Distribution-free Conformal Prediction for Ordinal Classification

Subhrasish Chakraborty   SC2325@NJIT.EDU
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, USA

Chhavi Tyagi   CT364@NJIT.EDU
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, USA

Haiyan Qiao   HQIAO@CSUSB.EDU
School of Computer Science & Engineering
California State University San Bernardino
San Bernardino, CA 92407, USA

Wenge Guo   WENGE.GUO@NJIT.EDU
Department of Mathematical Sciences
New Jersey Institute of Technology
Newark, NJ 07102, USA

Abstract

Conformal prediction is a general distribution-free approach for constructing prediction sets combined with any machine learning algorithm that achieve valid marginal or conditional coverage in finite samples. Ordinal classification is common in real applications where the target variable has natural ordering among the class labels. In this paper, we discuss constructing distribution-free prediction sets for such ordinal classification problems by leveraging the ideas of conformal prediction and multiple testing with FWER control. Newer conformal prediction methods are developed for constructing contiguous and non-contiguous prediction sets based on marginal and conditional (class-specific) conformal p-values, respectively. Theoretically, we prove that the proposed methods respectively achieve satisfactory levels of marginal and class-specific conditional coverages. Through simulation study and real data analysis, these proposed methods show promising performance compared to the existing conformal method.

Keywords: Conformal prediction, ordinal classification, multiple testing, FWER control, marginal coverage, class-specific conditional coverage

1. Introduction

Ordinal classification, also known as ordinal regression or ordinal prediction, is a machine learning task that involves predicting a target variable with ordered categories (McCullagh, 1980; Agresti, 2010). In ordinal classification, the target variable has a natural ordering or hierarchy among its categories, but the intervals between the categories may not be evenly spaced or defined. Unlike regular classification, where the classes are nominal and

unordered, ordinal classification takes into account the ordering relationship between the classes. This makes it suitable for situations where the outcome variable has multiple levels of severity, satisfaction ratings, or rankings. Here are a few examples of ordinal classification: customer satisfaction levels, movie ratings, disease severity levels, and education levels (Agresti, 2010).

In ordinal classification, the goal is to learn a model that can accurately predict the ordinal variable's value given a set of input features. The model needs to understand the ordering of the classes and make predictions that respect this order. In the literature, some conventional classification algorithms have been adapted or modified to address ordinal classification, for example, ordinal logistic regression, SVM, decision trees, random forest, and neural networks (Harrell, 2015; da Costa et al., 2010; Kramer et al., 2001; Janitza et al., 2016; Cheng et al., 2008). Some alternative methods are also specifically developed for ordinal classification problems by fully exploiting the ordinal structure of the response variables (Frank and Hall, 2001; Cardoso and da Costa, 2007; Gutiérrez et al., 2015). However, these existing methods can only provide point prediction, which is not adequate in some high stakes areas such as medical diagnosis and automatic driving. Uncertainty quantification (UQ) techniques aim to go beyond point predictions and provide additional information about the reliability of these predictions. There are various techniques for UQ in machine learning, including Bayesian methods, calibration, and conformal prediction (Hüllermeier and Waegeman, 2021).

Conformal prediction is a unique distribution-free UQ technique that provides a prediction set rather than a point prediction for the true response with guaranteed coverage (Vovk et al., 1999, 2005; Shafer and Vovk, 2008; Angelopoulos and Bates, 2021; Fontana et al., 2023). It can be used as a wrapper with any underlying algorithm. In this paper, we use the conformal prediction technique to construct prediction sets for ordinal classification problems. By combining the ideas of conformal prediction and multiple testing, two new conformal prediction methods are introduced for constructing contiguous and noncontiguous prediction sets. Firstly, the problem of ordinal classification is reformulated as a problem of multiple testing; Secondly, for each constructed hypothesis, the marginal and conditional conformal p-values are respectively calculated; Thirdly, based on these marginal (conditional) conformal p-values, three multiple testing procedures are developed for controlling marginal (conditional) familywise error rate (FWER); Finally, based on the testing outcomes of these procedures, the prediction sets are constructed and proved having guaranteed marginal (conditional) coverage. There are almost no works of applying conformal prediction to address ordinal classification in the literature. To our knowledge, Lu et al. (2022) is the only existing work, in which, a new (split) conformal prediction method is developed for constructing adaptive contiguous prediction region. This method is proved to have guaranteed marginal coverage, however, it cannot guarantee to have more desired conditional coverage. Moreover, it does not work well for high dimensional data. Compared to the method introduced in Lu et al. (2022), our proposed methods generally show via theoretical and numerical studies better performance in the settings of higher dimensions and in terms of class-specific conditional coverage; especially for the conditional conformal p-values based methods, they are proved to have guaranteed conditional coverage.

The rest of this paper is structured as follows. In Section 2, we briefly introduce split-conformal prediction and review related works, followed by Section 3 which presents the

development of our proposed conformal methods using the idea of multiple testing. Section 4 provides numerical studies to evaluate the performance of the proposed methods compared to the existing method. Some discussions are presented in Section 5 and all proofs are deferred to the appendix section.

2. Preliminaries

In this section, we briefly describe the conformal prediction framework and review the related literature.

2.1. Conformal Prediction

Conformal prediction is a general statistical technique to construct prediction sets for any supervised learning method. The main advantage of this approach is that it is distribution-free and can work with any underlying algorithm. Conformal prediction is broadly of two types – full conformal prediction and split-conformal prediction. The full conformal prediction uses all the observations to train the underlying algorithms (Vovk et al., 2005). In contrast, split-conformal prediction (Papadopoulos et al., 2002) involves splitting the training data into proper training data to train the underlying algorithm and calibration data to calculate the threshold for forming prediction sets. Our proposed methods are based on the split-conformal method. Consider a multi-class classification problem with feature space $\mathcal{X}$ and labels $\mathcal{Y} = \{1, 2, \cdots, K\}$. Given the training observations $(X_i, Y_i)_{i=1}^{2n}$ and a test input X_{2n+1}, the goal is to find a prediction set $C(X_{2n+1}) : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ that contains the unknown response Y_{2n+1} with enough statistical coverage.

