Conformal Prediction: A Review and an Application in Deep Learning-based Image Classification

Tianyu Wang

Department of Mathematics and Statistics

McGill University, Montreal

August, 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

Abstract

Conformal prediction is a valuable framework for uncertainty quantification. This method produces prediction results with coverage guarantee under minimal assumptions. The thesis reviews recent developments in conformal prediction. It summarizes different calibration strategies applicable to conformal prediction algorithms, as well as various extensions of conformal prediction framework that are conditional valid or valid under non-exchangeable conditions. Finally, the thesis attempts to explore the feasibility of logits-based nonconformity scores. We examined two logits-based methods adapted from existing conformal prediction algorithms based on estimated probabilities. An empirical comparison of these methods was conducted through image classification application.

Abrégé

La prédiction conforme est un cadre précieux pour la quantification de l'incertitude. Cette méthode produit des résultats de prédiction avec une garantie de couverture sous des hypothèses minimales. Cette thèse passe en revue les développements récents de la prédiction conforme. Elle résume différentes stratégies de calibration applicables aux algorithmes de prédiction conforme, ainsi que diverses extensions du cadre de la prédiction conforme qui sont valides sous conditions ou valides dans des conditions non échangeables. Enfin, la thèse tente d'explorer la faisabilité des scores de non-conformité basés sur les logits. Nous avons examiné deux méthodes basées sur les logits, adaptées d'algorithmes de prédiction conforme existants, basées sur des probabilités estimées. Une comparaison empirique de ces méthodes a été réalisée à travers une application de classification d'images.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Abbas Khalili, for his continuous guidance, patience, and support throughout my studies. I am deeply thankful for his mentorship, which has enabled me to come to McGill and further my study. I also would like to thank Professor Mehdi Dagdoug for taking time to read my thesis and give me valuable feedback. I also wish to extend my thanks to my supervisor and the Department of Mathematics and Statistics for their financial support of my study. Lastly, I am profoundly grateful to my parents for their unwavering support and companionship throughout this journey.

Table of Contents

List of Tables

Chapter 1

Introduction

Conformal prediction is a powerful and flexible framework that allows for the quantification of uncertainty in predictive modeling. By constructing prediction sets instead of single-point predictions, conformal prediction provides a measure of confidence that is valid assuming only exchangeable assumption. This makes it particularly useful in prediction where reliability and robustness are paramount. The primary goal of conformal prediction is to ensure that the constructed prediction sets contain the true label with a specified probability. While the framework offers strong guarantees, real-world data do not always follow the exchangeable assumption. This thesis explores both the theoretical underpinnings of conformal prediction and its extensions to more complex scenarios where traditional assumptions do not hold.

The following chapters of this thesis are organized as follows:

Chapter 2 motivates the discussion of conformal prediction methods through two examples and delves into the challenges faced by basic conformal prediction algorithms in real-world applications.

Chapter 3 introduces the fundamental concepts of conformal prediction and discusses its significance in the context of statistical learning. This chapter explores various extensions of the conformal prediction framework, including various calibration strategies, conditionally valid conformal predictors, and conformal predictors for covariate shift and

distributional shift. These variants expand the applicability of conformal prediction to a wider range of scenarios.

Chapter 4 presents practical application of conformal prediction on image classification tasks using convolutional image classifiers. The chapter explores the feasibility of designing logits-based nonconformity scores for deep learning-based image classification tasks. It compares existing probability-based methods: Regularized Adaptive Prediction Sets (RAPS) [3] and Adaptive Prediction Sets (APS) [33] with two logits-based conformal methods: APS-Logits and RAPS-Logits which are adapted from the previous two methods.

This thesis aims to contribute to the growing body of research on conformal prediction by offering both theoretical insights and practical applications. The subsequent chapters will provide a detailed exploration of these topics.

Chapter 2

Conformal Prediction Examples

Prediction has always been an important goal of mathematical models. Machine learning algorithms provide us vast methods to obtain point predictions producing solutions to various interesting tasks such as the applications of neural network in facial recognition, random forest algorithm for credit scoring, support vector machine in protein or cancer classification, and so on. However, it remains a challenge to measure the reliability of the predictions made by these "black box" algorithms. Quantifying the uncertainty becomes an important aspect for prediction algorithms to produce meaningful results for real-world applications such as decision-making in the health care field where prediction errors may lead to serious consequences.

