.

[8] X. Xu and S.-L. Huang, "Maximal correlation regression," *IEEE Access*, vol. 8, pp. 26 591–26 601, 2020.

[9] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "An information-theoretic approach to universal feature selection in high-dimensional inference," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1336–1340.

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[11] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating deep neural networks using focal loss," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 288–15 299, 2020.