Conformal Risk Control for Ordinal Classification

Abstract

As a natural extension to the standard conformal prediction method, several conformal risk control methods have been recently developed and applied to various learning problems. In this work, we seek to control the conformal risk in expectation for ordinal classification tasks, which have broad applications to many real problems. For this purpose, we firstly formulated the ordinal classification task in the conformal risk control framework, and provided theoretic risk bounds of the risk control method. Then we proposed two types of loss functions specially designed for ordinal classification tasks, and developed corresponding algorithms to determine the prediction set for each case to control their risks at a desired level. We demonstrated the effectiveness of our proposed methods, and analyzed the difference between the two types of risks on three different datasets, including a simulated dataset, the UTKFace dataset and the diabetic retinopathy detection dataset.

1 INTRODUCTION

In many decision making settings, a black box machine learning system is no longer adequate. Instead, we expect our system to not only make the predictions but also to quantify uncertainties and to control their risks Hüllermeier and Waegeman [2021]. This is especially true in certain high risk areas such as medical diagnosis and automatic driving.

One solution to the problem is conformal prediction Vovk et al. [1999], which gains a lot of attention recently due to its many advantages. It is distribution free, rigorous in statistics, and is easy to integrate with many machine learning models. The goal of conformal prediction is to create uncertainty sets for predictions made by these models, so that a certain coverage or risk requirement can be satisfied.

To extend the notion of error by conformal prediction, recently a new framework called conformal risk control has been developed Angelopoulos et al. [2022a]. Compared with a traditional conformal prediction method, conformal risk control generalizes the miscoverage to errors incurred by any bounded non-increasing loss functions, which offers a lot of more flexibility to problems where other metrics are valued over the miscoverage rate. Yet, it still remains a question how to define proper loss functions and to derive the corresponding prediction set for specific problems.

In this work, we focus on the ordinal classification McCullagh [1980], which is widely applied to many real life problems. In these problems, there exist a relative ordering among different classes in their label space. This ordinal nature brings unique challenges to the measurement of the prediction errors. For example, in a task of computer-aided medical diagnosis (CAMD) Juri Yanase [2019], it is much more harmful to mis-diagnose a severe condition than a mild condition, which indicates that different weights to different classes should be considered. While in a task of predicting a company's revenue range, it is desired that the predicted revenue range should be as close as possible to the actual range, which indicates that the distance between the actual range and the prediction set should be considered. Both can be captured by the conformal risk control framework.

For this purpose, we develop the conformal risk control method specifically for the ordinal classification problems. Our goal is to construct proper prediction sets in the ordinal setting so that their expected loss of prediction can be controlled. Our major contributions are four-fold.

- Formulated the ordinal classification problem in the risk control framework, along with three conditions for an ideal prediction set in the ordinal setting.
- Provided the upper and lower bounds of risk for the proposed risk control method.
- Proposed two different types of risk for constructing prediction sets in the ordinal classification setting, and developed corresponding algorithms to find the optimal prediction sets.
- Demonstrated the effectiveness of the method on both simulated and real data, and compared the difference between the two types of risks.

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

1.1 RELATED WORK

Conformal prediction was firstly developed by Vladimir Vovk and collaborators Vovk et al. [1999] Vovk et al. [2005]. Shafer and Vovk [2008], Angelopoulos and Bates [2021], Fontana et al. [2023] provide a good introduction and survey to this field of work. Our work are primarily based on split conformal prediction Papadopoulos et al. [2002] Lei et al. [2015]. Other types of conformal prediction methods have been developed, such as cross-conformal prediction Vovk [2015] Vovk et al. [2018] and CV+/Jackknife+ Barber et al. [2021] Kim et al. [2020]. Recently conformal prediction methods have been extended to the settings of non-exchangeability Tibshirani et al. [2019], Cauchois et al. [2020], Podkopaev and Ramdas [2021], Gibbs and Candes [2021], Barber et al. [2022], and have improved conditional coverage Vovk [2012], Lei and Wasserman [2014], Barber et al. [2021], Bian and Barber [2022].

Our work is most relevant to risk control in expectation Angelopoulos et al. [2022a] and ordinal conformal prediction Lu et al. [2022]. The former introduced a general framework of conformal risk control in expectation and discussed possible applications to several problems such as tumor segmentation, multi-label classification, hierarchical image classification and question answering. The latter developed a general method of ordinal conformal prediction sets with guaranteed marginal coverage for rating the disease severity in medical image. In this paper, we use the framework of conformal risk control in expectation, with the focus on ordinal classification problems.

Other relevant work includes PAC-type risk control and its applications Bates et al. [2021], Angelopoulos et al. [2021], Park et al. [2019], Park et al. [2021], Angelopoulos et al. [2022b], and Schuster et al. [2022] and conformal prediction methods for conventional binary or multi-class classification Lei [2014], Hechtlinger et al. [2018], Sadinle et al. [2019], Romano et al. [2020], Cauchois et al. [2021], Angelopoulos et al. [2020], and Kuchibhotla and Berk [2023].