Reliable prediction intervals with guarantee in coverage probability can be obtained in various ways. For example, one classic method for producing reliable prediction intervals for regression problems is using the linear regression models [28]. Given response vector $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and predictor variables $\mathbf{x_j} = (1, x_{1j}, \ldots, x_{nj})^T \in \mathbb{R}^n$, $j = 1, \ldots, p$. Let $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_p}) \in \mathbb{R}^{n \times (p+1)}$ denote the $n \times (p+1)$ design matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denote the regression parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ denote the error term that is independent of design matrix $\mathbf{X}$ with $\epsilon_i \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$. A multiple linear regression model relating response variable $\mathbf{y}$ to predictor variables $\mathbf{X}$ can be expressed

as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

By the least squares method the regression coefficient $\boldsymbol{\beta}$ can be estimated by solving the following minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta \in \mathbb{R}^{p+1}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

given the inverse matrix $(\mathbf{X^T X})^{-1}$ exists, we obtain the least square estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$. The unknown parameter σ^2 can be estimated as the residual mean square as follows

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n - (p+1)}.$$

Then, to predict a future observation y_0 corresponding to predictor variables $\mathbf{x_0} = (1, x_{01}, \ldots, x_{0p})^T$ we can use the point prediction $\hat{y_0} = \mathbf{x_0}^T \hat{\boldsymbol{\beta}}$ based on the statistic

$$t = \frac{\hat{y_0} - y_0}{\hat{\sigma}\sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}}$$

we can construct a 100(1-α) percent prediction interval for y_0:

$$\hat{y_0} - t_{\alpha/2, n-(p+1)} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0)} \leq y_0 \leq \hat{y_0} + t_{\alpha/2, n-(p+1)} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0)},$$

where t denotes the $(1 - \alpha/2)$ quantile of t-distribution with n-(p+1) degrees of freedom. The prediction interval produced by fitting a linear regression model is one way of quantifying the uncertainty of prediction to satisfy a desired confidence level.

For classification problem, one example which gives reliable prediction is the logistic regression model. Suppose the response variable $\mathbf{y} = (y_1, \ldots, y_n)^T \in \{0, 1\}^n$ is binary with $y_j \sim Bin(1, p_j)$ and the corresponding covariate is $\mathbf{x_j} = (1, x_{j1}, \ldots, x_{jk})^T \in \mathbb{R}^{k+1}$, $j = 1, \ldots, n$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ denotes the regression parameters then the logistic

regression models relates the log odds ratio linearly to the covariates as follows:

$$\log\left(\frac{p_j}{1-p_j}\right) = \mathbf{x_j}^T\boldsymbol{\beta}.$$

By numerically maximizing the log likelihood function of logistic regression

$$l(\boldsymbol{\beta}) = log(\prod_{j=1}^{n} p(\mathbf{x_j})^{y_j}(1-p(\mathbf{x_j}))^{1-y_j}) = \sum_{j=1}^{n} -\log(1+e^{\mathbf{x_j}^T\boldsymbol{\beta}}) + \sum_{j=1}^{n} y_j(\mathbf{x_j}^T\boldsymbol{\beta})$$

we can obtain the maximum likelihood estimates of coefficients denoted by $\hat{\boldsymbol{\beta}}$. For a new example with explanatory variable $\mathbf{x_0} = (1, x_{01}, ..., x_{0k})^T$, the estimated conditional probability is

$$P(Y=1|X=x_0) = \frac{e^{\mathbf{x_0}^T\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x_0}^T\hat{\boldsymbol{\beta}}}}.$$

The conditional probability give us information on the probabilities of the new observation falling into each category. If the estimated conditional probability is greater than the decision boundary 0.5 (i.e. $P(Y=1|X) > 0.5$), then the new observation is predicted to belong to the class $\hat{Y}=1$, otherwise, we predict $\hat{Y}=0$. Note that although both examples successfully generate prediction (intervals) with informative confidence level, they both require distributional assumptions on the data, which, in reality, can be easily violated.

