# LEARNING FROM INTERVAL-VALUED DATA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The classification problem concerning crisp-valued data has been well resolved. However, interval-valued data, where all of the observations' features are described by intervals, is also a common type of data in real-world scenarios. For example, the data extracted by many measuring devices are not exact numbers but intervals. In this paper, we focus on a highly challenging problem called *learning from interval-valued data* (LIND), where we aim to learn a classifier with high performance on interval-valued observations. First, we obtain the estimation error bound of the LIND problem based on Rademacher complexity. Then, we give the theoretical analysis to show the strengths of multi-view learning on classification problems, which inspires us to construct a new framework called *multi-view interval information extraction* (Mv-IIE) approach for improving classification accuracy on interval-valued data. The experiment comparisons with several baselines on both synthetic and real-world datasets illustrate the superiority of the proposed framework in handling interval-valued data. Moreover, we describe an application of the Mv-IIE framework that we can prevent data privacy leakage by transforming crisp-valued (raw) data into interval-valued data.

## 1 INTRODUCTION

Machine learning methods for the classification problem (Pan et al., 2018; Li et al., 2018) have made great achievements in many areas, including medical imaging (Raghu et al., 2019), natural language processing (Otter et al., 2020), biology (Llorente et al., 2021) and computer vision (Tran et al., 2019). The well-known classification machine learning algorithms incorporate logistic regression (Efron, 1975; Kayabol, 2020), support vector machines (Noble, 2006; Kafai & Eshghi, 2019), random forests (Breiman, 2001; Biau, 2012) and neural networks (Anderson, 1995; Zhang et al., 2021). Moreover, the theoretical analysis of these well-known algorithms has been well researched by applying different types of complexity, such as Rademacher complexity (Bartlett et al., 2006; Mohri et al., 2012) and VC-dimension (Mohri et al., 2012; Daniely & Shalev-Shwartz, 2014). Most existing works for the classification problem only focus on crisp-valued data classification.

However, in many real-world scenarios, observations with crisp-valued features are not always available. Interval-valued data (Dombi, 1990; Billard & Diday, 2003) is a common type of data where all of the observations' features are described by intervals, not crisp-valued numbers. For example, the mushroom dataset (see Table 1) is a real-world interval-valued dataset described by five interval-valued features and one category variable. Moreover, the data extracted by many measuring devices are not exact numbers but intervals because there are only a limited number of decimals available on most of these measuring devices. Existing well-known machine learning methods cannot be directly used to handle interval-valued data. Recently, some researchers have begun exploring imprecise data from different perspectives, such as superset label learning and data disambiguation Cour et al. (2011); Lin & Cercone (2012); Hüllermeier (2014); Liu & Dietterich (2014). Unfortunately, the existing research related to analyzing interval-valued data mainly focuses on decision-making (Jahanshahloo et al., 2006), clustering analysis (De Carvalho & Tenório, 2010), regression analysis (Hao, 2009; Utkin & Coolen, 2011; Souza et al., 2017), and feature selection (Li et al., 2022), yet less on classification tasks (Utkin & Coolen, 2011). Besides, limited research on interval-valued classification only gives some simple framework and no relevant experimental analysis on real-world interval-valued datasets.

In this paper, we focus on a highly challenging problem called *learning from interval-valued data* (LIND), where we aim to learn a classifier that can obtain high classification accuracy on interval-

Table 1: Some instances of the mushroom dataset. The first column of this table shows the name of each instance. The 2nd-6th columns of this table are five interval-valued features of the mushroom dataset, and the last column of this table shows the category of each instance (label).

| Name | Pw(cm) | Sl(cm) | St(cm) | Sma(cm) | Smi($\mu$m) | Category |
|---|---|---|---|---|---|---|
| Arorae | [3, 8] | [4, 9] | [0.5, 2.5] | [4.5, 5] | [3, 3.5] | Agaricus |
| Moronii | [6, 12] | [2, 7] | [1.5, 3] | [6, 7.5] | [4, 5] | Agaricus |
| Appendiculatus | [7, 14] | [5, 9] | [3, 6] | [11.5, 13.5] | [3.5, 4.5] | Boletus |
| Fragans | [6, 15] | [4, 10] | [1, 3.5] | [13, 17.5] | [5, 7.8] | Boletus |
| Augusta | [6, 12] | [9, 17] | [1, 2] | [9.5, 11.5] | [8.5, 10] | Amanita |

valued observations. Throughout existing research involving interval-valued data, no research discusses a theory regarding the interval-valued data classification problem. To fill this gap, we first present theoretical analysis to obtain the estimation error bound of the LIND problem based on Rademacher complexity (Theorem 1). This Rademacher complexity-based bound demonstrates that we can always train a classifier with high classification accuracy when enough interval-valued instances can be collected. Next, we provide a theorem to show the strengths of multi-view learning in addressing classification problems (Theorems 4 and 5). This theorem inspires us to propose a new framework called the *multi-view interval information extraction* (Mv-IIE) approach using multi-view learning (Blum & Mitchell, 1998; Zhang et al., 2018a; Wang et al., 2021a;b).

The proposed framework, which comprises two main parts (Figure 1), applies multi-view learning to classify crisp-valued information extracted from the interval-valued observations. The *first part* is used to extract crisp-valued information from the interval-valued observations. The most commonly used method is to take the midpoint of the intervals to extract crisp-valued information, however using this method will result in the loss of a lot of critical information from the intervals. For example, suppose we have two intervals $\bar{x}_1 = [1, 5]$ and $\bar{x}_2 = [2, 4]$, we will obtain the same crisp-valued information $x = 3$ from different intervals by taking the midpoint of the intervals. However, $\bar{x}_1$ clearly has a larger interval than $\bar{x}_2$ has, thus it is improper to consider them as the same instance in the view of midpoint. Therefore, in this paper, we propose a membership function-based method (Dombi, 1990; Delgado et al., 1998; Oussalah, 2002) to extract multi-view information (crisp-valued). The *second part* is a multi-view classifier to handle the extracted multi-view information. In this paper, support vector machines, random forests and neural networks are used as the basic structures of the multi-view classifier. This multi-view classifier guided by the proposed theorem is trained on the view-fusion representation vectors constructed by integrating an appropriate number of candidate views (more details and motivation are discussed in Section 4).

Finally, we compare the performance of the Mv-IIE framework with several baselines on both synthetic and real-world datasets. The experiment results illustrate the superiority of the proposed model in handling interval-valued data. Moreover, we detail an application of the Mv-IIF framework that we present a novel framework for protecting data privacy called *interval privacy-preserving* (INPP), see Section 5.4. Through experiments on one real-world dataset, it demonstrates that applying INPP can prevent raw (crisp-valued) data leakage while ensuring high performance.

## 2 Problem Setting

In this section, we introduce the problem of learning from interval-valued data.

Let $\bar{\mathbf{x}} = (\bar{x}_1, \cdots, \bar{x}_p)^\top$ be a $p$-dimension interval-valued vector, where $\bar{x}_j = [x_j^{\mathrm{l}}, x_j^{\mathrm{r}}], j \in [p]$. Here, we denote $[p] = \{1, \cdots, p\}$. $\bar{\mathbb{R}}$ is denoted as the set of all real-valued intervals (closed) and $\bar{\mathbb{R}}^p$ is denoted as the set of all $p$-dimension interval-valued vector, i.e., $\bar{\mathbb{R}} = \{[x^{\mathrm{l}}, x^{\mathrm{r}}] : x^{\mathrm{l}}, x^{\mathrm{r}} \in \mathbb{R}, x^{\mathrm{l}} \leq x^{\mathrm{r}}\}$ and $\bar{\mathbb{R}}^p = \{([x_1^{\mathrm{l}}, x_1^{\mathrm{r}}], \cdots, [x_p^{\mathrm{l}}, x_p^{\mathrm{r}}])^\top : x_j^{\mathrm{l}}, x_j^{\mathrm{r}} \in \mathbb{R}, x_j^{\mathrm{l}} \leq x_j^{\mathrm{r}}, j \in [p]\}$.

**Key Definitions.** In this part, we introduce some basic definitions to identify the LIND problem. We first show the definition of the interval-valued random variable.

**Definition 1** (Interval-valued Random Variable)**.** *Suppose $X^{\mathrm{l}}, X^{\mathrm{r}} \in \mathbb{R}$ are two real-valued random variables (Jeffreys, 1998) defined in $\mathbb{R}$. We define $\bar{X} = [X^{\mathrm{l}}, X^{\mathrm{r}}] \in \bar{\mathbb{R}}$ as an interval-valued random variable, as long as $X^{\mathrm{l}} \leq X^{\mathrm{r}}$. Then, a $p$-dimension interval-valued random vector $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top \in \bar{\mathbb{R}}^p$ is a $k$-tuple of the interval-valued random variables, where $\bar{X}_j$ ($j \in [p]$) is an interval-valued random variable.*

The interval-valued random variable is a natural extension of the ordinary real-valued random variable. Then, we define $\bar{\mathcal{D}}$ as the interval probability distribution of $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ (denoted as $\bar{\mathbf{X}} \sim \bar{\mathcal{D}}$). Note that the strict definitions related to interval probability distribution and i.i.d. interval-valued random vectors are given in Appendix B. Next, we introduce the definition of the interval expectation for an interval-valued random vector.

**Definition 2** (Interval Expectation)**.** *Suppose $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top \sim \bar{\mathcal{D}}$ is an interval-valued random vector. We denote $\mathbf{X}^{\mathrm{l}} = (X_1^{\mathrm{l}}, \cdots, X_p^{\mathrm{l}})^\top$ and $\mathbf{X}^{\mathrm{r}} = (X_1^{\mathrm{r}}, \cdots, X_p^{\mathrm{r}})^\top$, which are two real-valued random vectors following probability distribution $\mathcal{D}^{\mathrm{l}}$ and $\mathcal{D}^{\mathrm{r}}$. Then, the interval expectation of an interval-valued random vector $\bar{\mathbf{X}}$ is defined as,*

$$\mathbb{E}_{\bar{\mathcal{D}}}[\bar{\mathbf{X}}] = \frac{1}{2}\int \mathbf{x} d\mathcal{D}^{\mathrm{l}}(\mathbf{x}) + \frac{1}{2}\int \mathbf{x} d\mathcal{D}^{\mathrm{r}}(\mathbf{x}) = \frac{1}{2}\mathbb{E}[\mathbf{X}^{\mathrm{l}}] + \frac{1}{2}\mathbb{E}[\mathbf{X}^{\mathrm{r}}].$$

Based on the above definitions and the introduction of ordinary classification problems with crisp-valued observations (Mohri et al., 2012), we can identify the LIND problem.

**Learning from Interval-valued Data:** Let $\bar{\mathcal{X}} \subset \bar{\mathbb{R}}^p$ be the input space and $\mathcal{Y} = [K]$ be the output space. Suppose $\bar{S} = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^m$ is a sample drawn i.i.d. from $\bar{\mathcal{D}}$, where $\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \cdots, \bar{x}_{ip})^\top \in \bar{\mathcal{X}}$ and $y_i = f(\bar{\mathbf{x}}_i) \in \mathcal{Y}$ be the ground-truth function. Let $\mathcal{H} \subset \{\boldsymbol{h} : \bar{\mathcal{X}} \to \mathbb{R}^K\}$ be the hypothesis space of the LIND problem and for any $\boldsymbol{h} \in \mathcal{H}$,

$$\begin{aligned} \boldsymbol{h}(\bar{\mathbf{x}}_i): \quad & \bar{\mathcal{X}} \to \mathbb{R}^K \\ & \bar{\mathbf{x}}_i \to (h_1(\bar{\mathbf{x}}_i), \cdots, h_K(\bar{\mathbf{x}}_i))^\top. \end{aligned}$$

Without loss of generality, we suppose that $\sum_{k=1}^K h_k(\bar{\mathbf{x}}_i) = 1$ and each $h_k(\bar{\mathbf{x}}_i)$ represents the probability of instance $\bar{\mathbf{x}}_i$ belonging to the $k$-th category. Therefore, we have $\sup_{\boldsymbol{h}\in\mathcal{H}} \| \boldsymbol{h} \|_\infty \leq 1$. Let $\mathcal{L}_{\mathcal{H}} = \{\ell(\boldsymbol{h}(\bar{\mathbf{x}}), y) : \bar{\mathbf{x}} \in \bar{\mathcal{X}}, \boldsymbol{h} \in \mathcal{H}, y \in \mathcal{Y}\}$ be the class of functions with respect to the loss $\ell$ and $\mathcal{H}$, where $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$. Based on the ordinary classification problem, we denote $R_{\bar{\mathcal{D}}}(\boldsymbol{h}) = \mathbb{E}_{\bar{\mathcal{D}}}[\ell(\boldsymbol{h}(\bar{\mathbf{x}}), y)]$ as the risk of the LIND problem. Therefore, the aim of the LIND problem is to find the optimal classifier $\boldsymbol{h}^* \in \mathcal{H}$ such that $\boldsymbol{h}^* = \arg\min_{\boldsymbol{h}\in\mathcal{H}} R_{\bar{\mathcal{D}}}(\boldsymbol{h})$.

**Remark 1:** Most previous works considering that the expectation of an interval is itself an interval (Aumann, 1965) were primarily focused on the operation of interval-valued data. However, in this paper, we focus on learning this type of data (interval) from a machine learning perspective. Therefore, we give a different definition of the expectation of an interval (Definition 2).

## 3 Theoretical Analysis

This section presents the main theoretical outcome of the LIND problem (all proofs and further analysis are shown in Appendix B).

Let $\bar{S}_{\bar{X}} = \{\bar{\mathbf{x}}_i\}_{i=1}^m$ be a sample drawn i.i.d. from $\bar{\mathcal{D}}$. Based on the Rademacher complexity of $\mathcal{H}$ with respect to $\bar{S}_{\bar{X}}$ (see Definition 7 in Appendix B), we can obtain the following theorem.

**Theorem 1.** *Suppose that $\sup_{\|\boldsymbol{h}\|_\infty \leq 1} \max_y \ell(\boldsymbol{h}, y) \leq C_\ell$, and all functions in $\mathcal{L}_{\mathcal{H}}$ are $L_\ell$-Lipschitz functions. For any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $\boldsymbol{h} \in \mathcal{H}$:*

$$|R_{\bar{\mathcal{D}}}(\boldsymbol{h}) - \widehat{R}_{\bar{\mathcal{D}}}(\boldsymbol{h})| \leq 2\sqrt{2}L_\ell \widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}) + 3C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}}. \tag{1}$$

This theorem presents a generalization bound of the discrepancy between the risk and empirical risk of $\boldsymbol{h}$ based on empirical Rademacher complexity. $\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H})$ is in the order of $O(1/\sqrt{m})$ under some certain restrictions of $\mathcal{H}$ (Bach et al., 2004; Cortes et al., 2010; Kloft et al., 2011), for example $\mathcal{H}$ has limited-VC dimension or $\mathcal{H}$ is a kernel class with bounded trace. According to Eq. (1) and if $\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}) = O(1/\sqrt{m})$, we notice that as $m \to \infty$, $R_{\bar{\mathcal{D}}}(\boldsymbol{h}) \to \widehat{R}_{\bar{\mathcal{D}}}(\boldsymbol{h})$. Therefore, this bound demonstrates that we can always well handle the LIND problem when enough interval-valued instances can be collected.