The split-conformal procedure suggests to split $2n$ observations to n training observations, i.e., $(X_i, Y_i)_{i=1}^{n}$, which are used to train $\hat{f}$, an underlying classifier such that $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$ and the remaining n observations $(X_i, Y_i)_{i=n+1}^{2n}$ for calibration. The central part of this technique involves calculating the conformity scores for each observation, which measures how much the test observation conforms with the calibration observations. There can be several choices of conformity scores for multi-class classification problem, including posterior class probability, cumulative probability, and regularized cumulative probability (Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2020). Given the score function $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, the conformity score for the ith calibration observation is defined as $s_i = s(X_i, Y_i), \ \ i = n + 1, \cdots, 2n$.

For a test input X_{2n+1}, we compute the conformity score for each class label. Therefore, for a class label $y \in \mathcal{Y}$ the conformity score corresponding to (X_{2n+1}, y) is $s_{2n+1} = s(X_{2n+1}, y)$. By using the conformity scores obtained for the calibration observations and the test input coupled with a given label y, we can calculate the conformal p-value to test whether the unknown true label Y_{2n+1} corresponding to the test input X_{2n+1} is y or not. The (marginal) conformal p-value is defined as,

$$p(X_{2n+1}, y) = \frac{\sum\limits_{i=n+1}^{2n} \mathbb{I}\left(s_{2n+1} \leq s_i\right) + 1}{n + 1}. \tag{1}$$

The final step involves constructing the prediction set $C(X_{2n+1}) = \{y : p(X_{2n+1}, y) \geq \alpha\}$, which satisfies

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})) \geq 1 - \alpha, \tag{2}$$

when the calibration and test observations $(X_i, Y_i)_{i=n+1}^{2n+1}$ are exchangeably distributed, where $\alpha \in (0, 1)$ is a pre-specified mis-coverage level. Equation (2) is called *marginal validity* of the prediction set $C(X_{2n+1})$. It guarantees that the true label Y_{2n+1} is contained in the prediction set with $100(1 - \alpha)\%$ confidence. Vovk (2012) introduced another type of conformal p-value which is called as the conditional conformal p-value. Let $\mathcal{D}_{cal} = \{n + 1, \ldots, 2n\}$ denote the indices of the calibration observations $(X_i, Y_i)_{i=n+1}^{2n}$. For a test input X_{2n+1} and any class $y = 1, \ldots, K$, the (class-specific) conditional conformal p-value given $Y_{2n+1} = y$ is defined as

$$p(X_{2n+1}|y) = \frac{\sum_{i \in \mathcal{I}_y} \mathbb{I}\{s_i \leq s_{2n+1}\} + 1}{n_y + 1}, \tag{3}$$

where $\mathcal{I}_y = \{i \in \mathcal{D}_{cal} : Y_i = y\}$, $n_y = |\mathcal{I}_y|$ is the size of $\mathcal{I}_y$, and $s_i = s(X_i, y)$ for $i = n + 1, \ldots, 2n + 1$.

In general, the concept of conditional coverage such as *object conditional validity* and *class-specific conditional validity* are more relevant to practical applications (Vovk, 2012; Lei, 2014; Barber et al., 2021b). If conditional coverage is guaranteed, our predictions are valid for a more specific sub-population, not merely on average over a general population. Specifically, in classification problems, *class-specific conditional validity* provides conditional coverage for each given class, which is defined as

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})|Y_{2n+1} = y) \geq 1 - \alpha \tag{4}$$

for any $y = 1, \ldots, K$. Proposition 1 and 2 below ensure that the marginal and conditional conformal p-values are valid, which result in desired marginal and (class-specific) conditional coverage.

To simplify the notation, we let $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, 2n + 1$ and denote the conformity scores of the calibration data, $\{Z_i\}_{i=n+1}^{2n}$, as s_i's, and the conformity score of the test data, $Z_{2n+1} = (X_{2n+1}, Y_{2n+1})$, as s_{2n+1}, where Y_{2n+1} is unknown. These notations are used in all propositions and theorems presented in this paper.

Proposition 1 *Suppose that $\{Z_i\}_{i=1}^{2n+1}$ where $Z_i = (X_i, Y_i)$ are exchangeable random variables, then the marginal conformal p-values defined below as,*

$$p(Z_{2n+1}) = \frac{\sum_{i=n+1}^{2n} \mathbb{I}(s_i \leq s_{2n+1}) + 1}{n + 1} \tag{5}$$

is valid in the sense that for any $t \in [0, 1]$, we have

$$\mathbb{P}(p(Z_{2n+1}) \leq t) \leq t.$$

Moreover, if the conformity scores $\{s_i\}_{i=n+1}^{2n+1}$ are distinct surely, we have,

$$p(Z_{2n+1}) \sim U\left\{\frac{1}{n+1}, \cdots, 1\right\}$$

Proposition 2 *Suppose that* $\{Z_i\}_{i=1}^{2n+1}$ *where* $Z_i = (X_i, Y_i)$ *are exchangeable random variables, then for any* $y \in \mathcal{Y}$, *given* $\mathcal{I}_y \subseteq \mathcal{D}_{cal}$ *and* $Y_{2n+1} = y$, *the corresponding conditional conformal p-value as defined in equation (3), is conditionally valid in the sense that for any* $t \in [0, 1]$,

$$\mathbb{P}\left(p(X_{2n+1}|y) \leq t \mid \mathcal{I}_y, Y_{2n+1} = y\right) \leq t.$$

Moreover, if $\{s_i\}_{i \in \mathcal{I}_y \cup \{2n+1\}}$ *are distinct surely, we have that conditional on* $\mathcal{I}_y$ *and* $Y_{2n+1} = y$,

$$p(X_{2n+1}|y) \sim U\left\{\frac{1}{n_y + 1}, \cdots, 1\right\}.$$

2.2. Related work

The framework of Conformal prediction was introduced by Vladimir Vovk and his collaborators Vovk et al. (1999, 2005) and has found many applications in classification problems. Shafer and Vovk (2008) and Angelopoulos and Bates (2021) provided a tutorial introduction and brief literature review on this field. Several conformal methods have been developed to address binary classification (Lei, 2014) and multi-class classification problems (Hechtlinger et al., 2018; Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2020; Tyagi and Guo, 2023). (Marginal) Coverage guarantees of all these methods are established under the assumption of exchangeability. Recently conformal prediction methods have been extended to the settings of non-exchangeability (Tibshirani et al., 2019; Cauchois et al., 2022; Podkopaev and Ramdas, 2021; Gibbs and Candes, 2021; Barber et al., 2022), and have provided stronger conditional coverage guarantee (Vovk et al., 2003; Vovk, 2012; Lei and Wasserman, 2014; Barber et al., 2021a).