Comparing to existing predicting methods mentioned above, a relatively new method conformal prediction (CP) gives solution to the problem of quantifying the uncertainty prediction in an innovative way with minimal assumption on the data. The concept was first introduced by Vladimir Vovk, Alexander Gammerman, Craig Saunders, and Vladimir Vapnik in the years 1996-1999 [2]. The method is powerful as it is distribution-free and model-agnostic while still provides guarantee in coverage rate on prediction intervals. These advantages enable CP to solve multiple tasks including regression, classification, time-series forecasting, outlier detection and so on. Moreover, CP can be combined with many "black box" prediction algorithms to convert their point predictions into prediction intervals or prediction sets with guaranteed coverage rate. Its ability to trans-

form heuristic uncertainty into rigorous uncertainty made it an increasingly popular area of research. Over the past few decades, conformal prediction has been through several major development. Jing Lei and Larry Wasserman [25] created a general framework for distribution-free predictive inference in regression. Another group of researchers including Rina Barber, Emmanuel Candes, Ryan Tibshirani, and so on [6] further contributed to extend CP to covariate-shift and distribution-shift settings. As more researchers join this interesting research area and more powerful conformal predictors were invented, increasingly large amount of research have been conducted to extend and refine this method to be more adaptive and robust in different scenarios.

2.1 Motivating Examples

2.1.1 Application in Regression Problem

In this section, we will compare the classic linear regression model based conformal prediction algorithm: least square confidence machine (LSCM) [43] which is a special case of the ridge regression confidence machine that will be introduced in 3.1.1, full conformal prediction (FCP) method introduced in [20] and nearest neighbour conformal regressor under three settings. In setting A, we generated independently and identically distributed (i.i.d.) examples that meet regularity conditions. In setting B, we generated data from mixture of i.i.d. bivariate normal random vectors so that the examples are exchangeable but not i.i.d. In setting C, we generated time series data from the AR(1) process and thus the data is nonexchangeable.

Under three settings, we compare the performances of three different prediction algorithms in terms of their validity and efficiency. The first algorithm RRCM was introduced in [43] to illustrate how conformal predictor framework apply to regression problems. This algorithm adopts ridge regression as the underlying point predictor to produce nonconformity scores. It uses the residuals $|y_i - \hat{y}_i|$ as a natural choice of nonconformity score measure in regression setting. The prediction results include all values of $y \in \mathbb{R}$ such

Setting A

	RRCM	Full Conformal	1-NNR
Coverage	0.94	0.90	0.93
Length	3.29	3.53	4.70
Time	1.44	0.76	1.70

Table 2.1: Comparison of prediction results of three conformal algorithms under setting A.

that $p_y > \epsilon$ where p_y is the p-value defined as the fraction of nonconformity scores that is greater than or equal to the nonconformity score of new example (x_{new}, y). Detailed explanation of RRCM can be found in section 3.1.1. We implemented the algorithm by identifying all the values of y that give a p-value greater than the desired miscoverage level, which is $\alpha = 0.1$ in our example. The second method, full conformal prediction (FCP) [20], is similar to RRCM as it also uses linear regression model as a point predictor and residuals as nonconformity scores. The slight difference between two algorithms is that RRCM searches through the entire real line to include all possible values of y satisfying the criteria, while full conformal prediction creates a grid of y values to estimate quantiles of the nonconformity scores and improves computational efficiency. Thirdly, we showed the performance of k nearest neighbour conformal predictor that uses k nearest neighbour algorithm as the underlying predictor where we take k=1. More details of the conformal prediction framework is presented in section 3.1.

The results from the following three tables show that all three methods successfully guarantee the expected coverage rate of 90%. The RRCM and 1-NNR methods have higher coverage levels than the full conformal prediction method, but the sizes of their prediction sets are larger than that of the full conformal method. We will discuss more on the efficiency and validity tradeoff in the next chapter. Also, RRCM and 1-NNR are significantly more computationally inefficient compared to full conformal prediction. Thus, although basic conformal prediction algorithms produce reliable prediction intervals as expected, there still exist spaces to modify their design to improve the efficiency and reduce computational cost while maintaining the coverage guarantee.

Setting B

	RRCM	Full Conformal	1-NNR
Coverage	0.95	0.90	0.95
Length	4.85	2.85	4.26
Time	1.32	0.46	1.78

Table 2.2: Comparison of prediction results of three conformal algorithms under setting B.