2 METHOD

2.1 PROBLEM FORMULATION

Let's consider an ordinal classification problem with K classes, i.e., $\mathcal{Y} = \{0, ..., K-1\}$, where there exist an ordering among these classes. Let X_{test} be an input data, $Y_{test} \in \mathcal{Y}$ be its corresponding ground truth label, $C(X_{test}) \subseteq \{1, ..., K\}$ be a prediction set for X_{test}. Assume we also have a calibration dataset $\{(X_i, Y_i)\}_{i=1}^n$ that are drawn exchangeably from the same unknown distribution P_{XY}.

Let $L(Y_{test}, C(X_{test}))$ be a certain loss function defined as

$$L(Y_{test}, C(X_{test})) = g(Y_{test}, C(X_{test}))\mathbf{1}(Y_{test} \notin C(X_{test})) \quad (1)$$

where g is a weight function, and it measures the loss if the true label Y_{test} does not fall into the prediction set $C(X_{test})$; its value decreases as the prediction set $C(X_{test})$ grows. The choice of the weight function may impact the resulting prediction sets. In our work, we propose two forms of weight functions: a weight-based loss function and a divergence-based loss function. The former assigns different weights to different classes, while the latter incurs a loss proportional to the divergence between the true label and the prediction set.

Our goal is to construct $C(X_{test})$ so that the expected loss can be controlled at a specified α level, i.e.,

$$E[L(Y_{test}, C(X_{test})] \leq \alpha \quad (2)$$

Aside from the risk control by (2), we argue that an ideal prediction set $C(X_{test})$ in an ordinal setting should also satisfy the following three conditions:

C1. $C(X_{test})$ should be a contiguous range of classes on $\mathcal{Y}$, i.e., $C(X_{test}) = [l, u]$, where $l \leq u \in \mathcal{Y}$, and the interval $[l, u]$ stands for $\{l, l+1, ..., u-1, u\}$. This condition is needed because there is no real ordinal problem where a prediction set covers the neighboring classes while the one in the middle is skipped.

C2. The prediction set should cover the point prediction $\hat{y} \in \mathcal{Y}$, i.e., $\hat{y} \in C(X_{test})$. Here $\hat{y}$ is the class assignment when the classification model predicts a single label. This condition is needed to ensure that a conformal prediction set provides a consistent result with a regular point prediction.

C3. The prediction sets should be nested for different levels of α, i.e., $C_1(X_{test}) \subseteq C_2(X_{test})$ if $\alpha_1 \geq \alpha_2$, where C_i is the prediction set corresponding to level α_i. This is a technical assumption but it aligns well with our intuition that as the risk level increases, the prediction set should gradually reduces in a smooth way.

2.2 RISK CONTROL

To derive predictions sets for different level of risk, we can simply split the task as two steps: (1) construct a sequence of sets C_λ indexed by a threshold $\lambda \in \Lambda$; (2) determine the appropriate value $\hat{\lambda}$ of λ such that the risk of the corresponding prediction set $C_{\hat{\lambda}}$ is controlled at a desired risk level α. In this subsection, we focus on step 2 of this task by developing a general approach for determining $\hat{\lambda}$ with proven risk control of $C_{\hat{\lambda}}$, whereas in the subsequent subsections 2.3 and 2.4, we develop optimal algorithms for constructing a sequence of prediction sets C_λ specifically tailed to two different types of loss functions arising in the ordinal classification.

Specifically, consider that $C_\lambda = [l(\lambda), u(\lambda)]$ and the loss function $L(\lambda)$ is defined as in (1). Suppose that $l(\lambda)$ is decreasing and $u(\lambda)$ is increasing in λ, both right-continuous,

and there exists a value $\lambda_0 \in \Lambda$ such that $[l(\lambda_0), u(\lambda_0)] = \mathcal{Y}$. We also suppose that the weight function $g(Y, C)$ in (1) is decreasing in C in the sense that if $C_1 \subseteq C_2 \subseteq \mathcal{Y}$, we have $g(Y, C_1) \geq g(Y, C_2)$. Based on these assumptions, it's easy to check that the C_λ and $L(\lambda)$ specifically defined for the ordinal classification satisfy the following properties: (i) $L(\lambda)$ is non-increasing in λ, right-continuous; (ii) $\inf_\lambda L(\lambda) = 0$ and $\sup_\lambda L(\lambda) \leq B < \infty$ almost surely.

Let $L_i(\lambda) = L(Y_i, C_\lambda(X_i))$ be the loss value of the calibration observation (X_i, Y_i) for $i = 1, \ldots, n$. For any desired risk level upper bound α, pick the value $\hat{\lambda}$ of λ below so that the risk of $C_{\hat{\lambda}}$ is controlled:

$$\hat{\lambda} = \inf\{\lambda : \sum_{i=1}^{n} L_i(\lambda) \leq (n+1)\alpha - B\}. \tag{3}$$

Theorem 1. *Let $\hat{\lambda}$ be the value defined in equation (3), under the assumptions stated as above, we have $E[L(Y_{test}, C_{\hat{\lambda}}(X_{test})] \leq \alpha$. Moreover, if the losses $L_i(\lambda)$ are almost surely distinct for any value of λ, we have $\alpha - \frac{2B}{n+1} \leq E[L(Y_{test}, C_{\hat{\lambda}}(X_{test})]$.*

Proof. The first part of Theorem 1 directly follows from Theorem 1 in Angelopoulos et al. [2022a].