In addition, we prove two theorems (See Appendix B.2 for details) to illustrate the advantage of using multi-view learning to address the LIND problem in terms of error rate and estimation error bound. Theorem 4 shows that the error rate of a multi-view prediction function is lower than

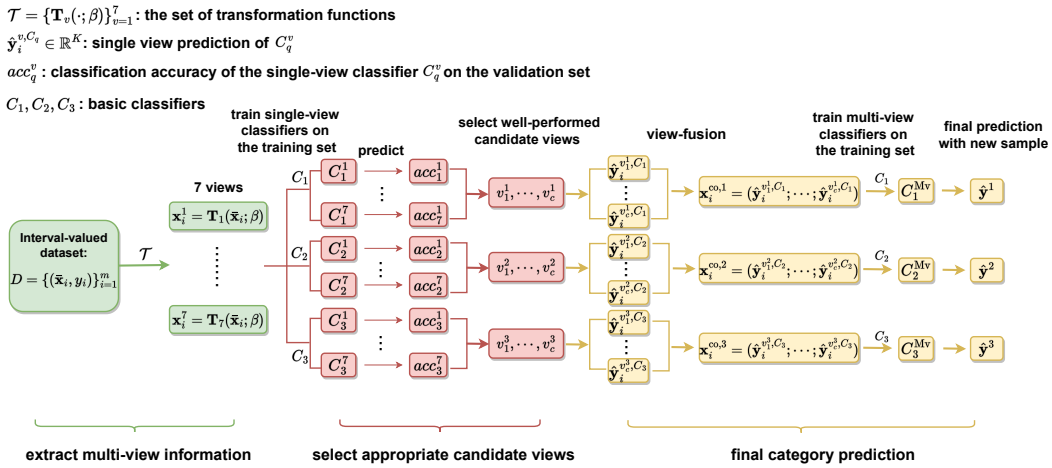



Figure 1: **Mv-IIE** framework. The first part (denoted in **green**) is to extract the multi-view information from the interval-valued dataset $D$. Then, the multi-view classifier with two structures is used to handle the extracted multi-view information. The first structure (denoted in **red**) is used to select well-performed candidate views. The second structure (denoted in **yellow**) aims to train the final multi-view classifiers by using the view-fusion representation vectors.

that of any single-view prediction function under some certain restrictions, which means that using multi-view methodology can reduce the error rate of the predict function for the classification tasks. Theorem 5 demonstrates that we can obtain tighter estimation error bounds by applying the multi-view methodology. Inspired by the theoretical analysis of Theorems 4 and 5, we decide to find appropriate multi-view features that can achieve well and similar performance on some specific classifiers to train our multi-view classifiers.

## 4 Construct Model for Interval-valued Data Classification

In this section, a new framework called *multi-view interval information extraction* (Mv-IIE) approach is presented to address the LIND problem. The structure of the Mv-IIE framework is shown in Figure 1. We describe this proposed framework in detail in the following paragraph.

We denote $D = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^m$ as the interval-valued dataset, where $\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \cdots, \bar{x}_{ip})^\top \in \bar{\mathbb{R}}^p$ is the interval-valued feature vector and $y_i \in [K]$ is the label. First, we construct a set of membership function-based transformation functions to extract multi-view information (crisp-valued) from the interval-valued dataset $D$ (denoted in green in Figure 1). $\mathcal{T} = \{\mathbf{T}_v(\cdot;\beta)\}_{v=1}^7$ is denoted as the set of transformation functions, where

$$\mathbf{T}_1 = \mathrm{MOM} \circ \mathbf{F}_1, \mathbf{T}_2 = \mathrm{COG} \circ \mathbf{F}_1, \mathbf{T}_3 = \mathrm{COG} \circ \mathbf{F}_2, \mathbf{T}_4 = \mathrm{ALC} \circ \mathbf{F}_1,$$
$$\mathbf{T}_5 = \mathrm{ALC} \circ \mathbf{F}_2, \mathbf{T}_6 = \mathrm{VAL} \circ \mathbf{F}_1, \mathbf{T}_7 = \mathrm{VAL} \circ \mathbf{F}_2.$$

Here, $\mathbf{F}_1(\cdot;\beta), \mathbf{F}_2(\cdot;\beta)$ are two functions that are used to transfer a interval-valued feature vector into a triangular fuzzy vector and a Gaussian fuzzy vector, and MOM, COG, ALC, VAL (Delgado et al., 1998; Oussalah, 2002) are four different membership function-based defuzzification methods (see Appendix C for details). $\mathbf{F}_1(\cdot;\beta), \mathbf{F}_2(\cdot;\beta)$ are defined as:

$$\mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta) = (F_\tau(\bar{x}_{i1};\beta), \cdots, F_\tau(\bar{x}_{ip};\beta))^\top, \tau = 1, 2.$$
$$F_1(\bar{x}_{ij};\beta) = \mathrm{Tr}(x_{ij}^{\mathrm{l}}, \beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, x_{ij}^{\mathrm{r}}),$$
$$F_2(\bar{x}_{ij};\beta) = \mathrm{Ga}(\beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, S_{1j}, S_{2j}),$$
$$S_{1j} = \sqrt{\mathrm{Var}(A_j)}, S_{2j} = \sqrt{\mathrm{Var}(B_j)},$$
$$A_j = \{x_{ij}^{\mathrm{l}} : i \in [m], (\bar{\mathbf{x}}_i, y_i) \in D\}, B_j = \{x_{ij}^{\mathrm{r}} : i \in [m], (\bar{\mathbf{x}}_i, y_i) \in D\}, j \in [p],$$

where $\mathrm{Tr}(x_{ij}^{\mathrm{l}}, \beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, x_{ij}^{\mathrm{r}})$ and $\mathrm{Ga}(\beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, S_{1j}, S_{2j})$ are represented two types of fuzzy numbers (see Appendix C for details). Through the above process, one interval-valued feature $\bar{\mathbf{x}}_i$ can be transferred into seven different parts $\mathbf{X}_i^{\mathrm{Mv}} = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^7)$, where for any $i \in [m], v \in [7], \mathbf{x}_i^v = \mathbf{T}_v(\bar{\mathbf{x}}_i;\beta), \mathbf{T}_v \in \mathcal{T}$. Then, we obtain the multi-view information $D_{\mathrm{Mv}} = \{(\mathbf{x}_i^1, y_i, 1), \cdots, (\mathbf{x}_i^7, y_i, 7)\}_{i=1}^m$ by using the above mentioned method. For any $(\mathbf{x}_i^v, y_i, v) \in D_{\mathrm{Mv}}, y_i \in [K]$ is the category label, and $v \in [7]$ is the view label.

**Algorithm 1** Mv-IIE

**1: Input** data $D = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{m}$, the basic classifiers $\mathrm{C}_q, q \in [3]$;
**2: Initial** network parameters of $\mathrm{C}_3$ and split $D$ into a training set $D^{\mathrm{tr}}$ with size $m_1$, a validation set $D^{\mathrm{va}}$ with size $m_2$ and a test set $D^{\mathrm{te}}$ with size $m_3$;
**3: Compute** extract multi-view information : $\mathbf{x}_i^v = \mathbf{T}_v(\bar{\mathbf{x}}_i; \beta), \mathbf{T}_v \in \mathcal{T}, i \in [m], v \in [7]$;
**4: Train** single-view classifiers $\mathrm{C}_q^v, v \in [7], q \in [3]$ on the training set $\{(\mathbf{x}_i^v, y_i)|(\bar{\mathbf{x}}_i, y_i) \in D^{\mathrm{tr}}\}_{i=1}^{m_1}$;
**5: Compute** classification accuracy of the single-view classifier $\mathrm{C}_q^v, v \in [7], q \in [3]$ on the validation set $\{(\mathbf{x}_i^v, y_i)|(\bar{\mathbf{x}}_i, y_i) \in D^{\mathrm{va}}\}_{i=1}^{m_2}$;
**6: Select** $c$ candidate views for each $q \in [3]$, denoted as $\mathcal{V}^q = \{v_1^q, \cdots, v_c^q\}$, that achieve higher classification accuracy than the rest of the views;
**7: Compute** view-fusion representation vector :

$$\mathbf{x}_i^{\mathrm{co},q} = (\hat{\mathbf{y}}_i^{v_1^q,\mathrm{C}_q}; \cdots; \hat{\mathbf{y}}_i^{v_c^q,\mathrm{C}_q}), i \in [m], q \in [3],$$

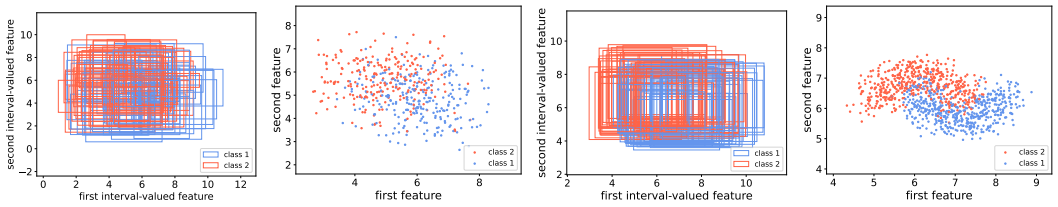
where $\hat{\mathbf{y}}_i^{v,\mathrm{C}_q} \in \mathbb{R}^K$ is the category prediction for the $v$-th view of the $i$-th data by applying $\mathrm{C}_q$;
**8: Train** multi-view classifiers $\mathrm{C}_q^{\mathrm{Mv}}, q \in [3]$ on the training set $\{(\mathbf{x}_i^{\mathrm{co},q}, y_i)|(\bar{\mathbf{x}}_i, y_i) \in D^{\mathrm{tr}}\}_{i=1}^{m_1}$;
**9: Select** the optimal hyperparameters that can obtain the highest classification accuracy on the validation set $\{(\mathbf{x}_i^{\mathrm{co},q}, y_i)|(\bar{\mathbf{x}}_i, y_i) \in D^{\mathrm{va}}\}_{i=1}^{m_2}$;
**10: Output** $\mathrm{C}_q^{\mathrm{Mv}}, q \in [3]$ with optimal hyperparameters and use these model to test the performance on the test set $\{(\mathbf{x}_i^{\mathrm{co},q}, y_i)|(\bar{\mathbf{x}}_i, y_i) \in D^{\mathrm{te}}\}_{i=1}^{m_3}$.

**Motivation of transformation functions construction:** The interval-valued features contain similar structures and properties with fuzzy numbers Delgado et al. (1998), which both exist a considerable amount of uncertainty. Further, the $\alpha$-cut of a fuzzy number $\widetilde{x}$ is defined as $\{t \in \mathbb{R}|\mu_{\widetilde{x}}(t) \leq \alpha\}$ ($\mu_{\widetilde{x}}(t)$ is the membership function of $\widetilde{x}$), which is a closed and bounded interval. Therefore, we design two fuzzilization methods to transfer the interval-valued features into two well-defined fuzzy numbers. Moreover, the four membership function-based methods can extract different crucial discriminant information from fuzzy numbers. For example, MOM finds the maximum membership level but ignores the changing trend of the membership function, while COG takes into account the trend and finds the centroid of the area bounded by the membership function. Through the above analysis, it inspired us to construct a set of transformation functions by fusing the two fuzzilization methods and the four membership function-based methods to extract multi-view discriminant information. Experimental results shown in Sections 5.2 and 5.3 verify the rationality and efficacy of the fuzzy transformation functions.

Next, we propose a multi-view classifier with two parts to train the multi-view information, which aims to minimize the empirical risk $\widehat{R}_{\mathcal{D}}(\boldsymbol{h}_{\mathrm{co}})$ in Section 3. The first part (denoted in red in Figure 1) is used to select appropriate multi-view information. We apply support vector machines, random forests and neural networks as three basic classifiers, which denoted as $\mathrm{C}_1, \mathrm{C}_2$ and $\mathrm{C}_3$. Then, we apply the three basic classifiers to train single-view classifiers $\mathrm{C}_q^v, v \in [7], q \in [3]$ on the training set, and we select several well-performed views with the number of $c$ as the candidate views for each basic classifier on the validation set (we set $c = 2$ for our experiments in this paper). This selected approach is inspired by the theoretical analysis of Theorem 4 and 5. Let $\hat{\mathbf{y}}_i^{v,\mathrm{C}_q} \in \mathbb{R}^K, i \in [m], v \in [7], q \in [3]$ denoted as the category prediction for the $v$-th view of the $i$-th data by applying the basic classifier $\mathrm{C}_q$, and $\mathcal{V}^q = \{v_1^q, \cdots, v_c^q\}, q \in [3]$ denoted as the selected candidate views for basic classifier $\mathrm{C}_q$. The second part (denoted in yellow in Figure 1) aims to train the final multi-view classifiers by using the selected candidate views. For each basic classifier $\mathrm{C}_q, q \in [3]$, the category predictions of the selected candidate views are integrated to obtain $\mathbf{x}_i^{\mathrm{co},q} = (\hat{\mathbf{y}}_i^{v_1^q,\mathrm{C}_q}; \cdots; \hat{\mathbf{y}}_i^{v_c^q,\mathrm{C}_q}), i \in [m], q \in [3]$ as view-fusion representation vector, and we use $\mathbf{x}_i^{\mathrm{co},q}$ as input and $\mathrm{C}_q$ as a classifier to train the multi-view classifier $\mathrm{C}_q^{\mathrm{Mv}}$ on the training set and select the optimal hyperparameters of $\mathrm{C}_q^{\mathrm{Mv}}$ on the validation set. Finally, the trained multi-view classifiers $\mathrm{C}_q^{\mathrm{Mv}}, q \in [3]$ with optimal hyperparameters are used to get the final category prediction $\hat{\mathbf{y}}^q$ on the test set. More detail of the Mv-IIE framework is shown in **Algorithm** 1.

## 5 Experiments

In this section, we compare the proposed model with several baselines on both synthetic and real-world datasets, and introduce an application of our method.




(a) $1^{\text{st}}$ synthetic dataset (b) $1^{\text{st}}$ synthetic dataset (c) $2^{\text{nd}}$ synthetic dataset (d) $2^{\text{nd}}$ synthetic dataset

Figure 2: Synthetic datasets. From (a) and (c), each rectangle represents one interval-valued instance. (b) and (d) plot the the center of the interval-valued data (rectangle) to show the separability of the synthetic dataset.

### 5.1 BASELINES

This section gives a brief introduction of seven baselines. The first two baselines called **DF-SVM** and **DF-MLP** are proposed in (Ma et al., 2022). Next three baselines called **L-IIE**, **U-IIE** and **M-IIE** that take the low bound, upper bound and midpoint values from intervals to train the three basic classifiers. The other two baselines called **Mv-IIE-2** and **Mv-IIE-3** are constructed based on our proposed framework in this paper. Instead of using the membership function-based method to extract multi-view information in the proposed framework, **Mv-IIE-2** uses the upper and lower bounds of intervals as two views. **Mv-IIE-3** uses the two views mentioned above and the midpoint of intervals as another view and integrates all these views to get the final prediction.

### 5.2 EXPERIMENTS ON SYNTHETIC DATASETS

In this section, we verify the efficacy of the proposed framework on three synthetic datasets. First, we introduce the process of the synthetic datasets generation.