Setting C

	RRCM	Full Conformal	1-NNR
Coverage	0.94	0.89	0.95
Length	4.78	3.21	3.41
Time	1.72	0.47	1.33

Table 2.3: Comparison of prediction results of three conformal algorithms under setting C.

2.1.2 Application in Classification Problem

When the label space is finite $|Y| < \infty$, we consider such situation as classification problem. For conformal classifiers, we no longer have the natural choice of residuals to measure nonconformity scores in most cases which requires us to define nonconformity measure A in other ways. [34] introduced three basic choices of nonconformity scores for binary classification problems, which are introduced in detail in Section 3.1.1. In their paper, the three conformal algorithms were applied to the Iris flower dataset to predict flower species using sepal length as feature. In this section, we adopted the same three methods on the banknote dataset where the entropy of banknote images is treated as feature to classify fake and authentic banknotes. The three methods use three different designs of the nonconformity scores based on three distinct underlying classifiers: 1-nearest neighbor, distance to the mean, and support vector machine (SVM). After obtaining the nonconformity score for each example, the p-value is defined in the same way as the fraction of nonconformity scores that are greater than or equal to that of the new example. In binary classification, we can obtain two p-values denoted by p_0 and p_1 for each class respectively. Then, for a specified significance level α we can obtain three possible prediction sets [34]:

• **Uncertain Prediction set** : $\{0, 1\}$ when $p_0 > \epsilon$ and $p_1 > \epsilon$.

• **Singleton Prediction set** : $\{0\}$ when $p_0 > \epsilon$ and $p_1 \leq \epsilon$; $\{1\}$ when $p_0 > \epsilon$ and $p_1 \leq \epsilon$.

• **Empty Prediction set** : $\emptyset$ when $p_0 \leq \epsilon$ and $p_1 \leq \epsilon$.

Note that both the empty set and the uncertain prediction set are equally uninformative. One way to obtain a point prediction is to choose the class that gives the highest p-value [4]. For classification problems, in addition to validity, it is also useful to look at *credibility* [43] of each prediction which is defined to be the largest p-value for the new example. For example, in our case of binary classification, credibility would be $max(p_0, p_1)$. A low credibility indicates that even the point prediction is unlikely to be observed. This can happen when the object is unusual for the chosen method and we should be careful with being overconfident about such prediction results.

The following two tables compare the prediction results of three conformal classifiers on Iris flower dataset and banknote dataset respectively. Table 2.4 presents the results from the classification example in [34]. Our example applied the conformal classifiers to the banknote dataset which consists of 1372 examples. Each example has a label $y_i \in \{0, 1\}$ indicating whether the banknote is fake $y_i = 0$ or authentic $y_i = 0$ and an object $x_i \in \mathbb{R}$ that depicts the entropy of image. The training data of 40 examples was sampled from the banknote dataset without replacement. We tested for one example in each repetition and summarized the total number of errors and correct predictions among a total of 1000 repetitions.

Comparing the performances of the three conformal classifiers on two different datasets, we notice that both results highlight conformal prediction's advantage in marginal coverage guarantee as all the percentages of correctly predicted examples are greater than or equal to the desired coverage level. However, in both examples, there still exists a large portion of uninformative prediction results which can be problematic in practice. Moreover, the SVM-based conformal classifier works significantly slower than the other two methods which suggests that basic conformal framework could be prevented from

	Nearest Neighbor	Distance to the Average	SVM
singleton hits	164	441	195
uncertain	795	477	762
total hits	959	918	957
empty	9	49	1
singleton errors	32	33	42
total errors	41	82	43
total examples	1000	1000	1000
%hits	96%	92%	96%
total singletons	196	474	237
%hits	84%	93%	82%

Table 2.4: Comparison of three conformal classifiers' performances on Iris flower dataset with α = 0.08. [34]

	Nearest Neighbor	Distance to the Average	SVM
singleton hits	52	6	23
uncertain	900	949	928
total hits	952	955	951
empty	0	35	27
singleton errors	48	10	22
total errors	48	45	49
total examples	1000	1000	1000
%hits	95.2%	95.5%	95.1%
total singletons	100	16	45
%hits	52%	37.5%	51.1%

Table 2.5: Comparison of three conformal classifiers' performances on Banknote dataset with α = 0.05.

collaborating with complex base classifiers due to computational inefficiency. Similar to the example in regression setting, although basic conformal prediction algorithms are promising in giving predictions with confidence, there still exist many limitations and shortcomings to overcome to make conformal prediction more practical and applicable in real-world scenarios. In the next chapter, we will review the recent developments in conformal prediction and introduce various improved CP methods that can address the problems detected from these examples.