To show the second part of Theorem 1, by using the similar argument as in the proof of Theorem 2 in Angelopoulos et al. [2022a], define

$$\hat{\lambda}' = \inf\{\lambda : \sum_{i=1}^{n+1} L_i(\lambda) \leq (n+1)\alpha - B\}. \tag{4}$$

It's easy to see that $\hat{\lambda} \leq \hat{\lambda}'$, due to the facts that $L(\lambda)$ is non-increasing in λ and $\inf_\lambda L(\lambda) = 0$.

Let E be the mutliset of the loss functions $\{L_1(\hat{\lambda}'), \ldots, L_{n+1}(\hat{\lambda}')\}$. Due to $L_{n+1}(\hat{\lambda}')|E \sim Uniform\{L_1(\hat{\lambda}'), \ldots, L_{n+1}(\hat{\lambda})\}$ by exchangeability, we have

$$\begin{aligned} E[L_{n+1}(\hat{\lambda}')|E] &= \frac{1}{n+1}\sum_{i=1}^{n+1} L_i(\hat{\lambda}') \\ &\geq \frac{1}{n+1}((n+1)\alpha - 2B) = \alpha - \frac{2B}{n+1}. \end{aligned}$$

Here, the inequality follows from (4) and the assumption that the losses $L_i(\lambda)$ are almost surely distinct for any value of λ. By taking the expectation of the above inequality and using the fact that $\hat{\lambda} \leq \hat{\lambda}'$, the desired result follows. □

Remark. *The risk control of the algorithm $C_{\hat{\lambda}}$ is established by using Theorem 1 of Angelopoulos et al. [2022a], however, its lower bound cannot be obtained by using Theorem 2 therein, since that result is based on the assumption that the L_i are i.i.d. from a continuous distribution, which is not satisfied in the ordinal classification setting.*

2.3 WEIGHT-BASED RISK

For the weight-based risk function, the weight function g is independent of the choice of $C(X_{test})$, i.e., $g(y, C(X_{test})) = h(y)$. This type of functions are particularly suitable to the cases where we'd intentionally adjust the importance of certain classes. One example of $h(y)$ is constant for any label y, for which the risk corresponds to the conventional mis-coverage rate. Another example of $h(y)$ is $h(y) = y$, which assigns a higher weight to a higher class.

2.3.1 An Oracle Method

Suppose we have the oracle access to the true conditional probability distribution $f(i|x)$ of $Y_{test} = i$ given $X_{test} = x$, where $i \in \mathcal{Y}$. By using this weight function, the conditional risk given $X_{test} = x$ can be written as:

$$\begin{aligned} &E[L(Y_{test}, C(X_{test}))|X_{test} = x] \\ &= \sum_i h(i)f(i|x) - \sum_{i \in [l,u]} h(i)f(i|x). \end{aligned} \tag{5}$$

We denote the first term $\sum_i h(i)f(i|x)$ of (5) by $D(x)$, which is the upper bound of the conditional risk given $X_{test} = x$, i.e., $D(x) \leq \sum_i \max\{h(i)\}f(i|x) = \max\{h(i)\}$. For the simplicity of expression, we normalize $h(y)$ so that $\max\{h(i)\} = 1$, in which case $D(x) \leq 1$. Therefore, the range of the risk value is $[0, 1]$.

Clearly, controlling the conditional risk in (5) is equivalent to

$$\sum_{i \in [l,u]} h(i)f(i|x) \geq D(x) - \alpha, \tag{6}$$

which suggests the following rule to derive the optimal prediction set $[l, u]$ for $X_{test} = x$,

$$\begin{aligned} &(l, u) = \\ &\underset{0 \leq l \leq u \leq K-1}{\operatorname{argmin}} \{u - l : \sum_{i \in [l,u]} h(i)f(i|x) \geq D(x) - \alpha\}. \end{aligned} \tag{7}$$

2.3.2 Marginal Risk Control

Due to $f(i|x)$ is unknown in practice, we therefore replace it by the classification model output as an estimate of the probability. Thus, we can define a sequence of prediction set indexed by a threshold value, i.e.,

$$\begin{aligned} (l(\lambda), u(\lambda)) = \underset{0 \leq l \leq \hat{y} \leq u \leq K-1}{\operatorname{argmin}} &\{u - l : \\ &1 - \sum_{i \in [l,u]} h(i)\hat{f}(i|x) \leq \lambda\} \end{aligned} \tag{8}$$

where $\hat{f}(i|x)$ is the model score assigned to x on class i, $\hat{y} = \operatorname{argmax}_i\{h(i)\hat{f}(i|x)\}$ is the point prediction value.

Remark. *We seek to control the marginal risk in this work, which is weaker than the conditional risk control, since it holds only on average Vovk [2012], Lei and Wasserman [2014], and Barber et al. [2021].*

Remark. *If $h(i)$ is constant, then (8) will reduce to the one proposed by Lu et al. [2022]. In other words, Lu et al's method is a special case of the weight-based risk control.*

Under the conditions C1–C3, we solve the problem of (8) as proposed by Algorithm 1. Basically, the algorithm works in a greedy way and runs at a complexity of $O(K)$. It starts with the point prediction label to grow the prediction set step by step until the risk requirement is satisfied, therefore it guarantees that the point prediction $\hat{y}$ is always covered within the prediction set. It also ensures the nested property since it does not shrink the prediction set in this process.