**Interval-valued Dataset Generation.** We use two different mechanisms to construct synthetic interval-valued datasets. In the first data-generation mechanism, we generate the crisp-valued dataset $\{(\mathbf{x}_i = (x_{i1}, x_{i2})^\top, y_i)\}_{i=1}^n$ in two categories by the double moon data generator. Then, we use the generated crisp-valued dataset to construct the first interval-valued dataset $\{\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \bar{x}_{i2})^\top, y_i\}_{i=1}^n$, where each $\bar{x}_{ij}$ is an interval characterized by $[x_{ij} - a_{ij}, x_{ij} + b_{ij}]$. We let $a_{ij}, b_{ij} \sim U[0, 4]$ and $n = 1000$ to generate the first synthetic dataset and let $a_{ij} \sim U[0.5, 1], b_{ij} \sim U[2, 4]$ and $n = 2000$ to generate the second synthetic dataset ($U[a, b]$ denotes the uniform distribution over $[a, b]$). Visualizations of the first two synthetic datasets are shown in Figure 2.

In the second data-generation mechanism, we first select one dataset (Letter Recognition dataset selected from the UCI Machine Learning Repository https://archive-beta.ics.uci.edu/) denoted as $D_{\mathrm{R}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^\top \in \mathbb{R}^p$, and $y_i \in [K]$. Then, we present one intervalization approach to generate the second synthetic interval-valued dataset (see Figure 3). We select the first $L$ features in $D_{\mathrm{R}}$ and find the maximum value $x_l^{\max}$ and minimum value $x_l^{\min}$ of each feature $l$, so for any $l \in [L], i \in [n], x_{ip} \in [x_l^{\min}, x_l^{\max}]$. We bisect the interval $[x_l^{\min}, x_l^{\max}]$ into $T$ intervals $[x_l^0, x_l^1], [x_l^2, x_l^3], \cdots, [x_l^{T-1}, x_l^T]$. We denote for any $l \in [L],\ t \in [L]$, and $k \in [K]$,

$$I_{lk}^t = \{(\mathbf{x}_i, y_i) \in D_R : x_{il} \in [x_l^{t-1}, x_l^t], y_i = k\}.$$

Finally, we transfer set $I_{lk}^t$ into an interval-valued data $(([x_1^{\mathrm{l}}, x_1^{\mathrm{r}}], \cdots, [x_p^{\mathrm{l}}, x_p^{\mathrm{r}}])^\top, k)$, where $x_j^{\mathrm{l}} = \min_{(\mathbf{x}_i, k) \in I_{lk}^t} x_{ij},\ x_j^{\mathrm{r}} = \max_{(\mathbf{x}_i, k) \in I_{lk}^t} x_{ij}, j \in [p]$. Then, let $L = 4, T = 12$, we generate the third synthetic interval-valued dataset by using the aforementioned data-generation mechanism.

**Experiment Results Analysis.** In our experiments, we compare the performance of the Mv-IIE framework with the seven baselines on the three generated synthetic datasets. All the experiment details are shown in Appendix D. The experimental results are shown in Table 2. From these results, it can be seen that the proposed model achieves the best classification accuracy on the three synthetic datasets. Further, results of the Wilcoxon rank-sum test Wilcoxon (1992) show that our approach outperforms **DF-MLP**, **L-IIE**, **U-IIE**, **M-IIE** and **Mv-IIE-2** significantly at the $0.05$ significance level in most cases. Further, our method outperforms **Mv-IIE-2** and **Mv-IIE-3**, which verifies the rationality of the theoretical analysis of Theorems 4 and 5 (see Section 3). All these results verify the superiority of the proposed model in addressing classification problems with interval-valued data.

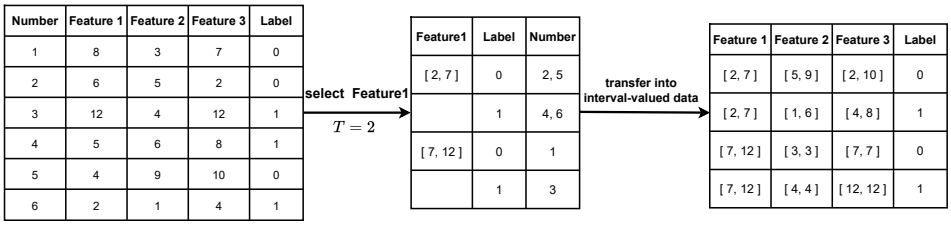



Figure 3: The intervalization approach.

Table 2: Experiment results (accuracy±standard deviation of accuracies) on the three synthetic datasets. The bold value represents the highest accuracy in each column. $p$ is the $p$-value of the Wilcoxon rank-sum test between the best performance and other algorithms. $^*$ represents $p < 0.05$, meaning that Mv-IIE outperforms other baselines significantly at the $0.05$ significance level (Rice, 2006).

| | | $1^{\text{st}}$ synthetic | | $2^{\text{nd}}$ synthetic | | $3^{\text{rd}}$ synthetic | |
|---|---|---|---|---|---|---|---|
| Algorithms | basic classifier | Test accuracy | $p$ | Test accuracy | $p$ | Test accuracy | $p$ |
| DF-SVM | | 71.30% ±1.62% | 0.73 | 97.82% ±0.61% | 0.43 | 94.26% ±2.10% | 0.59 |
| DF-MLP | | 70.60% ±1.67% | 0.35 | 97.13% ±1.04% | $0.034^*$ | 92.21% ±2.15% | $0.032^*$ |
| L-IIE | $C_1$ | 67.35% ±2.95% | $0.0051^*$ | 97.90% ±0.76% | 0.47 | 90.34% ±1.64% | $0.0012^*$ |
| | $C_2$ | 63.90% ±2.71% | $0.00016^*$ | 97.80% ±0.64% | 0.43 | 89.17% ±1.32% | $0.00067^*$ |
| | $C_3$ | 66.15% ±2.40% | $0.00016^*$ | 97.10% ±0.84% | $0.034^*$ | 90.49% ±1.88% | $0.0012^*$ |
| U-IIE | $C_1$ | 66.90% ±1.37% | $0.00016^*$ | 76.95% ±1.43% | $0.0016^*$ | 88.04% ±2.59% | $0.048^*$ |
| | $C_2$ | 65.20% ±2.61% | $0.00016^*$ | 75.68% ±1.31% | $0.0016^*$ | 89.31% ±2.94% | $0.00067^*$ |
| | $C_3$ | 66.65% ±1.21% | $0.00016^*$ | 75.95% ±1.63% | $0.0016^*$ | 87.01% ±2.90% | $0.00034^*$ |
| M-IIE | $C_1$ | 71.15% ±2.09% | 0.60 | 89.85% ±0.92% | $0.0016^*$ | 94.02% ±2.07% | $0.048^*$ |
| | $C_2$ | 70.25% ±2.17% | 0.11 | 88.58% ±0.72% | $0.0016^*$ | 90.54% ±1.92% | $0.0012^*$ |
| | $C_3$ | 69.95% ±2.11% | $0.048^*$ | 86.95% ±1.44% | $0.0016^*$ | 91.67% ±1.81% | $0.0094^*$ |
| Mv-IIE-2 | $C_1$ | 70.35% ±2.09% | 0.15 | 98.20% ±0.81% | 0.94 | 93.14% ±1.95% | $0.048^*$ |
| | $C_2$ | 69.85% ±2.92% | $0.045^*$ | 97.03% ±0.96% | $0.0011^*$ | 90.44% ±2.60% | $0.0011^*$ |
| | $C_3$ | 65.82% ±3.12% | $0.00016^*$ | 97.90% ±0.72% | 0.47 | 84.17% ±3.14% | $0.00034^*$ |
| Mv-IIE-3 | $C_1$ | 71.05% ±2.21% | 0.57 | 98.17% ±0.72% | 0.88 | 94.46% ±2.15% | 0.79 |
| | $C_2$ | 70.90% ±2.88% | 0.44 | 97.25% ±0.84% | $0.044^*$ | 91.37% ±2.70% | $0.0081^*$ |
| | $C_3$ | 66.25% ±1.95% | $0.00016^*$ | 97.85% ±0.65% | 0.43 | 80.59% ±4.43% | $0.00034^*$ |
| Mv-IIE | $C_1$ | 71.25% ±2.11% | 0.68 | **98.25% ± 0.69%** | — | **94.66% ± 1.81%** | — |
| | $C_2$ | **71.65% ± 2.05%** | — | 97.13% ±1.18% | $0.034^*$ | 90.49% ±2.51% | $0.0012^*$ |
| | $C_3$ | 71.05% ±1.67% | 0.57 | 98.05% ±0.72% | 0.74 | 86.96% ±1.35% | $0.00034^*$ |

Table 3: Experiment results (accuracy±standard deviation of accuracies) on the two real-world datasets. The bold value represents the highest accuracy in each column. $p$ is the $p$-value of the Wilcoxon rank-sum test between the best performance and other algorithms. $^*$ represents $p < 0.05$, meaning that Mv-IIE outperforms other baselines significantly at the $0.05$ significance level (Rice, 2006).

| | | Mushroom dataset | | Weather dataset | |
|---|---|---|---|---|---|
| Algorithms | basic classifier | Test accuracy | $p$ | Test accuracy | $p$ |
| DF-SVM | | 76.67% ±3.86% | $0.00067^*$ | 97.12% ±0.98% | 0.79 |
| DF-MLP | | 79.39% ±3.32% | $0.019^*$ | 96.83% ±0.98% | $0.048^*$ |
| L-IIE | $C_1$ | 71.18% ±6.30% | $0.00034^*$ | 93.56% ±0.96% | $0.00016^*$ |
| | $C_2$ | 76.36% ±6.62% | $0.00067^*$ | 93.33% ±0.73% | $0.00016^*$ |
| | $C_3$ | 76.19% ±4.13% | $0.00067^*$ | 93.54% ±0.85% | $0.00016^*$ |
| U-IIE | $C_1$ | 74.50% ±2.88% | $0.00054^*$ | 94.06% ±0.90% | $0.00016^*$ |
| | $C_2$ | 79.14% ±3.58% | $0.015^*$ | 93.15% ±1.16% | $0.00016^*$ |
| | $C_3$ | 76.01% ±4.29% | $0.00067^*$ | 93.93% ±1.03% | $0.00016^*$ |
| M-IIE | $C_1$ | 75.60% ±3.11% | $0.00054^*$ | 97.03% ±0.96% | 0.75 |
| | $C_2$ | 79.34% ±4.75% | $0.019^*$ | 97.08% ±0.73% | 0.77 |
| | $C_3$ | 76.01% ±4.29% | $0.00067^*$ | 96.85% ±0.88% | $0.048^*$ |
| Mv-IIE-2 | $C_1$ | 81.74% ±5.13% | $0.045^*$ | 96.76% ±0.91% | $0.047^*$ |
| | $C_2$ | 80.47% ±3.69% | $0.022^*$ | 95.50% ±1.25% | $0.0032^*$ |
| | $C_3$ | 76.55% ±5.58% | $0.00067^*$ | 93.72% ±1.00% | $0.00016^*$ |
| Mv-IIE-3 | $C_1$ | 82.34% ±3.79% | 0.56 | 97.12% ±0.90% | 0.79 |
| | $C_2$ | 82.57% ±4.54% | 0.69 | 97.10% ±0.74% | 0.77 |
| | $C_3$ | 73.79% ±4.69% | $0.00034^*$ | 93.52% ±1.16% | $0.00016^*$ |
| Mv-IIE | $C_1$ | 81.75% ±4.09% | $0.045^*$ | **97.26% ± 0.81%** | — |
| | $C_2$ | **83.69% ± 3.39%** | — | 96.46% ±0.73% | $0.041^*$ |
| | $C_3$ | 82.67% ±3.14% | 0.72 | 96.62% ±1.20% | $0.046^*$ |

Table 4: Experiment results (accuracy±standard deviation of accuracies) of the ablation study on the synthetic and real-world datasets. The bold value represents the highest accuracy in each column.

| Algorithms | Basic classifier | $1^{st}$ synthetic | $2^{nd}$ synthetic | $3^{rd}$ synthetic | Mushroom | Weather |
|---|---|---|---|---|---|---|
| view 1 | $C_1$ | 70.70% ±2.16% | 97.97% ±0.80% | 94.22% ±2.05% | 76.81% ±3.07% | 96.94% ±0.96% |
| | $C_2$ | 69.75% ±2.32% | 97.85% ±0.92% | 91.27% ±2.31% | 82.29% ±5.26% | 97.03% ±0.68% |
| | $C_3$ | 71.45% ±2.25% | 98.12% ±0.66% | 92.21% ±1.76% | 77.56% ±3.36% | 96.80% ±1.25% |
| view 2 | $C_1$ | 70.95% ±1.62% | 96.50% ±0.56% | 94.26% ±1.99% | 76.66% ±3.83% | 97.12% ±0.74% |
| | $C_2$ | 69.75% ±1.74% | 95.10% ±1.10% | 91.81% ±1.87% | 83.35% ±5.06% | 96.83% ±0.97% |
| | $C_3$ | 70.25% ±1.93% | 96.47% ±0.68% | 92.45% ±1.85% | 79.62% ±4.15% | 96.83% ±0.96% |
| view 3 | $C_1$ | 71.20% ±2.17% | 95.20% ±0.56% | 94.41% ±2.05% | 76.55% ±3.25% | 97.01% ±0.94% |
| | $C_2$ | 69.45% ±2.27% | 94.00% ±0.81% | 91.67% ±2.28% | 82.44% ±4.65% | 96.69% ±0.99% |
| | $C_3$ | 71.20% ±1.68% | 94.30% ±0.91% | 91.52% ±2.62% | 79.62% ±3.32% | 96.78% ±1.12% |
| view 4 | $C_1$ | 71.30% ±1.62% | 97.82% ±0.61% | 94.26% ±2.10% | 76.67% ±3.86% | 97.12% ±0.98% |
| | $C_2$ | 69.15% ±3.24% | 97.13% ±1.04% | 90.88% ±2.98% | 82.45% ±5.26% | 96.72% ±1.20% |
| | $C_3$ | 70.60% ±1.67% | 97.62% ±0.93% | 92.21% ±2.15% | 79.39% ±3.32% | 96.76% ±0.98% |
| view 5 | $C_1$ | 70.60% ±1.73% | 97.97% ±0.80% | 94.17% ±1.87% | 75.07% ±3.18% | 96.96% ±0.89% |
| | $C_2$ | 69.75% ±2.32% | 97.50% ±0.78% | 91.08% ±2.49% | 82.70% ±4.88% | 96.96% ±0.69% |
| | $C_3$ | 71.20% ±2.11% | 98.12% ±0.66% | 90.49% ±2.19% | 71.38% ±5.94% | 96.42% ±1.03% |
| view 6 | $C_1$ | 70.65% ±2.13% | 98.00% ±0.80% | 94.17% ±2.13% | 77.12% ±3.14% | 97.01% ±0.87% |
| | $C_2$ | 69.75% ±2.32% | 98.05% ±0.76% | 90.54% ±2.11% | 82.78% ±5.08% | 96.94% ±0.70% |
| | $C_3$ | 70.35% ±2.47% | 98.12% ±0.66% | 92.55% ±1.65% | 77.90% ±4.48% | 96.58% ±1.05% |
| view 7 | $C_1$ | 70.60% ±1.74% | 97.95% ±0.77% | 94.36% ±2.02% | 76.81% ±3.07% | 97.05% ±0.75% |
| | $C_2$ | 69.75% ±2.32% | 97.38% ±1.03% | 90.54% ±2.07% | 82.89% ±5.13% | 97.03% ±0.68% |
| | $C_3$ | 70.60% ±1.67% | 98.12% ±0.66% | 91.57% ±2.33% | 75.13% ±4.82% | 96.96% ±1.04% |
| Mv-IIE | $C_1$ | 71.25% ±2.11% | **98.25% ± 0.69%** | **94.66% ± 1.81%** | 81.75% ±4.09% | **97.26% ± 0.81%** |
| | $C_2$ | **71.65% ± 2.05%** | 97.13% ±1.18% | 90.49% ±2.51% | **83.69% ± 3.39%** | 96.46% ±0.73% |
| | $C_3$ | 71.05% ±1.67% | 98.05% ±0.72% | 86.96% ±1.35% | 82.67% ±3.14% | 96.62% ±1.20% |

### 5.3 Experiments on Real-world Datasets

This section illustrates the experimental results on two real-world datasets which are used to verify the efficacy of the proposed framework. The briefly introduction of the two real-world datasets used in our experiments is shown in Appendix E.