Chapter 3

Selective Review of Conformal Prediction

In this chapter, we formally introduce the conformal prediction algorithm in section 3.1 and then discuss desirable properties of conformal predictors in section 3.2. The rest of this chapter provides readers with various advanced and modified conformal methods designed to enhance the validity, efficiency, and robustness of the basic conformal predictor under different conditions.

3.1 Terminologies and Basic Procedures

Conformal prediction is a general framework that is model-agnostic, distribution-free, and produces prediction sets or intervals with guaranteed coverage rates for unobserved examples based on available data. The method is attractive as it can be combined with many black box algorithms to produce prediction sets with an automatic guarantee of coverage rate with minimal assumptions. To illustrate terminologies and basic procedures of conformal prediction, we introduce the full conformal prediction (FCP) or equivalently the transductive conformal prediction (TCP) which is the first and most basic

version of conformal prediction proposed by Vladimir Vovk and his collaborators in 2002 [40].

For each *example* $z_i \in Z$, we denote its *object* as $x_i \in X$, its *label* as $y_i \in Y$, where X is a non-empty measurable space, Y is a measurable space with at least two essentially different elements, and the example space $Z := X \times Y$ is a Cartesian product of X and Y. We denote a multi-set or bag of examples by $\wr z_1, ..., z_n \wr$. With a training set $\wr z_1, ..., z_n \wr$, we aim to predict for an unknown example Z_{n+1} under the assumption that all examples $Z_i = (X_i, Y_i)$ for $i = 1, ..., n+1$ are *exchangeable*.

Definition 1. *The variables $z_1, \ldots, z_N$ are exchangeable if, for every permutation τ of the integers $1, \ldots, N$, the variables $w_1, \ldots, w_N$, where $w_i = z_{\tau(i)}$, have the same joint probability distribution as $z_1, \ldots, z_N$.*

Note that the default randomness assumption for conformal prediction (i.e. exchangeability) is weaker than the standard i.i.d. assumption in machine learning which requires examples to be drawn from a power probability distribution P^∞ in Z^∞ where P is the unknown probability distribution of examples on Z. This is obvious because exchangeable variables are not necessarily independent of each other and, in the finite-sample case, there are many examples of exchangeable distribution Q on Z^N that is not of the form P^N. However, for the case of infinite-sample, the differences between the two assumptions become negligible according to the well-known De Finetti's theorem [19].

Theorem 1. *(De Finetti's Theorem) Each exchangeable probability distribution on Z^∞ is a mixture of power probability distribution P^∞, provided Z is a Borel space.*

Under the above setting, for any chosen miscoverage level $\alpha \in (0, 1)$, FCP gives a prediction set $\hat{C}_{n,\alpha}(X_{n+1})$ for the unknown label Y_{n+1} with following steps. Firstly, we choose a *nonconformity measure* $A : Z^* \times Z \to R$ which is a measurable function that measures the *nonconformity score* for each example in the training set as

$$\alpha_i := A(\wr z_1, ..., z_n \wr \setminus \wr z_i \wr, z_i) \text{ for } i = 1, ..., n.$$

The nonconformity score for the future example with a trial label y is denoted as

$$\alpha^y := A(\{z_1, ..., z_n\}, (X_{n+1}, y)) \text{ for each } y \in Y.$$

A higher nonconformity score indicates that the corresponding example is less conforming (i.e. 'stranger') compared to all examples in the bag. As its definition already implies, the nonconformity measure is assumed to be symmetric with respect to its first entry. In other words, the permutation of examples in the bag does not influence the nonconformity score outputs. A standard example of nonconformity measure is of the following form

$$A(\{z_1, ..., z_{i-1}, z_{i+1}, ..., z_n\}, z_i) := \Delta(\hat{h}(\{z_1, ..., z_{i-1}, z_{i+1}, ..., z_n\}), y_i),$$

where $\hat{h}$ is the underlying prediction model trained on $\{z_1, ..., z_{i-1}, z_{i+1}, ..., z_n\}$ and Δ is a metric. More examples and discussions on nonconformity scores will be shown in section 3.1.1.