Algorithm 1: Determine the prediction set for a given λ

Input: λ, $h(i)$, $\hat{f}(x|i)$ for $i \in \{0, ..., K-1\}$
Output: l, u

$s(i) \leftarrow h(i)\hat{f}(x|i)$ for $i \in \{0, ..., K-1\}$
$u, l \leftarrow \text{argmax}_i\{s(i)\}$
$sum \leftarrow 0$
while $sum < (1-\lambda)$ **do**
  if $s(l-1) > s(u+1)$ **then**
    $sum = sum + s(l-1), l = l-1$
  else
    $sum = sum + s(u+1), u = u+1$
return l, u

Theorem 2. *The prediction set derived by Algorithm 1 satisfies conditions C1–C3 and is optimal in the sense of satisfying (8).*

Proof. Let $C_\lambda = [l(\lambda), u(\lambda)]$ denote the prediction set derived by Algorithm 1 and C^*_λ be the optimal prediction set satisfying (8) and conditions C1–C3. It's easy to check that C_λ satisfies conditions C1–C3. In the following, we prove by contradiction that C_λ also satisfies (8). Assume that there exits a value λ_0 of λ such that $C_{\lambda_0} \neq C^*_{\lambda_0}$. Let $\lambda_1 = \sup\{\lambda : C_\lambda = C^*_\lambda\}$ and $C_{\lambda_1} = C^*_{\lambda_1} = [l_1, u_1]$.

Without loss of generality, suppose $s(l_1 - 1) > s(u_1 + 1)$. It's easy to see that for $\lambda \in (\lambda_1, \lambda_1 + s(u_1 + 1)]$, $C_\lambda = [l_1 - 1, u_1]$ and $C^*_\lambda = [l_1, u_1 + 1]$, due to the definition of λ_1. However, for $\lambda \in (\lambda_1 + s(u_1 + 1), \lambda_1 + s(l_1 - 1)]$, $C_\lambda = [l_1 - 1, u_1]$ keeps unchanged, but $C^*_\lambda = [l_1, u_1 + 1]$ needs to grow by one class so that it can satisfy (8), which contradicts the optimality assumption of C^*_λ. □

To calculate the actual value of λ, we can do a linear search. However, for an improved efficiency, we can also do a binary search as shown in Algorithm 2.

Algorithm 2: Determine the value of $\hat{\lambda}$ for a given α

Input: $\{x_i, y_i\}$ for $i \in \{0, ..., n\}$
Output: $\hat{\lambda}$
Parameters : precision δ

$\lambda_0 \leftarrow 0$, $\lambda_1 \leftarrow 0.5$
while $\Delta\lambda_k = |\lambda_k - \lambda_{k-1}| > \delta$ **do**
  $L(\lambda_k) \leftarrow 0$
  for $i \leftarrow 1$ **to** n **do**
    Calculate (l_i, u_i) for x_i and λ_k using Algorithm 1
    Calculate $L_i(\lambda_k)$ from (l_i, u_i) and y_i
    $L(\lambda_k) \leftarrow L(\lambda_k) + L_i(\lambda_k)$
  if $L(\lambda_k) > (n+1)\alpha - 1$ **then**
    $\lambda_{k+1} \leftarrow \lambda_k - \frac{\Delta\lambda_k}{2}$
  else
    $\lambda_{k+1} \leftarrow \lambda_k + \frac{\Delta\lambda_k}{2}$
return $\hat{\lambda} = \lambda_{k+1}$

2.4 DIVERGENCE-BASED RISK

For the divergence based risk, the loss is proportional to the distance between the true label and the prediction set, i.e., $g(Y_{test}, C(X_{test})) = \inf\{d(y, i) : i \in [l, u]\}$, where $d(\cdot, \cdot) \in \mathcal{Y} \times \mathcal{Y} \to \mathbf{R}^+$ is a certain distance measure on the label space. This type of functions are particularly suitable to the cases where we are more concerned with large prediction errors than the differences among the individual classes.

There could be different ways to define the difference measure $d(\cdot, \cdot)$. For example, $d(y, i) = |y - i|^p$. In this work, we only consider the case where $p = 1$. Furthermore, we normalize the loss value by $K - 1$, therefore,

$$g(Y_{test}, C(X_{test})) = \frac{1}{K-1} \inf\{|y-k| : k \in [l, u]\}. \quad (9)$$

Clearly, when $l = 0$ and $u = K - 1$, the prediction set covers the whole label space, we have $L(Y_{test}, C(X_{test})) = 0$. The upper bound of the risk, on the other hand, is 1, which happens when the true label and the prediction set are exactly on two opposite boundaries of the label space. Therefore $E[L(Y_{test}, C(X_{test}))|X_{test}] \in [0, 1]$.