Although various conformal prediction methods have been developed for conventional classification problems, however, to our knowledge, Lu et al. (2022) is the only reference that is specifically devoted to address ordinal classification problems using conformal prediction methods, in which an adaptive conformal method is developed for constructing contiguous prediction sets for ordinal response and is applied to AI disease rating in medical imaging. In addition, Xu et al. (2023) is a closely related reference in which newer methods are developed for two types of loss functions specially designed for ordinal classification in the more general framework of conformal risk control.

3. Method

In this section, we introduce several new conformal prediction methods for ordinal classification problems, in which there is a natural ordering among the classes labels. For simplicity, we assume a descending order of priority from class 1 to K in the response space $\mathcal{Y} = \{1, 2, \cdots, K\}$.

3.1. Problem Formulation

We formulate the ordinal classification problem as a multiple testing problem. Specifically, by using the One-vs-All (OVA) strategy (Rifkin and Klautau, 2004), for each class label,

we construct a hypothesis to test whether or not a given test input X_{2n+1} belongs to the particular class. The construction of the hypothesis is described as follows,

$$H_i : Y_{2n+1} = i \quad \text{vs} \quad H_i' : Y_{2n+1} \neq i, \tag{6}$$

for $i = 1, \cdots, K$. It is easy to see that all these hypotheses are random and there exists only one true null hypothesis, as each data point pertains uniquely to a single class label. To test each individual hypothesis H_i, we use the corresponding marginal conformal p-value $p(X_{2n+1}, i)$; to test $H_1, \ldots, H_K$ simultaneously, we consider the following three p-value-based testing procedures:

- **Procedure 1** : Test $H_1, H_2, \cdots, H_K$ sequentially. The test is performed as follows.

  - If $p(X_{2n+1}, 1) \leq \alpha$, reject H_1, move to test H_2 else stop testing;
  - For $i = 2, \cdots, K - 1$, if $p(X_{2n+1}, i) \leq \alpha$, reject H_i, move to test H_{i+1} else stop testing;
  - If $p(X_{2n+1}, K) \leq \alpha$, reject H_K else stop testing.

- **Procedure 2** : Test $H_K, H_{K-1}, \cdots, H_1$ sequentially. The test is performed as follows.

  - If $p(X_{2n+1}, K) \leq \alpha$, reject H_K, move to test H_{K-1} else stop testing;
  - For $i = K - 1, \cdots, 2$, if $p(X_{2n+1}, i) \leq \alpha$, reject H_i, move to test H_{i-1} else stop testing;
  - If $p(X_{2n+1}, 1) \leq \alpha$, reject H_1 else stop testing.

- **Procedure 3** : Single-step procedure with common critical value α. This procedure rejects any hypothesis H_i if and only if $p(X_{2n+1}, i) \leq \alpha$.

Procedure 1 and 2 are two pre-ordered testing procedures for which Procedure 1 follows the same testing order as that of the K classes whereas Procedure 2 uses the reverse order of these classes (Dmitrienko et al., 2009). Procedure 3 is actually a conventional Bonferroni procedure for a single true null. Since there is only one true null among the K tested hypotheses, by Proposition 1, we have that all these three (marginal) conformal p-value based procedures strongly control family-wise error rate (FWER), the probability of making at least one type I error, at a pre-specified level α (Dmitrienko et al., 2009). For each Procedure $i, i = 1, 2, 3$ defined above, the index set A_i of the accepted hypotheses is described as follows,

1. $A_1 = \{y_{min}, y_{min} + 1, \cdots, K\}$, where $y_{min} = \min\{y \in \mathcal{Y} : p(X_{2n+1}, y) > \alpha\}$;

2. $A_2 = \{1, 2, \cdots, y_{max}\}$, where $y_{max} = \max\{y \in \mathcal{Y} : p(X_{2n+1}, y) > \alpha\}$;

3. $A_3 = \{y : p(X_{2n+1}, y) > \alpha\}$.

3.2. Ordinal Prediction Interval

Based on the the acceptance sets A_1 and A_2 of Procedure 1 and 2 given as above, we can obtain a new acceptance region $A_{12} = A_1 \cap A_2 = \{y_{min}, \ldots, y_{max}\}$, which is used to define the prediction region $C(X_{2n+1})$ for the unknown response Y_{2n+1}. Specifically, the prediction region $C(X_{2n+1})$ consists of the class labels for which the corresponding hypotheses are both accepted by Procedure 1 and 2, resulting in a *contiguous set* of labels $\{y_{min}, \ldots, y_{max}\}$. This prediction region is referred to as a *prediction interval* in this context. The procedure for constructing the prediction interval $C(X_{2n+1})$ is summarized in Algorithm 1 and illustrated in Figure 1.

Remark 3 *The contiguous prediction interval A_{12} produced by Algorithm 1, based on the combination of Procedures 1 and 2, includes A_3 of Procedure 3. Moreover, to determine the contiguous interval A_{12}, we only need to identify the minimal and maximal elements of A_3. While this approach may seem mathematically simpler, our multiple testing approach offers certain advantages. By leveraging various ideas and tools from multiple testing, we can develop new CP methods for ordinal classification more easily and systematically.*

Algorithm 1: Ordinal Prediction Interval

Input: training set $\mathcal{D}_{train} = (X_i, Y_i)_{i=1}^n$, calibration set $\mathcal{D}_{cal} = (X_i, Y_i)_{i=n+1}^{2n}$, test input X_{2n+1}, underlying algorithm $\mathcal{A}$, conformity score function s, mis-coverage level α.

Output: Prediction interval, $C(X_{2n+1})$.