**Experiment Results Analysis.** All the experiment details on the two real-world datasets are shown in Appendix D. The experiment results on the two real-world datasets are shown in Table 3. From the results of classification accuracy and the Wilcoxon rank-sum test, it can be seen that the proposed model outperforms **DF-SVM**, **DF-MLP**, **L-IIE**, **U-IIE**, **M-IIE** and **Mv-IIE-2** significantly at the 0.05 significance level nearly in all cases. **DF-SVM** and **DF-MLP** perform much worse than our methods on the mushroom dataset because they ignore some crucial discriminant information from this dataset. In comparison, our methods via multi-view learning and fuzzy transformation functions can extract more discriminant information. In addition, although **Mv-IIE-3** applies 3 views, our method still get better outcomes than **Mv-IIE-3**. These results again demonstrate the superiority of our method in addressing classification problems with interval-valued data.

**Ablation Study.** To verify the advantage of using multi-view methodology, we apply all single-view classifiers ($C_q^v, v \in [7], q \in [3]$, see Section 4) to test classification performance on both synthetic and real-world datasets. All results are report in Table 4, which verifies the proposed framework's superiority and rationality in addressing interval-valued data classification problems.

### 5.4 Application

In this section, we describe an application of Mv-IIE, where a novel framework for protecting data privacy called *interval privacy-preserving* (INPP) is presented. The structure of the INPP framework is shown in Figure 4 (see Appendix D). There are three roles involved in each machine learning task: the input party (data owners), the computation party and the results' party. In such systems, the data owner(s) send their data to the computation party. Then, the computation party trains a model using these data and sends this model to the results' party. Finally, the results' party uses this model to predict new data. If all three roles are from the same entity, then privacy is naturally preserved. However, when these roles are from two or more entities, privacy-preserving is necessary. For example, an online clothing retailer wants to know different customers' preferences to adjust the quantity of each garment. In this situation, different customers play the first role and online clothing retailers play the second and third roles.

Table 5: Experiment results of INPP framework on letter recognition dataset. $R$ is equal to the ratio of the outcomes of INPP framework to the best outcome on the original dataset.

| Method | $L$ | $T$ | $q$ | Test accuracy | R | EN |
|---|---|---|---|---|---|---|
| original dataset | — | — | — | 95.86% ±0.19% | — | — |
| INPP | 6 | 15 | 0.20 | 88.85% ±0.71% | 92.69% | 66.82% |
| | 6 | 15 | 0.30 | 91.19% ±0.75% | 95.13% | 56.84% |
| | 6 | 15 | 0.50 | 93.24% ±0.50% | 97.27% | 37.19% |

In the proposed framework, we denote $D_{\mathrm{R}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ as the raw data from the data owner(s), where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^\top \in \mathbb{R}^p, y_i \in [K]$. First, the data owner(s) use the intervalization approach (see Figure 3) to transfer $D_{\mathrm{R}}$ into the interval-valued data $D_{\mathrm{EN}} = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{m}$ and an interval method the same as the first data-generation mechanism described in Section 5.2 to transfer $D_{\mathrm{C}}$, which contains $n*q$ instances randomly selected from $D_{\mathrm{R}}$, into the interval-valued data $D_{\mathrm{IN}} = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^{n*q}$. Then, the data owner(s) send these two interval-valued datasets $D_{\mathrm{EN}}, D_{\mathrm{IN}}$ to the computation party. Secondly, the computation party uses the interval-valued data $D_{\mathrm{EN}}$ to train Model 1 by applying Mv-IIE and the interval-valued data $D_{\mathrm{IN}}$ is used to fine-tune Model 1 to obtain Model 2. Then, the computation party sends Model 2 to the results' party. Finally, the results' party uses the same interval method described in Section 5.2 to transfer $\mathbf{x}_{\mathrm{new}}$ into $\bar{\mathbf{x}}_{\mathrm{new}}$ and uses Model 2 to predict $\bar{\mathbf{x}}_{\mathrm{new}}$ for new data prediction. According to the above methods, the intervalization process of our proposed framework is irreversible and the raw data is largely compressed. Therefore, the computation party and other parties cannot obtain the raw data from $D_{\mathrm{EN}}$ and $D_{\mathrm{IN}}$, so this process achieve the purpose of preventing data leakage. We define $\mathrm{EN} = 1 - (m + n * q)/n$, where $(m + n * q)$ is the amount of data that the computation party can receive from the data owner(s) and $n$ is the amount of raw data. A smaller EN means the computation party will receive more data from the data owner(s), so the computation party may receive more information about the raw data. Therefore, EN can be used to measure the degree of privacy-preserving by applying INPP. Greater EN means greater privacy protection by applying INPP.

Differential privacy (DP) Dwork et al. (2014); Papernot et al. (2018) and homomorphic encryption (Gilad-Bachrach et al., 2016; Zhang et al., 2018b; Lou & Jiang, 2021) are common used schemes to achieve privacy-preserving. DP and homomorphic encryption can be applied to the raw data or the algorithm, but our method only applies to the raw data. DP applied to the raw data is based on data-perturbation, and homomorphic encryption is based on data-encryption, but the amount of data is not changed. Moreover, if the keys of the encryption schemes are compromised, the information of the raw data will also be compromised. While our method compresses the raw data into interval-valued data with fewer instances through an irreversible process to protect data privacy. Further, DP and our approach can not be easily applied to image data, which is a meaningful problem worth considering in the future.

Experiments on one real-world dataset are conducted to verify the efficacy and feasibility of the INPP framework. We use four well-known machine learning methods (logistic regression, support vector machines, random forests and neural networks) to classify the original dataset and compare the best outcome of these four methods on the original dataset with the outcomes of the INPP framework. The experiment details of the INPP are shown in Appendix D. All the experiment results are shown in Table 5. We note that the proposed framework can achieve $93.24\%$ classification accuracy on the new data with $\mathrm{R} = 97.27\%$ when $L = 6, T = 15, q = 0.5$, which demonstrates that applying the proposed framework can prevent crisp-valued data leakage while ensuring high classification accuracy of the model that has been trained by the computation party.

## 6 Conclusion, Limitations and Future Work

In this paper, we focus on a highly challenging called LIND and obtain the estimation error bound of this problem based on Rademacher complexity. Moreover, we construct a new framework called Mv-IIE by applying multi-view learning for interval-valued data classification. Through experimental comparisons with seven baselines on both synthetic and real-world datasets, it demonstrates the superiority of the proposed model. Finally, we detail an application of the proposed framework that we can prevent crisp-valued data leakage by transforming crisp-valued data into interval-valued data. However, we only consider the situation where the observations with interval-valued features in the training and test sets are drawn from the same distribution in this paper. Therefore, we plan to consider more complicated issues related to interval-valued data analysis in future research, for example, covariate shift and domain adaptation with interval-valued observations.

## REPRODUCIBILITY STATEMENT

All the codes and processed data in this paper will be publicly released on our GitHub website after this paper is accepted. Moreover, we include all complete proofs proofs for our theoretical results in Appendix B, and additional experiment details in Appendix D.

## REFERENCES

James A Anderson. *An introduction to neural networks*. MIT press, 1995.

Robert J Aumann. Integrals of set-valued functions. *Journal of mathematical analysis and applications*, 12(1):1–12, 1965.

Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *ICML*, 2004.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13 (1):1063–1095, 2012.

Lynne Billard and Edwin Diday. From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487, 2003.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pp. 92–100, 1998.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *ICPR*, 2010.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *ICML*, 2010.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *COLT*, 2014.

Francisco de AT De Carvalho and Camilo P Tenório. Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems*, 161(23): 2978–2999, 2010.

Miguel Delgado, M Amparo Vila, and William Voxman. On a canonical representation of fuzzy numbers. *Fuzzy Sets and Systems*, 93(1):125–135, 1998.

Jozsef Dombi. Membership function as an evaluation. *Fuzzy Sets and Systems*, 35(1):1–21, 1990.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Bradley Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.

Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *ICML*. PMLR, 2016.

Pei-Yi Hao. Interval regression analysis using support vector networks. *Fuzzy Sets and Systems*, 160(17):2466–2485, 2009.

Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

Gholam Reza Jahanshahloo, F Hosseinzadeh Lotfi, and Mohammad Izadikhah. An algorithmic method to extend topsis for decision-making problems with interval data. *Applied Mathematics and Computation*, 175(2):1375–1384, 2006.

Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

Mehran Kafai and Kave Eshghi. Croification: Accurate kernel classification with the efficiency of sparse linear svm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):34–48, 2019.

Koray Kayabol. Approximate sparse multinomial logistic regression for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):490–493, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. $l_p$-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on k-means and smote. *arXiv preprint arXiv:1711.00837*, 2017.

Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In *NeurIPS*, 2018.

Wentao Li, Haoxiang Zhou, Weihua Xu, Xi-Zhao Wang, and Witold Pedrycz. Interval dominance-based feature selection for interval-valued ordered data. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Tung Yen Lin and Nick Cercone. *Rough sets and data mining: Analysis of imprecise data*. Springer Science & Business Media, 2012.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *ICML*. PMLR, 2014.

Dusthon Llorente, Mariana Ballesteros, Ivan DE JESUS Salgado Ramos, and Jorge Isaac Chairez Oria. Deep learning adapted to differential neural networks used as pattern classification of electrophysiological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Qian Lou and Lei Jiang. Hemet: A homomorphic-encryption-friendly privacy-preserving mobile neural network architecture. In *ICML*. PMLR, 2021.

Guangzhi Ma, Jie Lu, Feng Liu, Zhen Fang, and Guangquan Zhang. Multiclass classification with fuzzy-feature observations: Theory and algorithms. *IEEE Transactions on Cybernetics*, 2022. doi: 10.1109/TCYB.2022.3181193.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *ALT*, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Massachusetts Institute of Technology, 2012.

William S Noble. What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, 2006.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (2):604–624, 2020.

Mourad Oussalah. On the compatibility between defuzzification and fuzzy arithmetic operations. *Fuzzy Sets and Systems*, 128(2):247–260, 2002.

Yuqing Pan, Qing Mai, and Xin Zhang. Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association*, 114(527):1305–1319, 2018.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with pate. In *ICLR*, 2018.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *NeurIPS*, 2019.

John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

Leandro C Souza, Renata MCR Souza, Getúlio JA Amaral, and Telmo M Silva Filho. A parametrized approach for linear regression of interval data. *Knowledge-Based Systems*, 131: 149–159, 2017.

Luan Tran, Kihyuk Sohn, Xiang Yu, Xiaoming Liu, and Manmohan Chandraker. Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild. In *CVPR*, 2019.

Lev V Utkin and Frank PA Coolen. Interval-valued regression and classification models in the framework of machine learning. In *ISIPTA*, volume 11, pp. 371–380, 2011.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021a.

Shiping Wang, Zhaoliang Chen, Shide Du, and Zhouchen Lin. Learning deep sparse regularizers with applications to multi-view clustering and semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021b.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in Statistics*, pp. 196–202. Springer, 1992.

Changqing Zhang, Ehsan Adeli, Tao Zhou, Xiaobo Chen, and Dinggang Shen. Multi-layer multi-view classification for alzheimer's disease diagnosis. In *AAAI*, 2018a.

Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S Yu. From canonical correlation analysis to self-supervised graph neural networks. In *NeurIPS*, 2021.

Qiao Zhang, Cong Wang, Hongyi Wu, Chunsheng Xin, and Tran V Phuong. Gelu-net: A globally encrypted, locally unencrypted deep neural network for privacy-preserved learning. In *IJCAI*, 2018b.

Table 6: Main notations and their descriptions.

| Notation | Description |
|---|---|
| **• Spaces and Labels** | |
| $\bar{\mathbb{R}} = \{[x^{\mathrm{l}}, x^{\mathrm{r}}] \vert x^{\mathrm{l}}, x^{\mathrm{r}} \in \mathbb{R}, x^{\mathrm{l}} \leq x^{\mathrm{r}}\}$ | the set of all real-valued intervals |
| $\bar{\mathbb{R}}^p = \{([x_1^{\mathrm{l}}, x_1^{\mathrm{r}}], \cdots, [x_p^{\mathrm{l}}, x_p^{\mathrm{r}}])^\top\}$ | the set of all $p$-dimension interval-valued vector |
| $\bar{\mathcal{X}} \subset \bar{\mathbb{R}}^p$ | input (feature) space of LIND problem |
| $\mathcal{X}_v \subset \mathbb{R}^p, v \in [c]$ | single-view input (feature) space |
| $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_c \subset \mathbb{R}^p \times \cdots \times \mathbb{R}^p$ | multi-view input (feature) space |
| $\mathcal{Y}$ | output (label) space |
| $[K] = \{1, \cdots, K\}$ | $1, \cdots, K$ represent the labels in $\mathcal{Y}$ |
| **• Distributions** | |
| $\bar{X} = [X^{\mathrm{l}}, X^{\mathrm{r}}]$ | interval-valued random variable |
| $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ | interval-valued random vector |
| $\mathbf{X}^{\mathrm{l}} = (X_1^{\mathrm{l}}, \cdots, X_p^{\mathrm{l}})^\top,$ | real-valued random vector |
| $\mathbf{X}^{\mathrm{r}} = (X_1^{\mathrm{r}}, \cdots, X_p^{\mathrm{r}})^\top$ | |
| $\mathcal{D}^{\mathrm{l}}, \mathcal{D}^{\mathrm{r}}$ | distribution of real-valued random vector $\mathbf{X}^{\mathrm{l}}, \mathbf{X}^{\mathrm{r}}$ |
| $\bar{\mathcal{D}}$ | interval distribution over $\bar{\mathcal{X}}$ |
| $\mathcal{D}$ | multi-view distribution over $\mathcal{X}$ |
| $\bar{S}_{\bar{X}} = \{\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \cdots, \bar{x}_{ip})^\top\}_{i=1}^m$ | a sample drawn i.i.d. from $\bar{\mathcal{X}}$ |
| $S_{X^v} = \{\mathbf{x}_i^v = (x_{i1}^v, \cdots, x_{ip}^v)^\top\}_{i=1}^m$ | the single-view sample drawn i.i.d. from $\mathcal{X}_v$ |
| $S_X = \{\mathbf{X}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^c)\}_{i=1}^m$ | the multi-view sample drawn i.i.d. from $\mathcal{X}$ |
| **• Loss Function ad Function Spaces** | |
| $\ell(\cdot, \cdot)$ | loss : $\mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$ |
| $\mathcal{H}$ | hypothesis space of the LIND problem |
| $\mathcal{H}_v$ | hypothesis space of $v$-th view, $v = 1, \cdots, c$ |
| $\mathcal{H}_{\mathrm{co}}$ | multi-view hypothesis space |
| $f_v$ | predict function of $\boldsymbol{h}_v \in \mathcal{H}_v$, $v = 1, \cdots, c$ |
| $f_{\mathrm{co}}$ | predict function of $\boldsymbol{h}_{\mathrm{co}} \in \mathcal{H}_{\mathrm{co}}$ |
| **• Risks and Complexities** | |
| $R_{\bar{\mathcal{D}}}(\boldsymbol{h})$ | risk of $\boldsymbol{h} \in \mathcal{H}$ |
| $R_{\mathcal{D}}(\boldsymbol{h}_{\mathrm{co}})$ | risk of $\boldsymbol{h}_{\mathrm{co}} \in \mathcal{H}_{\mathrm{co}}$ |
| $\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H})$ | empirical Rademacher complexity of $\mathcal{H}$ with respect to the sample $\bar{S}_{\bar{X}}$ |
| $\mathcal{R}_{\bar{S}_{\bar{X}}}(\mathcal{H})$ | Rademacher complexity of $\bar{\mathcal{H}}$ with respect to the sample $\bar{S}_{\bar{X}}$ |
| $\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)$ | Rademacher complexity of $\mathcal{H}_v$ with respect to the sample $S_{X^v}$ |
| $\mathcal{R}_{S_X}(\mathcal{H}_{co})$ | Rademacher complexity of $\mathcal{H}_{co}$ with respect to the sample $S_X$ |

# A NOTATIONS

In this section, we summarize important notations in Table 6.