2.4.1 An Oracle Method

Suppose we have the oracle access to the true conditional probability distribution $f(i|x)$ of Y_{test} given $X_{test} = x$. Then the conditional risk for a given $X_{test} = x$ can be written as:

$$E[L(Y_{test}, C(X_{test}))|X_{test} = x] = \frac{1}{K-1}\left(\sum_{i<l}(l-i)f(i|x) + \sum_{i>u}(i-u)f(i|x)\right) \quad (10)$$

It is easy to to see that the first term of (10) increases as l increases, while the second term decrease as u increases. The optimal prediction set $[l^*, u^*]$ for $X_{test} = x$ is therefore given by

$$(l^*, u^*) = \underset{0 \le l \le u \le K-1}{\operatorname{argmin}} \{u - l : \sum_{i<l} (l-i) f(i|x) + \sum_{i>u} (i-u) f(i|x) \le (K-1)\alpha\} \tag{11}$$

2.4.2 Marginal Risk Control

Similarly, we use the classification model output as an estimate of $f(i|x)$ and seek to control the marginal risk instead. Under this setting, we define a sequence of prediction set $[l(\lambda), u(\lambda)]$ indexed by a threshold value λ, i.e.,

$$(l(\lambda), u(\lambda)) = \underset{0 \le l \le \hat{y} \le u \le K-1}{\operatorname{argmin}} \{u - l : \sum_{i<l} (l-i) \hat{f}(i|x) + \sum_{i>u} (i-u) \hat{f}(i|x) \le \lambda\}, \tag{12}$$

where $\hat{f}(i|x)$ is the model score assigned to x on class i, and $\hat{y} = \operatorname{argmax}_i \{\hat{f}(i|x)\}$ is the point prediction value.

To determine the prediction set for a given λ, it helps to firstly look into how risk changes when the prediction set size adjusts by 1. For the simplicity of expression, let $R(l, u)$ be the risk incurred by a prediction set $C(X_{test)} = [l, u]$, which can be calculated by (10). It's easy to see that

$$R(l+1, u) - R(l, u) = \sum_{i<l+1} (l+1-i) \hat{f}(i|x) - \sum_{i<l} (l-i) f(\hat{i}|x) = \sum_{i \le l} \hat{f}(i|x), \tag{13}$$

Similarly,

$$R(l, u-1) - R(l, u) = \sum_{i>u-1} (i-(u-1)) \hat{f}(i|x) - \sum_{i>u} (i-u) \hat{f}(i|x) = \sum_{i \ge u} \hat{f}(i|x), \tag{14}$$

Therefore, the risk change incurred by every single adjustment of the prediction set boundary can be easily calculated beforehand at a complexity of $O(K)$. With this insight, we propose Algorithm 3 below to determine the prediction set under conditions C1–C3.

Basically, this is a greedy algorithm that runs at a complexity of $O(K)$. The algorithm starts with the class that has the maximum value of $\hat{f}(i|x)$, therefore it guarantees that the point prediction $\hat{y}$ is always covered within the prediction set. The algorithm grows the prediction set one class a step. At each step, it adjusts the prediction boundary by maximizing the risk reduction of that step, until it reaches the set that satisfies the risk requirement. Since it does not shrink the prediction set in this process, the nest property is also ensured.

Algorithm 3: Determine the prediction set for a given λ

Input: λ, $\hat{f}(x|i)$ for $i \in \{0, ..., K-1\}$
Output: l, u

```
Calculate head(j) = Σ_{i≤j} f(i|x),
  tail(j) = Σ_{i≥j} f(i|x) for i ∈ {0, ..., K−1}
l, u ← argmax_i{f̂(x|i)}
sum ← R(l, u)
while sum > λ do
    if head(l − 1) ≥ tail(u + 1) then
        sum = sum − head(l − 1)
        l = l − 1
    else
        sum = sum − tail(u + 1)
        u = u + 1
return l, u
```

To calculate the value of $\hat{\lambda}$ for a given α, we follow the same binary search process as used by Algorithm 2.

Theorem 3. *The prediction set derived by Algorithm 3 satisfies conditions C1–C3 and is optimal in the sense of satisfying (12).*

The proof is similar to the one for Theorem 2, therefore is omitted here.

3 EXPERIMENTS

To demonstrate the effectiveness of the proposed algorithms, we evaluated them on three ordinal classification tasks, including a task to classify a simulated ordinal data, a task to predict the age group of the person based on their face images, and a task to predict the severity of patients' diabetic retinopathy based on their retina images.

For all experiments in this work, data are equally split for validation and for test. The softmax output of the ordinal classification model are fed as the input to our algorithm, and the results are averaged over 100 random trials.

3.1 SIMULATED DATA

We simulated a 10-class ordinal data on a 2-D plate. Each class has 2,000 data points sampled from a Gaussian distribution, where the ith class is centered at the coordination $[i, i]$, with a randomly generated covariance matrix. Figure 1 displays the distribution of the data.