Train a classifier $\mathcal{A}$ on $\mathcal{D}_{train}$;

for $(X_i, Y_i) \in \mathcal{D}_{cal}$ **do**
| Compute conformity score $s_i = s(X_i, Y_i)$;
end

For each $y \in \mathcal{Y}$, compute conformity score $s_{2n+1} = s(X_{2n+1}, y)$ and corresponding conformal p-value $p(X_{2n+1}, y)$ using equation (1);

$y_{min} = \min\{y \in \mathcal{Y} : p(X_{2n+1}, y) > \alpha\}$;

$y_{max} = \max\{y \in \mathcal{Y} : p(X_{2n+1}, y) > \alpha\}$;

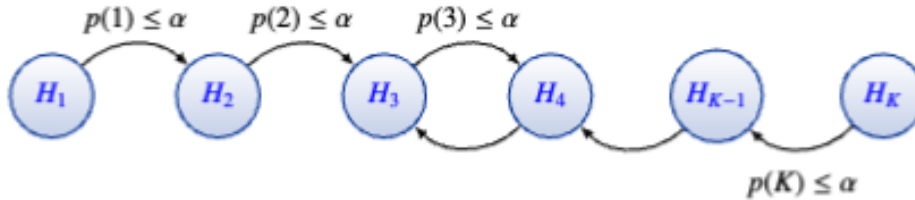Prediction interval, $C(X_{2n+1}) = \{y_{min}, \ldots, y_{max}\}.\}$

Figure 1: Graphical representation of testing K pre-ordered hypotheses using Procedures 1 and 2, which indirectly retrieve the prediction interval produced by the Ordinal Prediction Interval (OPI) method. Here, $p(i)$ represents the conformal p-values.

3.3. Ordinal Prediction Set

Our second method for constructing ordinal prediction regions is based on Procedure 3. In this method, the prediction region $C(X_{2n+1})$ is defined simply using the acceptance region A_3 of Procedure 3, that is, $C(X_{2n+1}) = \{y \in \mathcal{Y} : p(X_{2n+1}, y) > \alpha\}$. Specifically, the prediction region consists of any class labels for which the corresponding hypotheses are not rejected by Procedure 3, resulting in a *non-contiguous set* of labels. This prediction region is referred to as a *prediction set* in this context. The procedure for constructing the prediction set is detailed in Algorithm 2 below.

Algorithm 2: Ordinal Prediction Set

Input: training set $\mathcal{D}_{train} = (X_i, Y_i)_{i=1}^{n}$, calibration set $\mathcal{D}_{cal} = (X_i, Y_i)_{i=n+1}^{2n}$, test observation X_{2n+1}, underlying algorithm $\mathcal{A}$, conformity score function s, mis-coverage level α.

Output: Prediction set, $C(X_{2n+1})$.

Train a classifier $\mathcal{A}$ on $\mathcal{D}_{train}$;

for $(X_i, Y_i) \in \mathcal{D}_{cal}$ **do**
| Compute conformity score $s_i = s(X_i, Y_i)$;
end

For each $y \in \mathcal{Y}$, compute conformity score $s_{2n+1} = s(X_{2n+1}, y)$ and corresponding conformal p-value $p(X_{2n+1}, y)$ using equation (1);

Prediction Set, $C(X_{2n+1}) = \{y : p(X_{2n+1}, y) > \alpha\}$.

In the following, we present two results regarding the FWER control of Procedure 1-3 and (marginal) coverage guarantees of Algorithm 1-2 introduced as above.

Proposition 4 *Suppose that $(X_i, Y_i)_{i=n+1}^{2n+1}$ are exchangeable random variables, then Procedure 1-3 based on marginal conformal p-values, all strongly control the FWER at level α, i.e., $FWER \leq \alpha$. Specifically, if the conformity scores $\{s_i\}_{i=n+1}^{2n+1}$ are distinct surely, then for Procedure 3, we also have, $FWER \geq \alpha - \frac{1}{n+1}$.*

Theorem 5 *Suppose that $(X_i, Y_i)_{i=n+1}^{2n+1}$ are exchangeable random variables, then the prediction region $C(X_{2n+1})$ determined by Algorithm 1 and 2 both satisfy*

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})) \geq 1 - \alpha.$$

Specifically, for $C(X_{2n+1})$ determined by Algorithm 2, if the conformity scores $\{s_i\}_{i=n+1}^{2n+1}$ are distinct surely, we have

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})) \leq 1 - \alpha + \frac{1}{n+1}.$$

3.4. Class-specific conditional coverage

To achieve more desired class-specific conditional coverage for our constructed prediction intervals and prediction sets, in Procedure 1-3 we use (class-specific) conditional conformal p-values $p(X_{2n+1}|y)$ as described in equation (3), instead of the marginal conformal p-values $p(X_{2n+1}, y)$ for simultaneously testing $H_1, \ldots, H_K$ formulated in equation (6). It is shown in Proposition 4 below that all these three modified procedures strongly control the conditional familywise error rate (FWER) at level α, i.e., $\text{FWER}_y = \mathbb{P}(V > 0|Y_{2n+1} = y) \leq \alpha$, for any $y \in \mathcal{Y}$, where V is the number of type 1 errors. This result in turn leads to that the prediction regions $C(X_{n+1}|y)$ constructed by Algorithm 1 and 2 based on $p(X_{2n+1}|y)$ satisfy more desired (class-specific) conditional coverage, as stated in Theorem 2.

Proposition 6 *Under the same exchangeability assumption as in Proposition 2, Procedure 1-3 based on conditional conformal p-values $p(X_{2n+1}|y)$ all strongly control the conditional FWER at level α, i.e., for any $y \in \mathcal{Y}$,*

$$FWER_y = \mathbb{P}\{\text{reject } H_y|Y_{2n+1} = y\} \leq \alpha.$$

Specifically, if the conformity scores $\{s_i\}_{i \in \mathcal{I}_y \cup \{2n+1\}}$ are distinct surely, then for Procedure 3 based on $p(X_{2n+1}|y)$, we have that for any $y \in \mathcal{Y}$ and $\mathcal{I}_y \subseteq \mathcal{D}_{cal}$.

$$\mathbb{P}(\text{reject } H_y|Y_{2n+1} = y, \mathcal{I}_y) \geq \alpha - \frac{1}{n_y + 1}.$$

Theorem 7 *Under the same exchangeability assumption as in Theorem 1, the prediction region $C(X_{2n+1}|y)$ determined by Algorithm 1 or 2 based on conditional conformal p-values $p(X_{2n+1}|y)$ satisfies*

$$\mathbb{P}\left(Y_{2n+1} \in C(X_{2n+1}|y) \,\big|\, Y_{2n+1} = y\right) \geq 1 - \alpha$$

for any $y \in \mathcal{Y}$. Specifically, for the prediction set $C(X_{n+1}|y)$ determined by Algorithm 2 based on $p(X_{2n+1}|y)$, if the conformity scores $\{s_i\}_{i \in \mathcal{I}_y \cup \{2n+1\}}$ are distinct surely, we have

$$\mathbb{P}\left(Y_{2n+1} \in C(X_{2n+1}|y) \,\big|\, Y_{n+1} = y, \mathcal{I}_y\right) \leq 1 - \alpha + \frac{1}{n_y + 1}$$

for any $y \in \mathcal{Y}$ and $\mathcal{I}_y \subseteq \mathcal{D}_{cal}$.