To prove Theorem 4, 5 and Corollary 1, for any $\boldsymbol{h}_v \in \mathcal{H}_v$, we let

$$\begin{aligned} \boldsymbol{h}_v(\mathbf{x}_i^v): \quad & \mathcal{X}_v \to \mathbb{R}^K \\ & \mathbf{x}_i^v \to (h_{v1}(\mathbf{x}_i^v), \cdots, h_{vK}(\mathbf{x}_i^v))^\top. \end{aligned}$$

Without loss of generality, we suppose that $\sum_{k=1}^{K} h_{vk}(\mathbf{x}_i^v) = 1$ and the predict function $f_v$ of $\boldsymbol{h}_v$ is defined as

$$f_v(\mathbf{x}_i^v) = \arg\max_{1\le k\le K} h_{vk}(\mathbf{x}_i^v).$$

Then, for any $\boldsymbol{h}_{\mathrm{co}} \in \mathcal{H}_{\mathrm{co}}$, we let

$$\begin{aligned}\boldsymbol{h}_{\mathrm{co}}(\mathbf{X}_i): \quad & \mathcal{X} \to \mathbb{R}^K \\ & \mathbf{X}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^c) \to (h_{\mathrm{co}}^1(\mathbf{X}_i), \cdots, h_{\mathrm{co}}^K(\mathbf{X}_i))^\top,\end{aligned}$$

where $\boldsymbol{h}_{\mathrm{co}}^q(\mathbf{X}_i) = \sum_{v=1}^{c} \mathbf{w}_v^{q\top} \boldsymbol{h}_v(\mathbf{x}_i^v), \mathbf{w}_v^q = (w_{v1}^q, \cdots, w_{vK}^q)^\top$ and without loss of generality, we suppose $\sum_{q=1}^{K} h_{\mathrm{co}}^q(\mathbf{X}_i) = 1$. Therefore, we have $\sup_{\boldsymbol{h}_{\mathrm{co}}\in\mathcal{H}_{\mathrm{co}}} \| \boldsymbol{h}_{\mathrm{co}} \|_\infty \le 1$. The predict function $f_{\mathrm{co}}$ of $\boldsymbol{h}_{\mathrm{co}}$ is defined as

$$f_{\mathrm{co}}(\mathbf{X}_i) = \arg\max_{1\le q\le K} h_{\mathrm{co}}^q(\mathbf{X}_i).$$

# B PROOFS AND FURTHER ANALYSIS

## B.1 PROOFS

In this section, we prove Theorem 1 in Section 3. To prove Theorem 1, we first give some related definitions and prove the Azuma's Inequality and McDiarmid's Inequality of interval-valued random variables.

### B.1.1 RELATED DEFINITIONS AND THEOREMS TO PROVE THEOREM 1

**Definition 3** (Interval Probability Density Function)**.** *Suppose $X^{\mathrm{l}}, X^{\mathrm{r}}$ are two real-valued random variables and have the same continuous pdf $p_X(x)$. We define $\bar{p}_{\bar{X}}(x)$ as the interval pdf of interval-valued random variable $\bar{X}$, where*

$$\bar{p}_{\bar{X}}(x) = \left[\min_{x\in[X^{\mathrm{l}},X^{\mathrm{r}}]} p_X(x), \max_{x\in[X^{\mathrm{l}},X^{\mathrm{r}}]} p_X(x)\right].$$

*Let $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ be a $p$-interval-valued random vector and the interval pdf of $\bar{X}_j$ is $\bar{p}_{\bar{X}_j}(x), j \in [p]$. Then, we denote the joint interval pdf of $\bar{\mathbf{X}}$ as*

$$\bar{p}_{\bar{\mathbf{X}}}(\mathbf{x}) = \left[\prod_{j=1}^{p} \min_{x_j\in[X_j^{\mathrm{l}},X_j^{\mathrm{r}}]} p_{X_j}(x_j), \prod_{j=1}^{p} \max_{x_j\in[X_j^{\mathrm{l}},X_j^{\mathrm{r}}]} p_{X_j}(x_j)\right], \mathbf{x} = (x_1, \cdots, x_p)^\top.$$

**Definition 4** (Interval Probability Distribution)**.** *Let $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ be a $p$-interval-valued random vector with the joint interval pdf $\bar{p}_{\bar{\mathbf{X}}}(\mathbf{x})$. Let $\mathbf{X}^{\mathrm{l}} = (X_1^{\mathrm{l}}, \cdots, X_p^{\mathrm{l}})^\top, \mathbf{X}^{\mathrm{r}} = (X_1^{\mathrm{r}}, \cdots, X_p^{\mathrm{r}})^\top$ be two real-valued random vectors following probability distribution $\mathcal{D}^{\mathrm{l}}, \mathcal{D}^{\mathrm{r}}$. We define $\bar{\mathcal{D}}$ as the interval probability distribution of $\bar{\mathbf{X}}$ (denoted as $\bar{\mathbf{X}} \sim \bar{\mathcal{D}}$), if*

$$\bar{\mathcal{D}}(\bar{\mathbb{R}}^p) = \bar{\int} \bar{p}_{\bar{\mathbf{X}}}(\mathbf{x}) d\mathbf{x} = 1,$$

*where $\bar{\int} \bar{p}_{\bar{\mathbf{X}}}(\mathbf{x}) d\mathbf{x} = \frac{1}{2}\int d\mathcal{D}^{\mathrm{l}}(\mathbf{x}) + \frac{1}{2}\int d\mathcal{D}^{\mathrm{r}}(\mathbf{x})$. Therefore, $\bar{\mathbf{X}} \sim \bar{\mathcal{D}}$ if and only if $\mathbf{X}^{\mathrm{l}} \sim \mathcal{D}^{\mathrm{l}}$ and $\mathbf{X}^{\mathrm{r}} \sim \mathcal{D}^{\mathrm{r}}$. Then, we denote $\mathbb{P}(\bar{\mathbf{X}} \in \bar{B}) = \bar{\mathcal{D}}(\bar{B})$ as the probability of the event $\{\bar{\mathbf{X}} \in \bar{B}\}$, where $\bar{B} \in \bar{\mathcal{B}}$ and $\bar{\mathcal{B}}$ is the Borel $\sigma$-algebra in $\bar{\mathbb{R}}^p$ (Jeffreys, 1998).*

**Definition 5.** *Let $\bar{\mathbf{X}} = (\bar{X}_1, \cdots, \bar{X}_p)^\top$ be a $p$-interval-valued random vector with the joint interval pdf $\bar{p}_{\bar{\mathbf{X}}}(\mathbf{x})$ and $\mathbf{X}^{\mathrm{l}} = (X_1^{\mathrm{l}}, \cdots, X_p^{\mathrm{l}})^\top \sim \mathcal{D}^{\mathrm{l}}, \mathbf{X}^{\mathrm{r}} = (X_1^{\mathrm{r}}, \cdots, X_p^{\mathrm{r}})^\top \sim \mathcal{D}^{\mathrm{r}}$ are two real-valued random vectors. Then, the probability with respect to the function $g : \bar{\mathcal{X}} \to \mathbb{R}_+$ is defined as:*

$$\mathbb{P}(g(\bar{\mathbf{X}}) \ge \varepsilon) = \frac{1}{2}\int_{\mathrm{A}} d\mathcal{D}^{\mathrm{l}}(\mathbf{x}) + \frac{1}{2}\int_{\mathrm{B}} d\mathcal{D}^{\mathrm{r}}(\mathbf{x}),$$

*where* $\mathrm{A} = \{\mathbf{X}^{\mathrm{l}} \in \mathbb{R}^p : g(\bar{\mathbf{X}}) \ge \varepsilon\}, \mathrm{B} = \{\mathbf{X}^{\mathrm{r}} \in \mathbb{R}^p : g(\bar{\mathbf{X}}) \ge \varepsilon\}$.

**Definition 6** (Independence)**.** *The interval-valued random vectors $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n$ are said to be (mutually) independent if and only if the real-valued random vectors $\mathbf{X}_1^{\mathrm{l}}, \cdots, \mathbf{X}_n^{\mathrm{l}}, \mathbf{X}_1^{\mathrm{r}}, \cdots, \mathbf{X}_n^{\mathrm{r}}$ are (mutually) independent. Then, we denote $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n$ as i.i.d. interval-valued random vectors if and only if $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_n$ are independent and have the same interval probability distribution.*

**Definition 7.** *The empirical Rademacher complexity of $\mathcal{H}$ with respect to $\bar{S}_{\bar{X}}$ is defined as:*

$$\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{h}\in\mathcal{H}} \tfrac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} \sigma_{ik} h_k(\bar{\mathbf{x}}_i)\right], \tag{2}$$

*where $\boldsymbol{\sigma} = [\sigma_{ik}]_{m\times K}$ is a $m \times K$ matrix, with $\sigma_{ik}$s independent random variables drawn from the Rademacher distribution, i.e. $\mathbb{P}(\sigma_{ik} = +1) = \mathbb{P}(\sigma_{ik} = -1) = \frac{1}{2}, i \in [m], k \in [K]$. The Rademacher complexity $\mathcal{R}_{\bar{S}_{\bar{X}}}(\mathcal{H})$ is equal to the interval expectation of $\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H})$.*

**Definition 8.** *A sequence of $V_1, V_2, \cdots$ is a martingale difference sequence with respect to interval-valued random variables $\bar{X}_1, \bar{X}_2, \cdots$ if for any $i > 0$, $V_i$ is a real-value function of $\bar{X}_1, \cdots, \bar{X}_i$ and $\mathbb{E}_{\bar{\mathcal{D}}}[V_{i+1}|\bar{X}_1, \cdots, \bar{X}_i] = 0$.*

**Theorem 2** (Azuma's Inequality of Interval-valued Random Variables)**.** *Let $V_1, V_2, \cdots$ be a martingale difference sequence with respect to the interval-valued random variables $\bar{X}_1, \bar{X}_2, \cdots$ and assume that for any $i > 0$ there is a constant $c_i \geq 0$ and $Z_i$, which is a real-value function of $\bar{X}_1, \cdots, \bar{X}_{i-1}$, satisfies*

$$Z_i \leq V_i \leq Z_i + c_i.$$

*Then for any $\varepsilon > 0$ and $m \in N^+$, the following inequalities hold:*

$$\begin{aligned} \mathbb{P}\left[\sum_{i=1}^{m} V_i \geq \varepsilon\right] &\leq \exp^{\frac{-2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}}, \\ \mathbb{P}\left[\sum_{i=1}^{m} V_i \leq -\varepsilon\right] &\leq \exp^{\frac{-2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}}. \end{aligned} \tag{3}$$

*Proof.* Suppose $\bar{X} = [X^{\mathrm{l}}, X^{\mathrm{r}}]$ is an interval-valued random variable. According to Definition 5, we have

$$\begin{aligned} \mathbb{P}(g(\bar{X}) \geq \varepsilon) &= \tfrac{1}{2}\left(\int_{\mathrm{A}} e^{-tg(\bar{X})} e^{tg(\bar{X})} d\mathcal{D}^{\mathrm{l}}(x) + \int_{\mathrm{B}} e^{-tg(\bar{X})} e^{tg(\bar{X})} d\mathcal{D}^{\mathrm{r}}(x)\right) \\ &\leq e^{-t\varepsilon}\tfrac{1}{2}\left(\int_{\mathrm{A}} e^{tg(\bar{X})} d\mathcal{D}^{\mathrm{l}}(x) + \int_{\mathrm{B}} e^{tg(\bar{X})} d\mathcal{D}^{\mathrm{r}}(x)\right) \\ &\leq e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tg(\bar{X})}]. \end{aligned}$$

By the convexity of $x \to e^x$, for any $x \in [a, b]$, the following holds:

$$e^{tx} \leq \tfrac{b-x}{b-a} e^{ta} + \tfrac{x-a}{b-a} e^{tb}.$$

Thus, using $\mathbb{E}_{\bar{\mathcal{D}}}[V_{i+1}|\bar{X}_1, \cdots, \bar{X}_i] = 0$, then

$$\begin{aligned} \mathbb{E}_{\bar{\mathcal{D}}}[e^{tV_{i+1}}|\bar{X}_1, \cdots, \bar{X}_i] &\leq \mathbb{E}_{\bar{\mathcal{D}}}\left[\tfrac{Z_{i+1}+c_{i+1}-V_{i+1}}{c_{i+1}} e^{tZ_{i+1}} + \tfrac{V_{i+1}-Z_{i+1}}{c_{i+1}} e^{t(Z_{i+1}+c_{i+1})}|\bar{X}_1, \cdots, \bar{X}_i\right] \\ &= \tfrac{Z_{i+1}+c_{i+1}}{c_{i+1}} e^{tZ_{i+1}} + \tfrac{-Z_{i+1}}{c_{i+1}} e^{t(Z_{i+1}+c_{i+1})} \leq e^{t^2 c_{i+1}^2/8}. \end{aligned}$$