Based on the nonconformity scores $\alpha_1, ..., \alpha_n, \alpha^y$ calculated as above, for each potential label $y \in Y$, we define the p-value for the unobserved example $z_{n+1} = (x_{n+1}, y)$ with a provisional label y as the fraction of examples that conform worse than or the same as z_{n+1}:

$$p^y := \frac{|\{i = 1, ..., n : \alpha_i \geq \alpha^y\}| + 1}{n + 1},$$

or as the smoothed p-value

$$p^y := \frac{|\{i = 1, ..., n : \alpha_i > \alpha^y\}| + \tau(|\{i = 1, ..., n : \alpha_i = \alpha^y\}| + 1)}{n + 1},$$

where τ is random variable uniformly distributed on $[0, 1]$. The difference between the two definitions is that the smoothed version treats ties more carefully by introducing a tie-breaking random variable τ to break ties uniformly. The smoothed version guarantees a coverage rate of exactly $1 - \alpha$, while the normal version guarantees that the coverage rate is at least $1 - \alpha$. Finally, all labels $y \in Y$ which gives a p-value p^y greater than the

miscoverage level α make up the conformal prediction set

$$\hat{C}_{n,\alpha}(X_{n+1}) := \{y | p^y > \alpha\}.$$

If the conformal prediction was used under an online setting [43], we assume that the true label of new example Y_{n+1} becomes known to us after we made the prediction and then enlarge the training set as $\{Z_1, ..., Z_n, Z_{n+1}\}$ for the next prediction. On the other hand, there is also much literature studying conformal prediction under the batch mode [20] where the size training sets remain unchanged.

Algorithm 1 Full Conformal Prediction [34]

Input: Examples $(x_i, y_i), i = 1, \ldots, n$; significance level $\alpha \in (0, 1)$; nonconformity measure A; object of the new example x_{n+1}; trial label y.
Output: Prediction interval $\hat{C}_{n,\alpha}(X_{n+1})$ for label of the new example Y_{n+1}.
For each trial label y, set provisionally $z_{n+1} = (x_{n+1}, y)$.
For each $z_i, i = 1, \ldots, n$ calculate its nonconformity score as

$$\alpha_i := A(\{z_1, \ldots, z_n\} \setminus \{z_i\}, z_i).$$

For new example (x_{n+1}, y), calculate its nonconformity score as

$$\alpha^y := A(\{z_1, \ldots, z_n\}, (x_{n+1}, y)).$$

For each provisional new example obtain its p-value as

$$p^y := \frac{|\{i = 1, \ldots, n | \alpha_i \geqslant \alpha^y\}| + 1}{n + 1}.$$

Include y in $\hat{C}_{n,\alpha}(X_{n+1})$ if and only if $p^y > \alpha$.

Note that in practice it may be unrealistic for practitioners to calculate p-value for every possible value of y. For example, in regression setting $Y = \mathbb{R}$, the algorithm can be implemented by checking test values of y on a fine grid and still preserves coverage guarantee [10]. Also, it is exactly the same to choose a conformity measure B at the beginning and then define the p-value as the fraction of conformity scores that is smaller than or equal to the conformity score of the new example. For example, we can set B as $1/A$ or $1 - A$ which will produce the same prediction set as using the nonconformity

measure A. In fact, the p-value in conformal prediction is consistent with the widely accepted concept of p-value in Neyman-pearson theory [34]. The prediction problem can be expressed as the following hypothesis test:

• **Null Hypothesis** : the bag of the first $n+1$ examples is $\wr z_1, ..., z_n, z_{n+1}\wr$ where $z_{n+1} = (x_{n+1}, y)$.

• **Test Statistic T** : the random value of nonconformity score of z_{n+1}.

Under the null hypothesis, T is equally likely to take any value of α_i. Thus, we obtain the p-value as

$$p_H = P(T \geq \alpha^y | \wr z_1, ..., z_n, z_{n+1}\wr) = p^y.$$

For each label y in the conformal prediction set $\hat{C}_{n,\alpha}(X_{n+1})$, we have p-value $p_H = p^y > \alpha$. In other words, we do not reject the null hypothesis that y is the label of the new example at significance level α.