4. Numerical Study

In this section, we evaluate the performance of our four proposed methods, Ordinal Prediction Interval (OPI) in Algorithm 1 based on marginal conformal p-values (marginal OPI), the OPI based on conditional conformal p-values (conditional OPI), Ordinal Prediction Set (OPS) in Algorithm 2 based on marginal conformal p-values (marginal OPS), and the OPS based on conditional conformal p-values (conditional OPS), in comparison with the existing counterpart developed in Lu et al. (2022), Ordinal Adaptive Prediction Set (OAPS), on simulated data and one real dataset. The comparison is based on the marginal coverage, average set size, and class-specific conditional coverage of the prediction regions for a pre-specified level α. The empirical metric we use to measure the class-specific conditional coverage (CCV) of the above methods is defined as

$$CCV = \max_{y \in \{1,2,\cdots,K\}} \{(1-\alpha) - P_y, 0\},$$

where P_y is the estimate of $\mathbb{P}(Y \in C(X_{2n+1}|y) \,|Y = y)$ and $C(X_{2n+1}|y)$ is the prediction region obtained from Algorithm 1 or 2 using conditional conformal p-values for any $y \in \mathcal{Y}$. Intuitively, the metric measures the maximum of the deviance of the conditional coverage for each of the classes from the desired level of conditional coverage $1 - \alpha$.

In the whole numerical investigations including simulation studies and real data analysis, we use the logistic regression algorithm as the underlying algorithm for our experiments and compute the conformity scores as estimated posterior probabilities of classes.

4.1. Simulations

We present the simulation study to evaluate the performance of our proposed methods along with the existing method. We consider two simulation settings below, a Gaussian mixture model and a sparse model.

1. **Gaussian mixture.** $(X|Y = k) \sim \pi_1 N_d(\mu_k, \Sigma) + \pi_2 N_d(\mu_{k+1}, \Sigma)$ for $k = 1, 2, 3$ and $(X|Y = 4) \sim \pi_1 N(\mu_4; \Sigma) + \pi_2 N(\mu_1; \Sigma)$ with $\pi_1 = 0.2, \pi_2 = 0.8$.

   In the above setting, we set $d = 2$, $\mu_1 = (-1, 0)$, $\mu_2 = (-1, -1)$, $\mu_3 = (0, -1)$, $\mu_4 = (1, -1)$, and Σ as the equal correlation matrix with correlation $= 0.1$. Figure 2 provides an illustration of a sample of synthetic data generated under the simulation setting.

2. **Sparse model**. The sparse model is generated with different dimensions of feature vector with $d = 5, 10, 20, 50$, or 100. The features are generated with $X_i \sim N(0, 1)$ with $Cov(X_i, X_j) = 0.5$ for $i \neq j$. The class labels are generated using the sigmoid function $f(x)$ and the the following decision rule,

   $$Y_i = k \quad if \quad \frac{k-1}{4} \leq f(x_i) < \frac{k}{4}, \ where \ k = 1, 2, 3, 4,$$

   where $f(x) = 1/(1 + e^{-\beta' x})$ with $\beta = (\beta_1, \cdots, \beta_d)'$ and $x = (x_1, \ldots, x_d)'$. The value of β is set as $(\beta_1, \ldots, \beta_5) = (1, 1, 1, -\sqrt{2}, 1)$, and $\beta_i = 0$ for any $5 < i \leq d$.
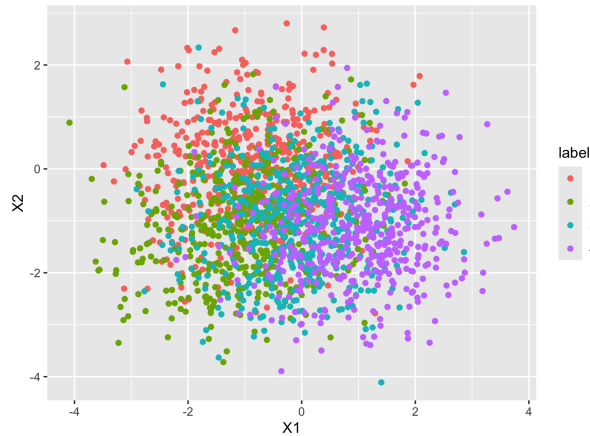
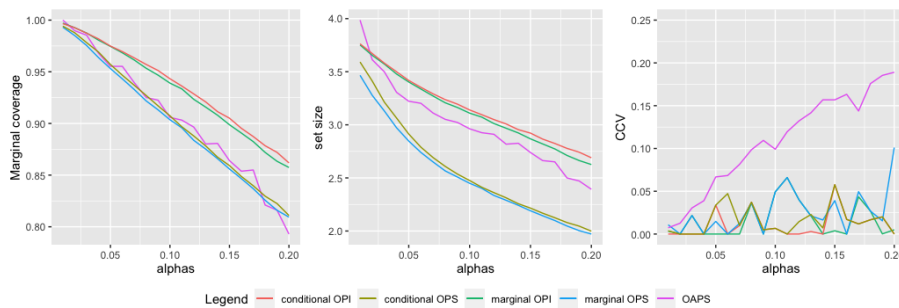Figure 2: The figure represents the bivariate synthetic data corresponding to simulation setting 1.

Figure 3: Performance comparison of four proposed methods OPS and OPI using marginal and conditional conformal p-values with the existing Ordinal APS under simulation setting 1 in terms of marginal coverage (left), average set size (middle) and class-specific conditional coverage (right) of the prediction sets.

The sample size for these two simulation settings is 2,000, out of which 500 samples have been used to train the classifier, 525 observations for calibration, and 975 for validation. The simulations are repeated 500 times, and the results are averaged to obtain the final performance metrics.