Let $S_k = \sum_{i=1}^{k} V_i$. Then, for any $t > 0$, we can write

$$\begin{aligned} \mathbb{P}[S_m \geq \varepsilon] &\leq e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tS_m}] \\ &= e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tS_{m-1}}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tV_m}|\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_{m-1}]] \\ &\leq e^{-t\varepsilon}\mathbb{E}_{\bar{\mathcal{D}}}[e^{tS_{m-1}}]e^{t^2 c_m^2/8} (\text{iterating previous argument}) \\ &\leq e^{-t\varepsilon} e^{t^2 \sum_{i=1}^{m} c_i^2/8} (\text{let } t = 4\varepsilon/\sum_{i=1}^{m} c_i^2) = e^{\frac{-2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}}, \end{aligned}$$

the second statement is shown in a similar way. □

**Theorem 3** (McDiarmid's Inequality of Interval-valued Random Variables)**.** *Let $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_m \in \bar{\mathcal{X}} \subset \bar{\mathbb{R}}^p$ be a set of $m \geq 1$ interval-valued random vectors and assume that there exist $c_1, c_2, \cdots, c_m > 0$ such that $f : \bar{\mathcal{X}}^m \to \mathbb{R}$ satisfies the following conditions:*

$$|f(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_i, \cdots, \bar{\mathbf{X}}_m) - f(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_i^{'}, \cdots, \bar{\mathbf{X}}_m)| \leq c_i,$$

*for any $i \in [m]$ and any points $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_i, \cdots, \bar{\mathbf{X}}_m, \bar{\mathbf{X}}_i^{'} \in \bar{\mathcal{X}}$. Let $f(\bar{S})$ denote $f(\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_m)$, then, for any $\varepsilon > 0$, the following inequalities hold:*

$$\begin{aligned} \mathbb{P}[f(\bar{S}) - \mathbb{E}_{\bar{S}}[f(\bar{S})] \geq \varepsilon] &\leq \exp \frac{-2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}, \\ \mathbb{P}[f(\bar{S}) - \mathbb{E}_{\bar{S}}[f(\bar{S})] \leq -\varepsilon] &\leq \exp \frac{-2\varepsilon^2}{\sum_{i=1}^{m} c_i^2}. \end{aligned} \tag{4}$$

*Proof.* Define a sequence of random variables $V_k, k \in [m]$, as follows:

$$\begin{aligned} &V = f(\bar{S}) - \mathbb{E}_{\bar{S}}[f(\bar{S})], \\ &V_1 = \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1] - E_{\bar{S}}[V], \\ &V_k = \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k] - \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}]. \end{aligned}$$

Note that $V = \sum_{i=1}^{m} V_i$. Furthermore, the interval-valued random vector $\mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k]$ is a function of $\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k$, therefore:

$$\mathbb{E}_{\bar{S}}[\mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k]|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}] = \mathbb{E}_{\bar{S}}[V|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}],$$

which implies $\mathbb{E}_{\bar{S}}[V_k|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}] = 0$. Thus, the sequence $(V_k), k \in [m]$ is a martingale difference sequence. Next, observe that, since $\mathbb{E}_{\bar{S}}[f(\bar{S})]$ is a scalar, $V_k$ can be expressed as follows:

$$V_k = \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_{k-1}].$$

Thus, we can define an upper bound $W_k$ and lower bound $U_k$ for $V_k$ by:

$$\begin{aligned} W_k &= \sup_{\bar{\mathbf{X}}} \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k], \\ U_k &= \inf_{\bar{\mathbf{X}}^{'}} \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}^{'}] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k], \end{aligned}$$

$$\begin{aligned} W_k - U_k &= \sup_{\bar{\mathbf{X}}, \bar{\mathbf{X}}^{'}} \{\mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}] - \mathbb{E}_{\bar{S}}[f(\bar{S})|\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}^{'}]\} \\ &\leq \tfrac{1}{2} \sup_{\bar{\mathbf{X}}, \bar{\mathbf{X}}^{'}} \{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^{m-k}}[|f(\bar{S}_1) - f(\bar{S}_2)|] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^{m-k}}[|f(\bar{S}_1) - f(\bar{S}_2)|]\} \\ &\leq c_k, \end{aligned}$$

where $\bar{S}_1 = (\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}, \bar{\mathbf{X}}_{k+1}, \cdots, \bar{\mathbf{X}}_m), \bar{S}_1 = (\bar{\mathbf{X}}_1, \cdots, \bar{\mathbf{X}}_k, \bar{\mathbf{X}}^{'}, \bar{\mathbf{X}}_{k+1}, \cdots, \bar{\mathbf{X}}_m)$. Thus, $U_k \leq V_k \leq W_k \leq U_k + c_k$. In the view of these inequalities, we can apply Theorem 2 to $V = \sum_{i=1}^{m} V_i$, which yields the result. □

### B.1.2 Proof of Theorem 1

For any sample $\bar{S} = \{\bar{\mathbf{z}}_i = (\bar{\mathbf{x}}_i, y_i)\}_{i=1}^m \sim \bar{\mathcal{D}}^m$ and any $\ell \in \mathcal{L}_{\mathcal{H}}$, we denote

$$\Phi(\bar{S}) = \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \{\mathbb{E}_{\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \tfrac{1}{m} \sum_{i=1}^{m} \ell(\bar{\mathbf{z}}_i)\} = \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \{\mathbb{E}_{\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}.$$

Let $\bar{S}$ and $\bar{S}^{'}$ be two samples differing by exactly one point, say $\bar{\mathbf{z}}_m$ in $\bar{S}$ and $\bar{\mathbf{z}}_m^{'}$ in $\bar{S}^{'}$. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$\Phi(\bar{S}^{'}) - \Phi(\bar{S}) \leq \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \{\widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})] - \widehat{\mathbb{E}}_{\bar{S}^{'}}[\ell(\bar{\mathbf{z}})]\} \leq \sup_{\ell \in \mathcal{L}_{\mathcal{H}}} \tfrac{\ell(\bar{\mathbf{z}}_m) - \ell(\bar{\mathbf{z}}_m^{'})}{m} \leq \tfrac{C_\ell}{m}.$$

Similarly, we can obtain $\Phi(\bar{S}) - \Phi(\bar{S}^{'}) \leq \frac{C_\ell}{m}$, thus $|\Phi(\bar{S}^{'}) - \Phi(\bar{S})| \leq \frac{C_\ell}{m}$. Based on Definition 2, $\Phi(\bar{S})$ is a function of random variables $\mathbf{X}_i^{\mathrm{l}}$ and $\mathbf{X}_i^{\mathrm{r}}$ and we have

$$\begin{aligned}\mathbb{E}_{\bar{S}'}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})]\} &= \tfrac{1}{2}\{\mathbb{E}_{\mathcal{D}^{\mathrm{l}}}[\tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i^{'})] + \mathbb{E}_{\mathcal{D}^{\mathrm{r}}}[\tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i^{'})]\} \\ &= \tfrac{1}{2}\{\tfrac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathcal{D}^{\mathrm{l}}}[\ell(\bar{\mathbf{z}}_i^{'})] + \tfrac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\mathcal{D}^{\mathrm{r}}}[\ell(\bar{\mathbf{z}}_i^{'})]\} \\ &= \tfrac{1}{2}\{\mathbb{E}_{\mathcal{D}^{\mathrm{l}}}[\ell(\bar{\mathbf{z}})] + \mathbb{E}_{\mathcal{D}^{\mathrm{r}}}[\ell(\bar{\mathbf{z}})]\} = \mathbb{E}_{\bar{S}'}[\ell(\bar{\mathbf{z}})].\end{aligned}$$

Then, by Theorem 3, for any $\delta > 0$, with probability at least $1 - \delta/2$, the following holds:

$$\Phi(\bar{S}) \leq \mathbb{E}_{\bar{S}}[\Phi(\bar{S})] + C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}},$$

$$\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] = \mathbb{E}_{\bar{S}}[\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\mathbb{E}_{\bar{S}'}[\ell(\bar{\mathbf{z}})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}] = \mathbb{E}_{\bar{S}}[\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\mathbb{E}_{\bar{S}'}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}].$$

Because

$$\begin{aligned}&\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\mathbb{E}_{\bar{S}'}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\} \\ =\ &\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\tfrac{1}{2}\{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]]\} \\ \leq\ &\tfrac{1}{2}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^m}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]]\} \\ \leq\ &\tfrac{1}{2}\{\mathbb{E}_{(\mathcal{D}^{\mathrm{l}})^m}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]] + \mathbb{E}_{(\mathcal{D}^{\mathrm{r}})^m}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}[\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]]\} \\ =\ &\mathbb{E}_{\bar{S}'}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}.\end{aligned}$$

Then, we have

$$\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] \leq \mathbb{E}_{\bar{S},\bar{S}'}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\widehat{\mathbb{E}}_{\bar{S}'}[\ell(\bar{\mathbf{z}}^{'})] - \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\} = \mathbb{E}_{\bar{S},\bar{S}'}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\tfrac{1}{m}\sum_{i=1}^{m}[\ell(\bar{\mathbf{z}}_i^{'}) - \ell(\bar{\mathbf{z}}_i)]\}.$$

We introduce Rademacher variables $\sigma_i$s, that are uniformly distributed independent random variables taking values in $\{-1, +1\}$,

$$\begin{aligned}\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] &\leq \mathbb{E}_{\bar{S},\bar{S}'}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\tfrac{1}{m}\sum_{i=1}^{m}[\sigma_i\ell(\bar{\mathbf{z}}_i^{'}) - \ell(\bar{\mathbf{z}}_i)]\}(\sup(U+V) \leq \sup U + \sup V) \\ &\leq \mathbb{E}_{\bar{S}'}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\tfrac{1}{m}\sum_{i=1}^{m}\sigma_i\ell(\bar{\mathbf{z}}_i^{'})\} + \mathbb{E}_{\bar{S}}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\tfrac{1}{m}\sum_{i=1}^{m}-\sigma_i\ell(\bar{\mathbf{z}}_i)\}.\end{aligned}$$

Because the definition of Rademacher complexity and the fact that the variables $\sigma_i$ and $-\sigma_i$ are distributed in the same way, then

$$\mathbb{E}_{\bar{S}}[\Phi(\bar{S})] \leq 2\mathbb{E}_{\bar{S}}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\tfrac{1}{m}\sum_{i=1}^{m}\sigma_i\ell(\bar{\mathbf{z}}_i)\} = 2\mathcal{R}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}).$$

Then using $\delta$ instead of $\delta/2$, with probability $1 - \delta$, the following holds :

$$\begin{aligned}\Phi(\bar{S}) &\leq 2\mathcal{R}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) + C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}} \\ \mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \tfrac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i) &\leq 2\mathcal{R}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) + C_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}.\end{aligned} \tag{5}$$

We observe that changing one point in $\bar{S}$ changes $\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}})$ by at most $C_\ell/m$. Then, again using Theorem 3, with probability $1 - \delta/2$ the following holds:

$$\mathcal{R}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) \leq \widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) + C_\ell\sqrt{\tfrac{\log(2/\delta)}{2m}}.$$

Then with probability at least $1-\delta$:

$$\begin{aligned}\Phi(\bar{S}) &\leq 2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) + 3C_{\ell}\sqrt{\frac{\log(1/\delta)}{2m}} \\ \mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \frac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i) &\leq 2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) + 3C_{\ell}\sqrt{\frac{\log(1/\delta)}{2m}}.\end{aligned} \tag{6}$$

Next we let,

$$\Psi(\bar{S}) = \inf_{\ell\in\mathcal{L}_{\mathcal{H}}}\{\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \frac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i)\} = -\sup_{\ell\in\mathcal{L}_{\mathcal{H}}}\{-\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] + \widehat{\mathbb{E}}_{\bar{S}}[\ell(\bar{\mathbf{z}})]\}.$$

In the same way, with probability at least $1-\delta$ the following holds:

$$\begin{aligned}\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \frac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i) &\geq -2\mathcal{R}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) - C_{\ell}\sqrt{\frac{\log(1/\delta)}{2m}} \\ \mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \frac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i) &\geq -2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) - 3C_{\ell}\sqrt{\frac{\log(2/\delta)}{2m}}.\end{aligned} \tag{7}$$

Since $\ell$ is Lipschitz continuous, according to Maurer (2016), we have

$$\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) \leq \sqrt{2}L_{\ell}\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}). \tag{8}$$

Following from Eqs. (5), (6), (7) and for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $\ell \in \mathcal{L}_{\mathcal{H}}$:

$$|\mathbb{E}_{\bar{\mathbf{z}}\sim\bar{\mathcal{D}}}[\ell(\bar{\mathbf{z}})] - \frac{1}{m}\sum_{i=1}^{m}\ell(\bar{\mathbf{z}}_i)| \leq 2\widehat{\mathcal{R}}_{\bar{S}}(\mathcal{L}_{\mathcal{H}}) + 3C_{\ell}\sqrt{\frac{\log(2/\delta)}{2m}}. \tag{9}$$

Using $R_{\bar{\mathcal{D}}}(\boldsymbol{h}) = \mathbb{E}_{\bar{\mathcal{D}}}[\ell(\boldsymbol{h}(\bar{X}), y)]$ and Eqs. (8) and (9), we have for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $\ell \in \mathcal{L}_{\bar{\mathcal{H}}}$:

$$|R_{\bar{\mathcal{D}}}(\boldsymbol{h}) - \widehat{R}_{\bar{\mathcal{D}}}(\boldsymbol{h})| \leq 2\sqrt{2}L_{\ell}\widehat{\mathcal{R}}_{\bar{S}_{\bar{X}}}(\mathcal{H}) + 3C_{\ell}\sqrt{\frac{\log(2/\delta)}{2m}}.$$

### B.2 Further Analysis

In this section, we consider why using multi-view learning to address the LIND problem in terms of error rate and estimation error bound.

Let $\mathcal{X}_v$ ($v \in [c]$) be the single-view input space and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_c$ be the muti-view input space. Let $S_X = \{\mathbf{X}_i = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^c)\}_{i=1}^m \subset \mathcal{X}$ be the multi-view sample drawn i.i.d. form $\mathcal{D}$, where $\mathbf{x}_i^v$ is the single-view observation, $\mathbf{X}_i \in \mathcal{X}$ and $y_i = f(\mathbf{X}_i) \in \mathcal{Y}$ is the ground-truth function. Let $\mathcal{H}_v$ be the hypothesis space of $v$-th view, where for any $\boldsymbol{h}_v \in \mathcal{H}_v$, $\boldsymbol{h}_v : \mathcal{X}_v \to \mathbb{R}^K$. Then, $f_v : \mathbb{R}^K \to \mathcal{Y}$ is a predict function induced by $\boldsymbol{h}_v$. Lastly, we set $\mathcal{H}_{\text{co}}$ to be the multi-view hypothesis space, where for any $\boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}$, $\boldsymbol{h}_{\text{co}} : \mathcal{X} \to \mathbb{R}^K$. Then, we can induce a predict function $f_{\text{co}} : \mathbb{R}^K \to \mathcal{Y}$ by $\boldsymbol{h}_{\text{co}}$.

#### B.2.1 Error Rate

First, we propose a notion called discrepancy set to measure the predict functions difference across different view. Then, we denote $\mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)$ as the discrepancy set between the predict functions $f_1, \cdots, f_c$ over $\mathcal{X}$, which is shown as follow:

$$\mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c) = \left\{\mathbf{X} = (\mathbf{x}^1, \cdots, \mathbf{x}^c) \in \mathcal{X} : \bigvee_{1\leq v_1<v_2\leq c} f_{v_1}(\mathbf{x}^{v_1}) \neq f_{v_2}(\mathbf{x}^{v_2})\right\},$$

here $\bigvee$ represents the logical relation "or". Next, we give the following assumption:

$$\text{For } \mathbf{X} = (\mathbf{x}^1, \cdots, \mathbf{x}^c), \text{ if } f_1(\mathbf{x}^1) = \cdots = f_c(\mathbf{x}^c), \text{ we have } f_{\text{co}}(\mathbf{X}) = f_1(\mathbf{x}^1). \tag{10}$$

This assumption means that if all single-view predictions are same, the multi-view predict function also has the same outcome, which is a trivial assumption. Then, we obtain the following theorem.