To classify the ordinal labels, we built a two-layer MLP with

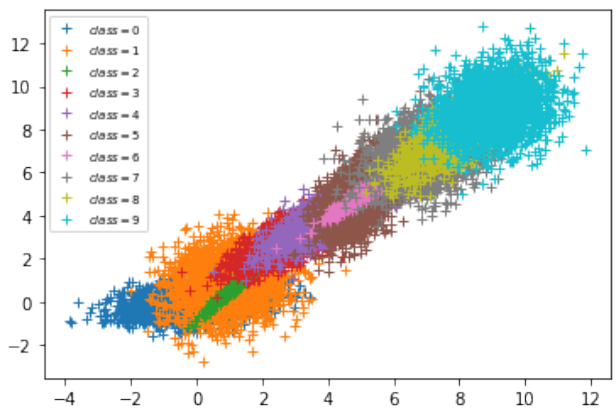


Figure 1: A simulated 10-class ordinal data

Table 1: Actual risk at different value of α on simulated data

α	0.02	0.08	0.14	0.20
S1	0.0199	0.0807	0.1403	0.1999
S2	0.0201	0.0802	0.1402	0.2001
S3	0.0201	0.0782	0.1406	0.1997
S4	0.0199	0.0463	0.0463	0.0464

50 neurons on the hidden layer. We use 14,000 data points for validation and test.

We evaluate the following 4 scenarios:

S1: weight-based risk with equal weights on all classes, which is essentially equivalent to the method proposed by Lu et al. [2022].

S2: weight-based risk with incremental weights on higher classes, where class i has a weight i;

S3: weight-based risk, with double weights on class 5∼9;

S4: divergence-based risk.

For all scenarios, we evaluate the algorithm with different α values. The actual risk calculated are given in Table 1, which shows that the proposed algorithm produces risk values very close to the α values for the first three scenarios.

For the last scenario, the actual risk saturates around 0.0463 since $\alpha = 0.08$. This is due to that the max divergence risk on this data set has been reached. Afterwards, the prediction set shrinks to a single class, therefore, the calculated risk no longer changes.

Need to point out that although the divergence-based risk has a same value range of $[0, 1]$ as the weight-based risk, it generally has a much smaller max risk value. In fact, the risk value of 1 can only be reached if every data point in the dataset has a true label that is on the opposite side of the predicted set on the label space, which is impossible in practice.

Figure [2] illustrates the trend of prediction set sizes at different values of α for these scenarios. It can be seen that as the value of α increases, the set sizes reduce. For the divergence-based risk, the set size shrinks to 1 around $\alpha = 0.05$, which aligns with the calculated max risk value 0.0463. The figure also shows that, for the weight-based risk, both the incremental weights and the double weights scenarios have larger prediction sets compared with the equal weights scenario on the same α value. This is because that the algorithm needs to extend the prediction set accordingly to reduce the overall risk.

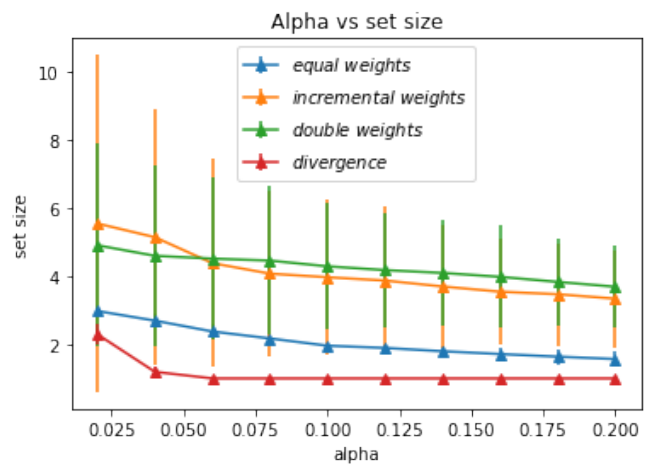
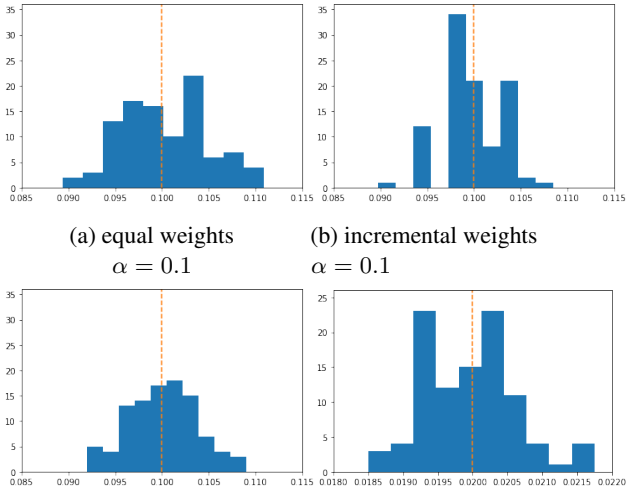


Figure 2: Prediction set sizes at different values of α on simulated data

Figure 3 illustrates the risk distribution at a fixed α for different scenarios, where α is fixed at 0.10 for the first three scenarios and at 0.02 for the last scenario. It shows that for all the scenarios, the algorithm controls the risk well within narrow ranges around the specified α values, which are shown by the orange dotted lines in the graphs.