Figure 3 displays the performance of our proposed methods along with the existing method under simulation setting 1. It can be seen from the left panel of Figure 3 that all these five methods empirically achieve the desired level of marginal coverage. The middle panel of Figure 3 compares the set sizes of the prediction regions corresponding to these five methods. It can be seen from the figure that the marginal OPS and the conditional OPS use shorter set sizes to attain the proper marginal coverage than the existing OAPS whereas the OAPS has shorter set sizes than the marginal OPI and the conditional OPI. Finally, while the marginal coverage is guaranteed by all the methods, the right panel of
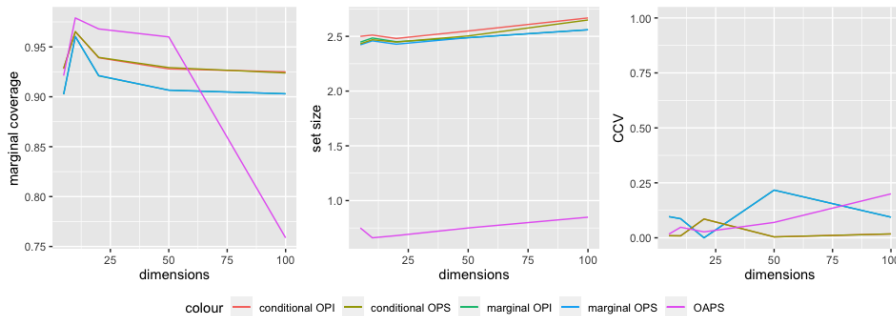
Figure 4: Performance comparison of four proposed methods with existing Ordinal APS under simulation setting 2 with different dimensions of inputs and fixed mis-coverage level $\alpha = 0.1$ in terms of marginal coverage (left), average set size (middle) and CCV (right) of the prediction sets.

Figure 3 shows their differences in class-specific conditional coverage; the existing OAPS exhibits the largest value of CCV compared to the proposed methods and among the four proposed methods, the conditional OPI and conditional OPS exhibit lower values of CCV than the marginal OPI and marginal OPS.

Figure 4 shows the performance of our proposed methods along with the existing method under simulation setting 2 with different dimensions of inputs. It can be seen from the left panel of this figure that all these methods empirically achieve desired marginal coverage for lower dimensions, however, the existing OAPS massively undercovers for higher dimensions and thus loses the control of mis-coverage rate. From the middle and right panels of Figure 4, we can also see that the conditional OPI and conditional OPS achieve lower values of CCV than the existing OAPS, although the OAPS has the lower set sizes than our proposed methods for various dimensions of input.

4.2. Application to real data

We also evaluate the performance of our proposed methods on a real dataset, Traffic accident data, which is publicly available on the website of Leeds City Council[1]. The real data consists of 1,908 traffic accidents that occurred in the year 2019. The objective is to predict the severity of the casualties, which are classified into three categories – mild, serious, and fatal based on the features available. In the numerical experiment, 500 observations are used to train the logistic regression model, 35% of the remaining observations are used for calibration, and 65% for validation. Figure 5 shows that all these methods empirically achieve desired marginal coverage for different levels of mis-coverage, however, the proposed marginal OPI and marginal OPS, and existing OAPS have lower set sizes than the conditional OPI and conditional OPS. It is also evident from Figure 5 that the proposed conditional OPS and conditional OPI both attain desired class-specific conditional coverage

---

1. The traffic accident dataset is taken from Leeds Council and its web link url is https://www.leeds.gov.uk/parking-roads-and-travel/connecting-leeds-and-transforming-travel/road-safety/road-traffic-collision-statistics.
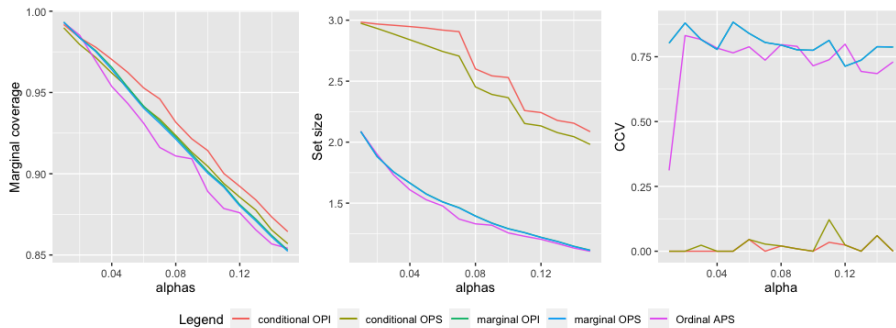
Figure 5: Performance comparison of four proposed methods OPS and OPI using marginal and conditional conformal p-values with the existing Ordinal APS in terms of marginal coverage (left), average set size (middle) and CCV (right) of the prediction regions for Traffic Accident Data.

unlike the existing method, OAPS, which seems to largely deviate from the desired level of conditional coverage.

5. Concluding Remarks

In this paper, we discussed the ordinal classification problem in the framework of conformal prediction and introduced two types of conformal methods, OPI and OPS, for constructing distribution-free *contiguous* prediction regions and *non-contiguous* prediction sets, respectively. These methods are developed by leveraging the idea of multiple testing with the FWER control and are specifically designed based on marginal conformal p-values and (class-specific) conditional conformal p-values, respectively. Theoretically, it was proved that the proposed methods based on marginal and conditional p-values respectively achieve satisfactory levels of marginal and class-specific conditional coverages. Through some numerical investigations including simulations and real data analysis, our proposed methods show promising results for the settings of higher dimensions and for class-specific conditional coverage.

This paper discussed constructing valid prediction set for single test input. It would be interesting to discuss how to construct (simultaneous) prediction sets for multiple test inputs with some overall error control such as false discovery rate (FDR) control. Another interesting extension might be to relax the conventional distributional assumption we used for classification problems, for which the training data and the test data follow from the same distribution. It will be interesting to see whether the proposed methods can be extended to the settings of distribution shift where the training and test data sets have different distributions.

References

Alan Agresti. *Analysis of ordinal categorical data.* John Wiley & Sons, 2010.

A.N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

R. Barber, E. Candès, A. Ramdas, and R. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, 10(2):455–482, 2021a.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021b.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Jaime Cardoso and Joaquim Pinto da Costa. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8:1393–1429, 2007.

M. Cauchois, S. Gupta, A. Ali, and J. Duchi. Predictive inference with weak supervision. *submitted arXiv:2201.08315*, 2022.

Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks*, pages 1279–1284. IEEE, 2008.