**Theorem 4.** *We assert that there exists a uniform constant $M \in (0,1)$ such that for any predict function $f_{\text{co}}$ satisfies assumption (10), if*

$$\mathbb{P}_{\mathcal{D}}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)) \leq M, \text{ where } y \text{ is the ground-truth label},$$

*we assert* $\text{err}(f_{\text{co}}) \leq \min_{v \in [c]} \text{err}(f_v)$, *where* $\text{err}(f_{\text{co}}) = \mathbb{P}_{\mathcal{D}}(f_{\text{co}}(\mathbf{X}) \neq y)$.

*Proof.* Without loss of generality, we suppose $\text{err}(f_1) \leq \cdots \leq \text{err}(f_c)$. First, we consider the case where $c = 2$. Then, we provide an upper bound on the error rate of $f_{\text{co}}$.

$$\begin{aligned} \text{err}(f_{\text{co}}) &= \mathbb{P}_{\mathcal{D}}(f_{\text{co}}(\mathbf{X}) \neq y) \\ &= \mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}^{C}(f_1, f_2)) + \mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, f_2)) \\ &\leq \tfrac{1}{2}[\text{err}(f_1) + \text{err}(f_2) - \mathbb{P}_{\mathcal{D}}(\mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, f_2))] + \mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | X \in \mathbf{D}_{\mathcal{F}}(f_1, f_2)), \end{aligned} \tag{11}$$

where $\mathbf{D}_{\mathcal{F}}^{C}(f_1, f_2)$ is denoted as the complement set of $\mathbf{D}_{\mathcal{F}}(f_1, f_2)$. According to Eq. (11) and $\text{err}(f_1) \leq \text{err}(f_2)$, if

$$\mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, f_2)) \leq \tfrac{1}{2}[\text{err}(f_1) - \text{err}(f_2) + \mathbb{P}_{\mathcal{D}}(\mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, f_2))],$$

we have $\text{err}(f_{\text{co}}) \leq \text{err}(f_1)$. Next, we consider the case where $c > 2$. For $c > 2$, we have $\boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}$,

$$h_{\text{co}}^{q}(\mathbf{X}) = \sum_{v=1}^{k+1} \mathbf{w}_v^{q\top} \boldsymbol{h}_v(\mathbf{x}^v) = \mathbf{w}_1^{q\top} \boldsymbol{h}_1(\mathbf{x}^1) + \sum_{v=2}^{c} \mathbf{w}_v^{q\top} \boldsymbol{h}_v(\mathbf{x}^v).$$

So exists $\alpha_q \in \mathbb{R}_+$, such that $\sum_{q=1}^{K} \alpha_q \sum_{v=2}^{c} \mathbf{w}_v^{q\top} \boldsymbol{h}_v(\mathbf{x}^v) = 1$, then exists

$$\boldsymbol{h}_{\text{co}}^{c-1} \in \mathcal{H}_{\text{co}}^{c-1}(\mathbf{x}^2, \cdots, \mathbf{x}^c), \text{ where } h_{\text{co}}^{c-1,q} = \alpha_q \sum_{v=2}^{c} \mathbf{w}_v^{q\top} \boldsymbol{h}_v(\mathbf{x}^v).$$

We combine the last $c-1$ views i.e., $\mathbf{X}' = (\mathbf{x}^2, \cdots, \mathbf{x}^c), \mathbf{X} = (\mathbf{x}^1, \mathbf{X}')$. So exists

$$\boldsymbol{h}_{\text{co}}^{c-1} \in \mathcal{H}_{\text{co}}^{c-1}(\mathbf{x}^2, \cdots, \mathbf{x}^c) \subset \mathcal{H}(\mathbf{X}'), \text{ such that } h_{\text{co}}^{q}(\mathbf{X}) = \mathbf{w}_1^{q\top} \boldsymbol{h}_1(\mathbf{x}^1) + \tfrac{1}{\alpha_q} h_{\text{co}}^{c-1,q}(\mathbf{X}').$$

Therefore we have $\boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}(\mathbf{x}^1, \mathbf{X}')$. Let $f_{\text{co}}^{c-1}(\mathbf{X}) = \arg\max_{1 \leq q \leq K} h_{\text{co}}^{c-1,q}(\mathbf{X})$ denoted as the predict function of $\boldsymbol{h}_{\text{co}}^{c-1}$. Because the conclusion is true when $c = 2$, so exists $M \in (0,1)$, such that

$$\text{if } \mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, f_{\text{co}}^{c-1})) \leq M, \text{we have } \text{err}(f_{\text{co}}) \leq \text{err}(f_1).$$

Because $\mathbf{D}_{\mathcal{F}}(f_1, f_{\text{co}}^{c-1}) \subset \mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)$, so

$$\mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, f_{\text{co}}^{c-1})) \leq \mathbb{P}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)).$$

Therefore, the conclusion is true when $c > 2$ which yields the result. □

We can easily find $M < 1$ that satisfies the condition in Theorem 4. According to Theorem 4, we always have the error rate of a multi-view prediction function $f_{co}$ is lower than that of any single-view prediction function $f_v, v \in [c]$ when $\mathbb{P}_{\mathcal{D}}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)) \to 0$, which means that using multi-view methodology can reduce the error rate of the predict function for the classification tasks. We can achieve $\mathbb{P}_{\mathcal{D}}(f_{\text{co}}(\mathbf{X}) \neq y | \mathbf{X} \in \mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)) \to 0$ by reducing the size of the discrepancy set $\mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)$. Based on the above theoretical analysis, we decide to find appropriate multi-view features that can achieve well and similar performance on all single-view classifiers to reduce the size of the discrepancy set $\mathbf{D}_{\mathcal{F}}(f_1, \cdots, f_c)$.

### B.2.2 Estimation Error Bound

$\mathcal{L}_{\mathcal{H}_{\text{co}}} = \{\ell(\boldsymbol{h}_{\text{co}}(\mathbf{X}), y) : \mathbf{X} \in \mathcal{X}, \boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}, y \in \mathcal{Y}\}$ be the class of functions with respect to the loss $\ell$ and $\mathcal{H}_{\text{co}}$, where $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$. The risk of $\boldsymbol{h}_{\text{co}}$ is denoted as $R_{\mathcal{D}}(\boldsymbol{h}_{\text{co}}) = \mathbb{E}_{\mathcal{D}}[\ell(\boldsymbol{h}_{\text{co}}(\mathbf{X}), y)]$. According to Theorem 3.1 and 3.2 in Mohri et al. (2012) and Theorem 2 in Maurer (2016), we obtain the following corollary.

**Corollary 1.** *Suppose that* $\sup_{\|\boldsymbol{h}_{\text{co}}\|_\infty \leq 1} \max_y \ell(\boldsymbol{h}_{\text{co}}, y) \leq C'_\ell$*, and all functions in* $\mathcal{L}_{\mathcal{H}_{\text{co}}}$ *are* $L_{\text{co}}$*-Lipschitz functions, and* $||\mathbf{W}||_2 \leq \Lambda$ *(*$\mathbf{W}$ *see Appendix B). For any* $\delta > 0$*, with probability at least* $1-\delta$*, each of the following holds for any* $\boldsymbol{h}_{\text{co}} \in \mathcal{H}_{\text{co}}$*:*

$$|R_{\mathcal{D}}(\boldsymbol{h}_{\text{co}}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{h}_{\text{co}})| \leq 2L_{\text{co}}\sqrt{\tfrac{2Kc\Lambda^2}{m}} + C'_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}. \tag{12}$$

*Proof.* According to Theorem 3.1, 3.2 in Mohri et al. (2012) and Theorem 2 in Maurer (2016), we have

$$|R_{\mathcal{D}}(\boldsymbol{h}_{\text{co}}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{h}_{\text{co}})| \leq 2\sqrt{2}L_{\text{co}}\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) + C'_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}. \tag{13}$$

Next, let

$$\mathbf{W} = \left(\mathbf{w}_1^{1\top}, \cdots, \mathbf{w}_c^{1\top}, \cdots, \mathbf{w}_1^{K\top}, \cdots, \mathbf{w}_c^{K\top}\right)^\top,$$
$$\mathbf{H} = \left(\sum_{i=1}^m \sigma_{i1}\boldsymbol{h}_1(\mathbf{x}_i^1)^\top, \cdots, \sum_{i=1}^m \sigma_{i1}\boldsymbol{h}_c(\mathbf{x}_i^c)^\top, \cdots, \sum_{i=1}^m \sigma_{iK}\boldsymbol{h}_1(\mathbf{x}_i^1)^\top, \cdots, \sum_{i=1}^m \sigma_{iK}\boldsymbol{h}_c(\mathbf{x}_i^c)^\top\right)^\top.$$

Then, we have

$$\begin{aligned}
\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) &= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_{\text{co}}\in\mathcal{H}_{\text{co}}} \sum_{i=1}^m \sum_{q=1}^K \sigma_{iq} h_{\text{co}}^q(\mathbf{X}_i)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_j\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} \sum_{i=1}^m \sum_{q=1}^K \sigma_{iq} \sum_{v=1}^c \mathbf{w}_v^{q\top}\boldsymbol{h}_v(\mathbf{x}_i^v)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} \langle\mathbf{W},\mathbf{H}\rangle] \\
&\leq \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v, ||\mathbf{W}||_2\leq\Lambda} ||\mathbf{W}||_2||\mathbf{H}||_2] \text{(using Cauchy-Schwarz inequality)} \\
&\leq \tfrac{\Lambda}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v} [\sum_{v=1}^c \sum_{q=1}^K || \sum_{i=1}^m \sigma_{iq}\boldsymbol{h}_v(\mathbf{x}_i^v)||_2^2]^{\frac{1}{2}}] \\
&\text{(using Jensen's inequality and } i\neq j \Rightarrow \mathbb{E}_{\vec{\sigma}}[\sigma_{ip}\sigma_{jp}] = 0) \\
&\leq \tfrac{\Lambda}{m}\left[\mathbb{E}_{\mathcal{D}}[\sup_{\boldsymbol{h}_j\in\mathcal{H}_j} K\sum_{i=1}^m \sum_{v=1}^c ||\boldsymbol{h}_v(\mathbf{x}_i^v)||_2^2]\right]^{\frac{1}{2}} \\
&\leq \tfrac{\Lambda}{m}\sqrt{Kcm} = \sqrt{\tfrac{Kc\Lambda^2}{m}}.
\end{aligned}$$

Then, we yield the final result

$$|R_{\mathcal{D}}(\boldsymbol{h}_{\text{co}}) - \widehat{R}_{\mathcal{D}}(\boldsymbol{h}_{\text{co}})| \leq 2L_{\text{co}}\sqrt{\tfrac{2Kc\Lambda^2}{m}} + C'_\ell\sqrt{\tfrac{\log(1/\delta)}{2m}}. \tag{14}$$

□

Corollary 1 presents a generalization bound of the discrepancy between the risk and empirical risk of $\boldsymbol{h}_{\text{co}}$. Finally, we give the following theorem to bound $\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}})$.

**Theorem 5.** *For any* $m \geq 1$*, we have* $\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) \leq \max_{v\in[c]} \mathcal{R}_{S_{X^v}}(\mathcal{H}_v)$*, where* $S_{X^v} = \{\mathbf{x}_i^v\}_{i=1}^m$*.*

*Proof.* Because $\sum_{q=1}^K \sum_{v=1}^c \sum_{k=1}^K w_{vk}^q h_{vk}(\mathbf{x}_i^v) = 1$ and for any $v \in [c], k \in [K], 0 \leq h_{vk}(\mathbf{x}_i^v) \leq 1$, so $\sum_{q=1}^K \sum_{v=1}^c \sum_{k=1}^K w_{vk}^q \leq 1$. Then,

$$
\begin{aligned}
\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}}) &= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_{\text{co}}\in\mathcal{H}_{\text{co}}}\sum_{i=1}^{m}\sum_{q=1}^{K}\sigma_{iq}h_{\text{co}}^{q}(\mathbf{X}_i)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v,||\mathbf{W}||_2\leq\Lambda}\sum_{i=1}^{m}\sum_{q=1}^{K}\sigma_{iq}\sum_{v=1}^{c}\mathbf{w}_v^{q\top}\boldsymbol{h}_v(\mathbf{x}_i^v)] \\
&= \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v,||\mathbf{W}||_2\leq\Lambda}\sum_{v=1}^{c}\sum_{q=1}^{K}\sum_{k=1}^{K}w_{vk}^{q}\sum_{i=1}^{m}\sigma_{iq}h_{vk}(\mathbf{x}_i^v)] \\
&\leq \tfrac{1}{m}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}[\sup_{\boldsymbol{h}_v\in\mathcal{H}_v}\max_{v\in[c],q\in[K]}\sum_{i=1}^{m}\sum_{k=1}^{K}\sigma_{ik}h_{vk}(\mathbf{x}_i^v)] \\
&\leq \max_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v) \\
&= \min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)+\max_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)-\min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)
\end{aligned}
$$

□

According to Theorem 5, if $\max_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)-\min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)\to 0$, we have $\mathcal{R}_{S_X}(\mathcal{H}_{\text{co}})\leq\min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)$, which demonstrates that we can obtain tighter estimation error bound by applying the multi-view methodology. Inspired by the above theoretical analysis, we achieve $\max_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)-\min_{v\in[c]}\mathcal{R}_{S_{X^v}}(\mathcal{H}_v)\to 0$ by finding appropriate multi-view features that can achieve similar performance on all single-view classifiers.

## C Membership Function-based Method

In this section, we give further details of the membership function-based method to extract multi-view information from interval-valued data.

First, we introduce two types of fuzzy number and four different defuzzification methods used to construct the membership function-based method. The first type of fuzzy number called triangular fuzzy number. A triangular fuzzy number $\widetilde{x}$ can be characterized by $\text{Tr}(a_1,b_1,a_2)$ and the membership function is shown as follows:

$$
\mu_{\widetilde{x}}(t)=\begin{cases}0, & t<a_1\\ \dfrac{t-a_1}{b_1-a_1}, & a_1\leq t<b_1\\ \dfrac{t-a_2}{b_1-a_2}, & b_1\leq t<a_2\\ 0, & t\geq a_2.\end{cases}
$$

Gaussian fuzzy number is the second type of fuzzy number. A Gaussian fuzzy number $\widetilde{x}$ can be characterized by $\text{Ga}(c,\delta_1,\delta_2)$ and the membership function is given in the following equation:

$$
\mu_{\widetilde{x}}(t)=\begin{cases}\exp(-(t-c)/2\delta_1)^2, & t<c\\ \exp(-(t-c)/2\delta_2)^2, & t\geq c.\end{cases}
$$

Next, we introduce the four different defuzzification methods.