We also looked into the difference between the weight-based risk and the divergence-based risk. For this purpose, we compare their predicted class ranges. Due to their risk may take very different values, for a meaningful comparison, we carefully choose their α values so that their averaged prediction sets size are both 3. Then we compare the distribution of the centroids of their prediction sets, which are shown in Figure 4. It can be seen that compared with the equal weights-based loss function, the divergence-based loss function tends to push the centroids toward the center of the label range. In other words, it is more centripetal than the weight-base risk. This can be explained by the fact that the divergence-based loss function punishes more on the extreme error cases where the true label lies on the opposite side of the prediction set.



(a) equal weights $\alpha = 0.1$ (b) incremental weights $\alpha = 0.1$

(c) double weights, $\alpha = 0.1$ (d) divergence, $\alpha = 0.02$

Figure 3: Risk distributions of different scenarios at a fixed α on the simulated dataset.

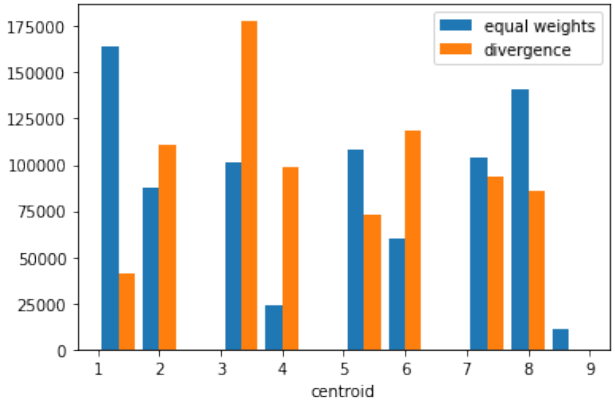


Figure 4: Distributions of prediction centroids on simulated data.

3.2 AGE RECOGNITION

For this task, we use the UTKFace dataset[1], which is a large-scale face dataset with over 20K images that cover a long age span (range from 0 to 116 years old). Each image has been annotated of the person's age. In our work, we kept those below age 100, and discretized the age into 20 groups, where each group covers 5 years in range, i.e., group 0 is for $0 \sim 4$, and group 1 is for $5 \sim 9$ and so on. Here are some examples of the images and their group labels.

We built an ordinal classifier to predict the age group using ResNet34 He et al. [2016], where we didn't particularly tune the hyper-parameters for a superb classification result. We use 17K images for validation and test, and evaluate the following scenarios:

[1]The data is available at https://susanqq.github.io/UTKFace/.

(a) age 8, group 1 (b) age 25, group 5 (c) age 42, group 8

(d) age 67, group 13 (e) age 86, group 17

Figure 5: Image examples in the UTKFace dataset.

Table 2: Actual risks at different values of α on UTKFace

α	0.005	0.01	0.025	0.05	0.10	0.20
S1	0.005	0.010	0.025	0.050	0.100	0.201
S2	0.005	0.010	0.025	0.050	0.100	0.200
S3	0.005	0.010	0.025	0.050	0.100	0.169
S4	0.005	0.010	0.016	0.016	0.016	0.016

- S1: weighted risk, with equal weights on all classes
- S2: weighted risk, with double weights on class 0∼3
- S3: weighted risk, with double weights on class 16∼19
- S4: divergence risk

Here shows in Table 2 the actual risks for different α values in these scenarios, which shows our algorithm produces the actual risks very close to the α value. It also shows that as α increases to certain values, the risks of the last two scenarios saturate to their max risk values. The trend of prediction set sizes over different α values and the risk distribution at a fixed α show similar conclusion as on the simulated data, and are therefore omitted here due to the length restriction.

We also fix the prediction set size to be 3 and compared the distributions of the prediction set centroids for the different scenarios, which are shown in Figure 6. From the distributions, we can see that by doubling the weights of the group $0 \sim 3$, it pushes the distribution towards the lower age end, while by doubling the weights of the group $16 \sim 19$, it pushes the distribution towards the higher age end. Similarly, the divergence loss function tends to be centripetal compared with other loss functions.

3.3 DIABETIC RETINOPATHY DETECTION

For this task, we use the diabetic retinopathy detection dataset[2]. It is a large set of over $35K$ retina images taken under a variety of imaging conditions. Each image has a

[2]The data is available at https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data.

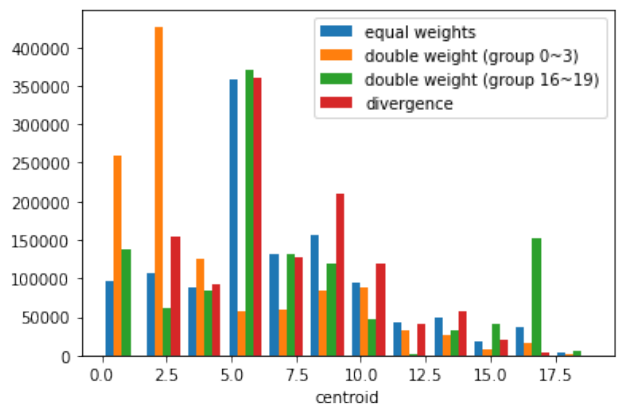
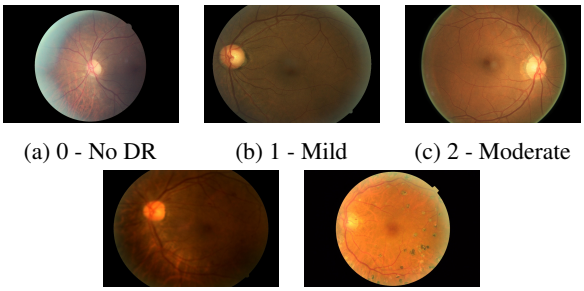


Figure 6: Distributions of prediction centroids on UTKFace

clinician rating on the presence of diabetic retinopathy (DR) using a scale of 0 to 4, where 0 means no DR while 4 means a proliferative DR. Figure 7 shows some examples of the image along with their labels.