Joaquim F Pinto da Costa, Ricardo Sousa, and Jaime S Cardoso. An all-at-once unimodal svm approach for ordinal classification. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 59–64. IEEE, 2010.

Alex Dmitrienko, Ajit C Tamhane, and Frank Bretz. *Multiple testing problems in pharmaceutical statistics*. CRC press, 2009.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *12th European Conference on Machine Learning, Proceedings 12*, pages 145–156. Springer, 2001.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1): 127–146, 2015.

Frank E Harrell. Ordinal logistic regression. In *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, pages 311–325. Springer, 2015.

Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2018.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Silke Janitza, Gerhard Tutz, and Anne-Laure Boulesteix. Random forest for ordinal responses: prediction and variable selection. *Computational Statistics & Data Analysis*, 96: 57–73, 2016.

Stefan Kramer, Gerhard Widmer, Bernhard Pfahringer, and Michael De Groeve. Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*, 47(1-2):1–13, 2001.

J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96, 2014.

C. Lu, A.N. Angelopoulos, and S. Pomerantz. Improving trustworthiness of ai disease rating in medical imaging with ordinal conformal prediction sets. *arXiv preprint arXiv:2207.02238*, 2022.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society B*, 42:109–142, 1980.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *13th European Conference on Machine Learning, Proceedings 13*, pages 345–356. Springer, 2002.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

Y. Romano, M. Sesia, and E. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591, 2020.

M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of American Statistical Association*, 114:223–234, 2019.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Chhavi Tyagi and Wenge Guo. Multi-label classification under uncertainty: a tree-based conformal prediction approach. In *Conformal and Probabilistic Prediction with Applications*, pages 488–512. PMLR, 2023.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world.* Springer, 2005.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Sixteenth International Conference on Machine Learning*, pages 444–453, 1999.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Conformal risk control for ordinal classification. In *Uncertainty in Artificial Intelligence*, pages 2346–2355. PMLR, 2023.

Appendix A. Mathematical Proofs

A.1. Proof of Proposition 1

Proof Suppose, for any given values of conformity scores, $v_1, \cdots, v_{n+1}$, they can be rearranged as $\tilde{v}_1 < \cdots < \tilde{v}_k$ with repetitions n_i of $\tilde{v}_i$ such that $\sum_{i=1}^k n_i = n+1$. Let E_v denote the event of $\{s_{n+1}, \cdots, s_{2n+1}\} = \{v_1, \cdots, v_{n+1}\}$. Then, under E_v, for $i = 1, \cdots, k$, we have

$$\mathbb{P}(s_{2n+1} = \tilde{v}_i | E_v) = \frac{n_i}{n+1},$$

due to the exchangeability of s_i's,

We also note that under E_v and $s_{2n+1} = \tilde{v}_i$ we have from equation (5),

$$p(Z_{2n+1}) = \frac{\sum_{l=1}^i n_l}{n+1}. \tag{7}$$

Then, for any $t \in [0,1]$ and $i = 1, \cdots, k$, we have

$$\mathbb{P}(p(Z_{2n+1}) \leq t | E_v, s_{2n+1} = \tilde{v}_i) = \begin{cases} 0 & \text{if } t < \frac{\sum_{l=1}^i n_l}{n+1}, \\ 1 & \text{o.w.} \end{cases} \tag{8}$$

Thus, for any $i = 1, \cdots, k$ and $\frac{\sum_{l=1}^{i-1} n_l}{n+1} \leq t < \frac{\sum_{l=1}^i n_l}{n+1}$, we have

$$\mathbb{P}(p(Z_{2n+1}) \leq t | E_v)$$
$$= \sum_{l=1}^k \mathbb{P}(p(Z_{2n+1}) \leq t | E_v, s_{2n+1} = \tilde{v}_l) \cdot \mathbb{P}(s_{2n+1} = \tilde{v}_l | E_v)$$
$$= \frac{\sum_{l=1}^{i-1} n_l}{n+1} \leq t.$$

By taking the expectation on the above inequality, it follows that the conformal p-value $p(Z_{2n+1})$ is marginally valid.

Specifically, if conformity scores $\{s_i\}_{i=n+1}^{2n+1}$ are distinct surely, then $k = n+1$ and $n_i = 1$ for $i = 1, \ldots, n+1$. Thus,

$$\mathbb{P}\left(p(Z_{2n+1}) \leq t | E_v\right) = \frac{i-1}{n+1}, \text{ if } \frac{i-1}{n+1} \leq t < \frac{i}{n+1},$$

that is,

$$p(Z_{2n+1})| E_v \sim U\left\{\frac{1}{n+1}, \cdots, 1\right\}.$$

This completes the proof. ∎

A.2. Proof of Proposition 2

Proof For any given $y \in \mathcal{Y}$, the corresponding (class-specific) conditional conformal p-value is given by

$$p(X_{2n+1}|y) = \frac{1}{n_y + 1} \left[\sum_{i \in \mathcal{I}_y} \mathbb{I}\{s_i \leq s_{2n+1}\} + 1 \right], \tag{9}$$

where $\mathcal{I}_y = \{i : (X_i, Y_i) \in \mathcal{D}_{cal}, Y_i = y\}$, $n_y = |\mathcal{I}_y|$, $s_i = s(X_i, y)$ for $i \in \mathcal{I}_y$, and $s_{2n+1} = s(X_{2n+1}, y)$. Given I_y and $Y_{2n+1} = y$, $(X_i)_{i \in \mathcal{I}_y} \cup \{X_{2n+1}\}$ are exchangeably distributed, which is due to the assumption that $(X_i, Y_i)_{i=1}^{2n+1}$ are exchangeably distributed. Using the similar arguments as in the proof of Proposition 1, for any given values of $v_1, \cdots, v_{n_y+1}$, suppose that they can be arranged as $\tilde{v}_1 < \cdots < \tilde{v}_l$ with repetitions m_i of $\tilde{v}_i$ such that $\sum_{i=1}^{l} m_i = n_y + 1$.