**MOM.** The first method is called *Mean/Middle of Maxima* (MOM) (Oussalah, 2002) which is widely-used due to its calculation simplicity. MOM is defined as:

$$
\text{MOM}(\widetilde{x})=\text{Mean}(t=\arg\max_t\mu_{\widetilde{x}}(t)). \tag{15}
$$

**COG.** *The Centre of Gravity* (COG) (Oussalah, 2002) is another widely-used defuzzification method. The definitions of COG for discrete and continuous membership functions are shown as follows:

$$
\text{COG}(\widetilde{x})=\frac{\sum t\mu_{\widetilde{x}}(t)}{\sum\mu_{\widetilde{x}}(t)}(\text{discrete})=\frac{\int t\mu_{\widetilde{x}}(t)dt}{\int\mu_{\widetilde{x}}(t)dt}(\text{continuous}). \tag{16}
$$

Table 7: Hyperparameters for the proposed method and four baselines

| Algorithm | Basic classifier | Hyperparameters | Ranges |
|---|---|---|---|
| DF-SVM | | regularization parameter, kernel type, shape parameter $\beta$ | $\{0.1, 0.2, \cdots, 1, 2, \cdots, 10\}$, {'linear', 'poly', 'rbf'}, $\{0, 0.1, \cdots, 1\}$ |
| DF-MLP | | learning rate, shape parameter $\beta$ | $\{0.001, 0.01, 0.1\}$, $\{0, 0.1, \cdots, 1\}$ |
| L-IIE, U-IIE, M-IIE, Mv-IIE-2, Mv-IIE-3 | SVM | regularization parameter, kernel type | $\{0.1, 0.2, \cdots, 1, 2, \cdots, 10\}$, {'linear', 'poly', 'rbf'} |
| | RF | min samples leaf, the number of trees | $\{1, \cdots, 10\}$, $\{5, 10, \cdots, 100\}$ |
| | Net | learning rate | $\{0.001, 0.01, 0.1\}$ |
| Mv-IIE | same above | same above, shape parameter $\beta$ | same above, $\{0, 0.1, \cdots, 1\}$ |

**ALC.** The third approach, called *averaging level cuts* (ALC) (Oussalah, 2002), is defined as the flat averaging of all midpoints of the $\alpha$-cuts.

$$\mathrm{ALC}(\widetilde{x}) = \tfrac{1}{2}\int_0^1 (\widetilde{x}_\alpha^L + \widetilde{x}_\alpha^U)d\alpha. \tag{17}$$

**VAL.** The final method is called *value of a fuzzy number* (VAL) (Delgado et al., 1998) which uses $\alpha$-levels as weighting factors in averaging the $\alpha$-cut midpoints. VAL is defined as :

$$\mathrm{VAL}(\widetilde{x}) = \int_0^1 \alpha(\widetilde{x}_\alpha^L + \widetilde{x}_\alpha^U)d\alpha. \tag{18}$$

We denote $D = \{(\bar{\mathbf{x}}_i, y_i)\}_{i=1}^m$ as the interval-valued dataset, where $\bar{\mathbf{x}}_i = (\bar{x}_{i1}, \cdots, \bar{x}_{ip})^\top \in \bar{\mathbb{R}}^p, y_i \in [K]$. Then, the construction process of the membership function-based method is introduced. We divide this method into two parts. In the first part, we use two functions $F_1(\cdot;\beta), F_2(\cdot;\beta)$ to transfer a interval-valued feature to a triangular fuzzy number and a Gaussian fuzzy number respectively. $F_1(\cdot;\beta), F_2(\cdot;\beta)$ are defined as:

$$\begin{gathered}
F_1(\bar{x}_{ij};\beta) = \mathrm{Tr}(x_{ij}^{\mathrm{l}}, \beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, x_{ij}^{\mathrm{r}}), \\
F_2(\bar{x}_{ij};\beta) = \mathrm{Ga}(\beta x_{ij}^{\mathrm{l}} + (1-\beta)x_{ij}^{\mathrm{r}}, S_{1j}, S_{2j}), \\
S_{1j} = \sqrt{\mathrm{Var}(A_j)}, S_{2j} = \sqrt{\mathrm{Var}(B_j)}, \\
A_j = \{x_{ij}^{\mathrm{l}} : i \in [m], (\bar{\mathbf{x}}_i, y_i) \in D\}, B_j = \{x_{ij}^{\mathrm{r}} : i \in [m], (\bar{\mathbf{x}}_i, y_i) \in D\}, j \in [p],
\end{gathered}$$

where $\beta \in [0,1]$ is a hyperparameter to control the shape of the membership function, $\mathrm{Var}(\cdot)$ is used to find the variance of the set. Using the above process, one interval-valued feature $\bar{\mathbf{x}}_i$ can be transferred into two fuzzy-valued features $\widetilde{\mathbf{x}}_i^1 = (\widetilde{x}_{i1}^1, \cdots, \widetilde{x}_{ip}^1)^\top$ and $\widetilde{\mathbf{x}}_i^2 = (\widetilde{x}_{i1}^2, \cdots, \widetilde{x}_{ip}^2)^\top$, where

$$\widetilde{\mathbf{x}}_i^\tau = \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta) = (F_\tau(\bar{x}_{i1};\beta), \cdots, F_\tau(\bar{x}_{ip};\beta))^\top, \tau = 1, 2.$$

In the second part, we use the four defuzzification methods to transfer the two fuzzy-valued features $\widetilde{\mathbf{x}}_i^1, \widetilde{\mathbf{x}}_i^2$ into eight crisp-valued features

$$\mathrm{MOM} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{COG} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{ALC} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{VAL} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \tau = 1, 2.$$

According to Eq. (15), we find that $\mathrm{MOM} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta) = \mathrm{MOM} \circ \mathbf{F}_2(\bar{\mathbf{x}}_i;\beta)$. Therefore, we can use the aforementioned membership function-based method to extract multi-view information, which contains seven parts: $\mathrm{MOM} \circ \mathbf{F}_1(\bar{\mathbf{x}}_i;\beta)$ and $\mathrm{COG} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{ALC} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \mathrm{VAL} \circ \mathbf{F}_\tau(\bar{\mathbf{x}}_i;\beta), \tau = 1, 2$. We denote $\mathcal{T} = \{\mathbf{T}_v(\cdot;\beta)\}_{v=1}^7$ as a set of transfer functions constructed by using the membership function-based method, where

$$\begin{gathered}
\mathbf{T}_1 = \mathrm{MOM} \circ \mathbf{F}_1, \mathbf{T}_2 = \mathrm{COG} \circ \mathbf{F}_1, \mathbf{T}_3 = \mathrm{COG} \circ \mathbf{F}_2, \mathbf{T}_4 = \mathrm{ALC} \circ \mathbf{F}_1, \\
\mathbf{T}_5 = \mathrm{ALC} \circ \mathbf{F}_2, \mathbf{T}_6 = \mathrm{VAL} \circ \mathbf{F}_1, \mathbf{T}_7 = \mathrm{VAL} \circ \mathbf{F}_2.
\end{gathered}$$

By applying the aforementioned transfer functions to extract crisp-valued information from the interval-valued data, one interval-valued feature $\bar{\mathbf{x}}_i$ can be transferred into seven different parts $\mathbf{X}_i^{\mathrm{Mv}} = (\mathbf{x}_i^1, \cdots, \mathbf{x}_i^7)$, where for any $i \in [m], v \in [7], \mathbf{x}_i^v = \mathbf{T}_v(\bar{\mathbf{x}}_i;\beta), \mathbf{T}_v \in \mathcal{T}$.

## D Experimental Details

In this section, the experiment details of all the baselines and our approach on both synthetic and real-world datasets are given. Moreover, the experiment details of the INPP framework are given.

We implement the model with PyTorch 1.9.0. All experiments are conducted on a NVIDIA Quadro GV100 GPU with 32 GB memory.

**Synthetic Datasets:** For **L-IIE**, **U-IIE**, **M-IIE**, **DF-MLP** and **Mv-IIE** with basic classifier $\mathrm{C}_3$, Adam (Kingma & Ba, 2015) is used as the optimization algorithm with momentum $= 0.9$, weight decay $= 0.0001$, and cross-entropy loss is used as the category label prediction loss. We set epochs equal to 200 and the mini-batch size equal to 200 for all datasets. The network structure of the basic classifier $\mathrm{C}_3$ is a two-layer network with ReLU and Dropout in all the layers ($100 \times 100 \times \#\text{classes}$). For each algorithm on each dataset, we randomly divide each dataset into a training set ($60\%$), a validation set ($20\%$) and a test set ($20\%$). First, we select the hyperparameters that can obtain the highest classification accuracy on the validation set. The hyperparameters that need to be selected are shown in Table 7. Then, the selected optimal hyperparameters are used to test the performance of each algorithm on the test set. In addition, the validation set is also used to select the candidate views of our proposed framework. We repeat the entire experiment process 10 times. Thus, the final results are shown in the form of "mean$\pm$ standard deviation". Classification accuracy is used to evaluate the performance of the proposed model. The definition of classification accuracy is shown as follows:

$$\text{Accuracy} = \frac{|\bar{\mathbf{x}} \in \bar{\mathcal{X}} : f(\bar{\mathbf{x}}) = \arg\min_{k\in[1,K]} h_k(\bar{\mathbf{x}})|}{|\bar{\mathbf{x}} \in \bar{\mathcal{X}}|},$$

where $f(\bar{\mathbf{x}})$ is the ground truth label of $\bar{\mathbf{x}}$, while $\boldsymbol{h}(\bar{\mathbf{x}}) = (h_1(\bar{\mathbf{x}}), \cdots, h_K(\bar{\mathbf{x}}))^\top$ is the label predicted by the presented algorithms and the baselines.

**Real-world Datasets:** The experiment details of the proposed method and the four baselines are basically the same as the synthetic datasets. We note that the mushroom dataset is an imbalanced dataset which means that each category contains a different number of instances. Therefore, we preprocess this dataset using a random oversampling technique (KMeansSMOTE (Last et al., 2017)) and use balanced accuracy (Brodersen et al., 2010) instead of ordinary classification accuracy to compare model performance on the mushroom dataset. The definition of balanced accuracy is shown as follows:

$$\text{Balanced Accuracy} = \tfrac{1}{K} \sum_{k=1}^{K} (\text{Recall of } k\text{-th class}),$$
$$\text{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}),$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative. After the process of the random oversampling technique, the data of each category in the mushroom dataset is expanded to 30. In addition, the Wilcoxon rank-sum test results of the method, which obtains the best performance, compared to the other methods are given on real-world datasets.

**INPP Framework:** The structure of INPP framework is shown in Figure 4. We randomly divide the original dataset (letter recognition dataset selected from the UCI Machine Learning Repository) into a raw dataset from the data owner(s) ($70\%$) and a new dataset ($30\%$) from the results' party. We choose $L = 6, T = 15$ and set $q = 0.20, 0.30, 0.50$. From Table 3, Mv-IIE with SVM-rbf (SVM with radial basis kernel function) achieve best outcomes on the second synthetic dataset. Therefore, we use SVM-rbf as the basic classifier of Mv-IIE in this experiment. The experimental details of Mv-IIE are the same as the aforementioned. The experiment details of the four well-known machine learning methods on the original dataset are the same as the experiment details of the four baselines on the synthetic datasets.

## E Details of The Two Real-world Datasets Descriptions

In this section, we briefly introduce the two real-world datasets used in the experiments.

**Mushroom Dataset :** The first dataset is extracted from https://www.mykoweb.com/CAF/, which contains 248 instances in 17 fungi species categories. There are five interval-valued variables: the pileus cap width $Pw$, the stipe length $Sl$, the stipe thickness $St$, the spores major axis length $Sma$, and the spores minor axis length $Smi$. Some instances of the mushroom dataset are shown in Table 1. The goal of our experiment on this dataset is to predict the species category of the California mushroom using five interval-valued features.

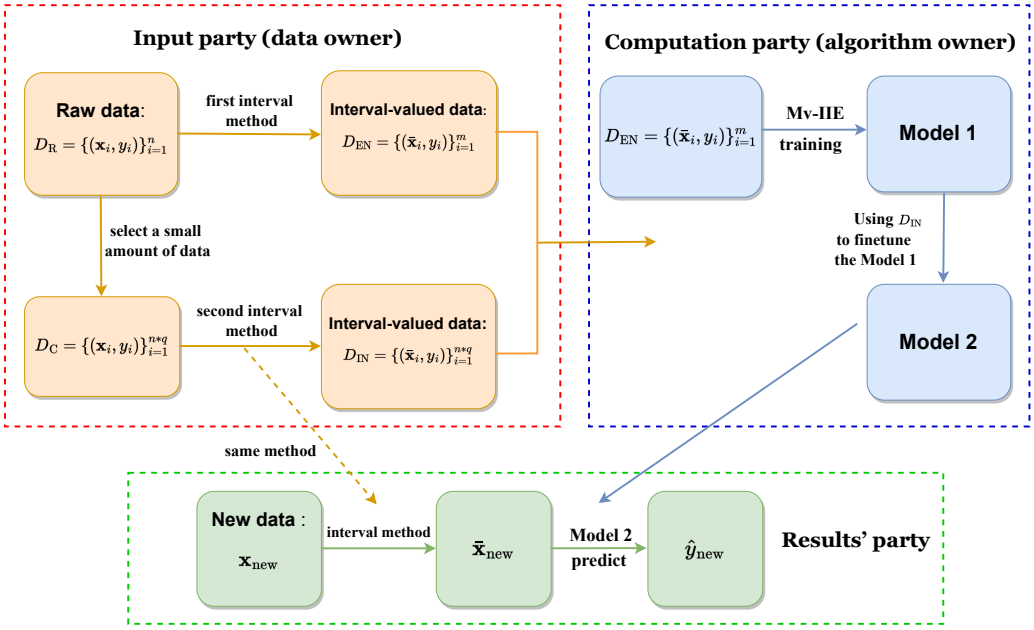



Figure 4: **INPP** framework: The input party (denoted in **orange**) applies two interval methods to transfer the raw data into two interval-valued datasets. The computation party (denoted in **blue**) uses $D_{\mathrm{EN}}$ to train Model 1 by applying Mv-IIE framework and $D_{\mathrm{IN}}$ is used to fine-tune Model 1 to obtain Model 2. The results' party (denoted in **green**) uses Model 2 for new data prediction.

Table 8: Some Instances of the Weather Dataset

| Local times | T | P0 | P | U | Td | Y |
|---|---|---|---|---|---|---|
| 31/12/2021 | [10.6, 13.3] | [757.8, 760.3] | [759.4, 762.1] | [81, 93] | [9.4, 11.1] | 1 |
| 24/12/2021 | [4.4, 12.2] | [757.3, 762.1] | [759.0, 763.6] | [40, 61] | [-5.0, 1.7] | 0 |
| 23/12/2021 | [-1.1, 5.0] | [763.4, 768.2] | [762.2, 769.9] | [38, 55] | [-10.0, 5.0] | 0 |
| 22/12/2021 | [2.8, 10.6] | [752.5, 761.6] | [754.0, 763.2] | [34, 93] | [-9.4, 2.2] | 1 |

**Weather Dataset :** The second dataset is the meteorological data of Washington (from January 1, 2016 to December 31, 2021), provided by the 'Reliable Prognosis' site (https://rp5.ru/), which contains 2191 instances. Each instance in this dataset is the meteorological data for one day in Washington, which is described by five interval-valued variables (air temperature $T$, atmospheric pressure at weather station level $P0$, atmospheric pressure reduced to main sea level $P$, humidity $U$ and dew-point temperature $Td$) and one category variable (Precipitation or not: $0 \equiv$ No Precipitation, $1 \equiv$ Precipitation). Some instances of this dataset are shown in Table 8. We aim to use the five interval-valued features for precipitation prediction.