3.1.1 Nonconformity Measure

The design of nonconformity measures is the core of a successful conformal prediction algorithm. A conformal prediction algorithm can be viewed as a wrapper that combines with nearly all machine learning algorithms. Although the validity of CP is automatically guaranteed regardless of what underlying model we adopt, a poor nonconformity measure will result in large prediction sets that convey no information to us. To obtain meaningful prediction result, we expect $\alpha_i \propto P(z_i \notin Z^*)$ and α^y follows the same distribution as $\alpha_1, ..., \alpha_n$. In this section, we provide readers with some common choices of nonconformity score functions for conformal predictors based on different underlying prediction rules under either regression or classification settings.

In general, nonconformity measures can be categorized into two groups: model-agnostic ones and model-dependent ones [1]. For regression problems, a basic choice of model-agnostic nonconformity score follows the scheme we introduced above with Δ chosen to

be the L1 norm

$$\alpha_i := \Delta(\hat{h}(\wr z_1, ..., z_n \wr), y_i) = |\hat{y}_i - y_i|,$$

which can be interpreted as the absolute residual, or in some cases as the absolute value of the deleted residual

$$\alpha_i := \Delta(\hat{h}(\wr z_1, ..., z_{i-1}, z_{i+1}, ..., z_n \wr), y_i) = |\hat{y}_{(-i)} - y_i|,$$

where $\hat{y}_{(-i)}$ indicates that the label is predicted by the prediction algorithm trained on the training dataset with z_i deleted.

Ridge Regression

One example in this category is the ridge regression confidence machine (RRCM) [43] which uses the following nonconformity scores:

$$\alpha_i = |e_i| = |\mathbf{x}_i'(\mathbf{X}_n'\mathbf{X}_n + a\mathbf{I}_p)^{-1}\mathbf{X}_n'\mathbf{y} - y_i|,$$

where $\mathbf{X}_n$ denotes the $n \times p$ design matrix, $\mathbf{y}$ denotes the vector of observed labels, a denotes the ridge parameter, and $\mathbf{I}_p$ denotes the identity matrix. Residual is a natural choice of nonconformity measure in a regression setting since labels of unusual examples will deviate more from their predicted labels (i.e. larger nonconformity scores). However, the drawback of this method is also apparent. All prediction intervals will have the same length in this way. To improve adaptivity, another choice of nonconformity measure of RRCM is based on studentized residuals:

$$\alpha_i = \frac{|e_i|}{\sqrt{1 - h_{ii}}},$$

where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}_n(\mathbf{X}_n'\mathbf{X}_n + a\mathbf{I}_p)^{-1}\mathbf{X}_n'$. Such modification allows instance-wise treatment and equalizes the variances of residuals to

produce more efficient predictions in certain cases [43]. More variants of locally weighted nonconformity measures can be found in [20] [29].

Nearest Neighbour Regression

For conformal predictors based on the k-nearest neighbors (kNN) algorithm, we can also construct nonconformity measures from the idea of absolute error. For regression, the simplest implementation would be predicting the label of each example as the mean or median of all the labels of its k-nearest neighbors [43]. In other words, given example (x_i, y_i) we find its k nearest neighbors $(x_{i_1}, y_{i_1}), ..., (x_{i_k}, y_{i_k})$ whose objects have the shortest distance from x_i among other training examples based on the chosen metric. Then, the predicted label denoted by $\hat{y}_{(-i)}$ can be either the mean or median of $y_{i_1}, ..., y_{i_k}$ based on the user's choice. Similar to the above, the nonconformity score can be expressed as

$$\alpha_i := |y_i - \hat{y}_{(-i)}|.$$

Moreover, [31] proposed several ways to normalize this nonconformity measure by assessing the expected accuracy of the nearest neighbor-based conformal prediction algorithm.