(a) 0 - No DR (b) 1 - Mild (c) 2 - Moderate

(d) 3 - Severe (e) 4 - Proliferative

Figure 7: Image examples in the DR dataset.

Similarly, we built an ordinal classifier to predict the age group using ResNet34, and didn't particularly tune the hyper-parameters for a superb classification result. We use 19K images for validation and test, and evaluate the following scenarios:

- S1: weighted risk, with equal weights on all classes
- S2: weighted risk, with double weights on class 3 & 4
- S3: divergence risk

We show the expected risk for different alpha values of these scenarios in Table 3. The trend of prediction set sizes over different α values and the risk distribution at a fixed α show similar conclusion as on the simulated data, and are therefore omitted here due to the length restriction.

We also fix the prediction set size to be 2 and compared the distribution of the prediction set centroids for the different risk functions, as shown in Figure 8. Similarly, compared with the equal weights, doubling weights pushes the distribution slightly towards the higher end, while divergence risk pushes the distribution slightly towards the center.

Table 3: Actual risk on different value of α on the DR dataset

α	0.02	0.08	0.14	0.20
S1	0.0198	0.0794	0.1402	0.1499
S2	0.0199	0.0801	0.1401	0.1485
S3	0.0200	0.0430	0.0428	0.0428

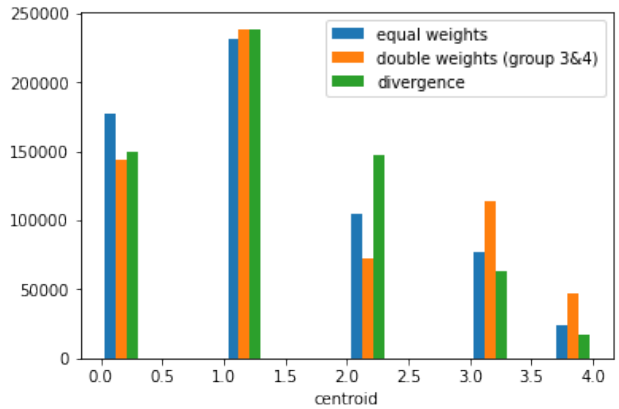


Figure 8: Distributions of prediction centroids on DR data

4 CONCLUSION & DISCUSSION

We formulated the ordinal classification task within the recently developed framework of conformal risk control in expectation, and introduced two types of loss functions specifically tailed to the learning task. Based on these two loss functions, we developed two conformal prediction algorithms, which are shown controlling the corresponding conformal risk at a desired level and are optimal in some sense. Simulation study and real data analysis showed effectiveness of the proposed algorithms.

With the proposed method, we can design weight functions accordingly to ensure the coverage that we want for the specific task. Between the two types of the proposed loss functions, the weight-based risk is more suitable to the cases where we want to adjust the importance of certain classes, while the divergence-based risk is more suitable to the cases where we are more concerned with large prediction errors than the differences among the individual classes. By comparison, the divergence-based risk is more centripetal since it pushes the centroids of the prediction sets toward the center of the label range, while the weight-based risk pushes the centroids toward the classes where they have higher weights.

There are a few important questions remain. Firstly, instead of the marginal coverage, how the choice of weight functions impacts the conditional coverage for each single class. Secondly, how to choose the risk threshold that is fit to the specific problem and weight function. We will investigate these questions in our future work.

References

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

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

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv:2208.02814*, 2022a.

Anastasios N Angelopoulos, Karl Krauth, Stephen Bates, Yixin Wang, and Michael I Jordan. Recommendation systems with distribution-free reliability guarantees. *arXiv preprint arXiv:2207.01609*, 2022b.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *Ann. Statist.*, 49:486–507, 2021.

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

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *arXiv preprint arXiv:2205.03647*, 2022.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *The Journal of Machine Learning Research*, 22(1):3681–3722, 2021.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 770–778, 2016.

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.

et al Juri Yanase. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 2019.

Byol Kim, Chen Xu, and Rina Barber. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33:4138–4149, 2020.

Arun K Kuchibhotla and Richard A Berk. Nested conformal prediction sets for classification with applications to probation data. *The Annals of Applied Statistics*, 17(1): 761–785, 2023.

Jing 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.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74: 29–43, 2015.

Charles Lu, Anastasios N. Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. *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 *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. PAC confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*, 2019.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC prediction sets under covariate shift. *arXiv preprint arXiv:2106.09848*, 2021.

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

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114 (525):223–234, 2019.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. *arXiv preprint arXiv:2207.07061*, 2022.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 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.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

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

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 37–51. PMLR, 2018.