Let E_v denote the event $\{s_i\}_{i \in \mathcal{I}_y} \cup \{s_{2n+1}\} = \{v_1, \cdots, v_{n_y+1}\}$. Then, given E_v, I_y, and $Y_{2n+1} = y$, we have

$$\mathbb{P}(s_{2n+1} = \tilde{v}_i | E_v, I_y, Y_{2n+1} = y) = \frac{m_i}{n_y + 1}$$

for $i = 1, \cdots, l$ and $y = 1, \cdots, K$, due to exchangeability of s_i, $i \in I_y \cup \{2n+1\}$ given I_y, which in turn is due to exchangeability of $(X_i, Y_i)_{i=1}^{2n+1}$. Note that given E_v, I_y, $Y_{2n+1} = y$, and $s_{2n+1} = \tilde{v}_i$, we have from equation (9),

$$p(X_{2n+1}|y) = \frac{\sum_{j=1}^{i} m_j}{n_y + 1}.$$

Thus, for any $t \in [0, 1]$ and $i = 1, \cdots, l$,

$$\mathbb{P}\left(p(X_{2n+1}|y) \leq t \big| E_v, I_y, Y_{2n+1} = y, s_{2n+1} = \tilde{v}_i\right) = \begin{cases} 0 & \text{if } t < \frac{\sum_{j=1}^{i} m_j}{n_y+1}, \\ 1 & \text{o.w.} \end{cases} \tag{10}$$

Then, for any given $i = 1, \cdots, l$ and $\frac{\sum_{j=1}^{i} m_j}{n_y+1} \leq t < \frac{\sum_{j=1}^{i+1} m_j}{n_y+1}$, we have

$$\mathbb{P}\left(p(X_{2n+1}|y) \leq t \big| E_v, I_y, Y_{2n+1} = y\right)$$

$$= \sum_{j=1}^{l} \mathbb{P}\left(p(X_{2n+1}|y) \leq t \big| E_v, I_y, Y_{2n+1} = y, s_{2n+1} = \tilde{v}_i\right) \cdot \mathbb{P}\left(s_{2n+1} = \tilde{v}_i \big| E_v, I_y, Y_{2n+1} = y\right)$$

$$= \frac{\sum_{j=1}^{i} m_j}{n_y + 1} \leq t.$$

By taking expectation, it follows that $p(X_{2n+1}|y)$ is conditionally valid given $Y_{2n+1} = y$. ∎

A.3. Proof of Proposition 3

Proof Consider Procedure 1-3 based on marginal conformal p-values. Note that among the tested hypotheses $H_1, \ldots, H_K$, there is exactly one hypothesis $H_{Y_{2n+1}}$ to be true. Thus, the FWER of Procedure 1-3 are all equal to

$$\mathbb{P}(\text{reject } H_{Y_{2n+1}}) \leq \mathbb{P}\left(p(X_{2n+1}, Y_{2n+1}) \leq \alpha\right) \leq \alpha,$$

where the last inequality follows by Proposition 1.

Specifically, for Procedure 3, if the conformity scores $\{s_i\}_{i=n+1}^{2n+1}$ are distinct surely, by Proposition 1, we have

$$\text{FWER} = \mathbb{P}(\text{reject } H_{Y_{2n+1}}) = \mathbb{P}\Big\{p(X_{2n+1}, Y_{2n+1}) \leq \alpha\Big\} \geq \alpha - \frac{1}{n+1},$$

the desired result. ∎

A.4. Proof of Theorem 1

Proof   Note that the prediction set derived from Algorithm 1 is given by $C(X_{2n+1}) = A_1 \cap A_2$. Thus, by Proposition 1,

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})) \geq \mathbb{P}(p(X_{2n+1}), Y_{2n+1}) > \alpha) \geq 1 - \alpha.$$

Similarly, for Algorithm 2, its prediction set is given by $C(X_{2n+1}) = \{y \in \mathcal{Y} : p(X_{2n+1}, y) > \alpha\}$. By Proposition 1, it is easy to check that

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})) = \mathbb{P}(p(X_{2n+1}, Y_{2n+1}) > \alpha) \geq 1 - \alpha.$$

Specifically, if the conformity scores $\{s_i\}_{i=1}^{2n+1}$ are distinct surely, for Algorithm 2, we have

$$\mathbb{P}(Y_{2n+1} \in C(X_{2n+1})) = 1 - \mathbb{P}(p(X_{2n+1}, Y_{2n+1}) \leq \alpha) \leq 1 - \alpha + \frac{1}{n+1}.$$

This completes the proof. ∎

A.5. Proof of Proposition 4

Proof   Consider Procedure 1-3 based on conditional conformal p-values. For any $y = 1, \cdots, K$, given $Y_{2n+1} = y$, the conditional FWER of Procedure 1-3 are all equal to

$$\mathbb{P}(\text{reject } H_y | Y_{2n+1} = y) \leq \mathbb{P}(p(X_{2n+1}|y) \leq \alpha | Y_{2n+1} = y) \leq \alpha,$$

where the inequalities follow the definitions of Procedure 1-3 and Proposition 2.

Specifically, for Procedure 3, if the conformity scores $\{s_i\}_{i \in \mathcal{I}_y \cup \{2n+1\}}$ are distinct surely, then by Proposition 2, the FWER conditional on I_y and $Y_{2n+1} = y$ is equal to

$$\mathbb{P}\Big\{\text{reject } H_y \,\big|\, I_y, Y_{2n+1} = y\Big\} = \mathbb{P}(p(X_{2n+1}|y) \leq \alpha | I_y, Y_{2n+1} = y) \geq \alpha - \frac{1}{n_y + 1}.$$

This completes the proof. ∎

19

A.6. Proof of Theorem 2

Proof By using Proposition 4 and the similar arguments as in the proof of Theorem 1, the prediction sets $C(X_{2n+1}|y)$ derived from Algorithm 1 and 2 based on the conditional conformal p-values $p(X_{2n+1}|y)$ all satisfy,

$$\mathbb{P}\Big\{Y_{2n+1} \in C(X_{2n+1}|y) \,\Big|I_y, Y_{2n+1} = y\Big\} \geq 1 - \alpha$$

for any $y = 1, \ldots, K$. Specifically, if the conformity scores $\{s_i\}_{i \in \mathcal{I}_y \cup \{2n+1\}}$ are distinct surely, for Algorithm 2, we have

$$\mathbb{P}\Big\{Y_{2n+1} \in C(X_{2n+1}|y) \,\Big|I_y, Y_{2n+1} = y\Big\} = \mathbb{P}\Big\{p(X_{2n+1}|y) > \alpha \,\Big|I_y, Y_{2n+1} = y\Big\}$$

$$\leq 1 - \alpha + \frac{1}{n_y + 1},$$

the desired result. ∎