Conformalized Quantile Regression

As the last example for regression, we introduce the nonconformity measure adopted by conformalized quantile regression (CQR) which is model-dependent and quantile-based and differs completely from the previous examples [13]. CQR combines quantile regression method with inductive conformal prediction which divides the training data into proper training and calibration sets and will be introduced in details in Section 3.3.1. First, two point predictors $\hat{f}^{\alpha/2}$and $\hat{f}^{1-\alpha/2}$ which estimate the $\alpha/2$ and $1-\alpha/2$ quantile of $Y|X = x$ respectively are trained on the proper training set. Then, nonconformity scores

are calculated for each example in the calibration set as follows:

$$R_i = max\{\hat{f}^{\alpha/2}(X_i) - Y_i, Y_i - \hat{f}^{1-\alpha/2}(X_i)\}.$$

Finally, the CQR prediction set becomes

$$\hat{C}_{n,\alpha}(x) = [\hat{f}^{\alpha/2}(x) - \hat{q}, \hat{f}^{1-\alpha/2}(x) + \hat{q}],$$

where $\hat{q} =$ the $\lceil (1-\alpha)(n+1) \rceil$ smallest value of R_i.

For conformal classification algorithms, [23] proposed several choices of model-agnostic nonconformity measures that can be used by neural network-based conformal classifiers such as the inverse probability or hinge loss

$$\Delta(\hat{h}(x_i), y_i) := 1 - \hat{P}_h(y_i|x_i),$$

where $\hat{P}_h(y_i|x_i)$ represents the amount of probability that underlying classifier h assigns to label y_i given that object is x_i and the margin nonconformity function

$$\Delta(\hat{h}(x_i), y_i) := \max_{y \neq y_i} \hat{P}_h(y|x_i) - \hat{P}_h(y_i|x_i).$$

Empirical comparison of the performance of different model-agnostic nonconformity measures can be found in [1].

For model-dependent nonconformity measures, we provide the following two examples.

Nearest Neighbour Classification

While using 1-nearest neighbour algorithm, we define

$$\alpha_i := \frac{min_{j=1,\ldots,n:j\neq i \& y_j = y_i} \Delta(x_i, x_j)}{min_{j=1,\ldots,n:y_j \neq y_i} \Delta(x_i, x_j)},$$

where Δ is a metric on X. In this way, we assign smaller nonconformity scores to more conforming examples whose object is closer to objects of training examples with the same label and further from examples with different labels.

Support Vector Machine

Another example of a model-dependent nonconformity measure in the case of classification is constructed for conformal predictors based on support vector machine. Suppose we are dealing with a binary classification problem with $Y = \{-1, 1\}$ and we adopt support vector machine method to find the optimal hyperplane on training data $z_1, z_2, ..., z_n$ to separate the two classes where objects are assumed to be vectors in a dot product space H. Then, this problem can be expressed as follows [13]:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C(\sum_{i=1}^{n} \xi_i), \tag{3.1}$$

subject to the constraints

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, i = 1, ..., n \tag{3.2}$$

$$\xi_i \geq 0, i = 1, ..., n, \tag{3.3}$$

where C is fixed positive constant, $w \in H$, and $\xi_i \in R$ handles margin violations. To solve this optimization problem using Lagrange multipliers we consider its dual problem and adopt a kernel trick. It suffices to find Lagrange multipliers α_i's for the following optimization problem:

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{j=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j), \tag{3.4}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, i = 1, ..., n \tag{3.5}$$

where $K(x_i, x_j) := F(x_i)$ with $F(x_j), F : X \to H$ mapping objects x_i into a dot product space H with $F(x_i) \in H$. In this way, the Lagrange multipliers α_i become natural choices for the nonconformity score of each example as they depict the property of each example with $\alpha_i = 0$ indicating example z_i is among the most conforming examples while $\alpha_i = C$ identifying example z_i to be one of the most nonconforming ones.

3.2 Evaluation Criteria

With the knowledge of the procedures of conformal prediction and some strategies for constructing nonconformity measures, we have enough ingredients to build different conformal predictors. However, it remains a problem for one to compare and choose among various conformal predictors. In this section, we will discuss the two most important evaluation criteria: validity and efficiency.

Validity

The most powerful characteristic of conformal prediction that distinguishes it from other prediction methods is its ability to guarantee a confidence level we specified initially. The confidence mentioned here is consistent with the concept of confidence interval as conformal prediction intervals guarantee to cover the true value of label y with the desired coverage rate on average. However, it is also important to note that basic conformal prediction algorithm lacks the ability to ensure within-category coverage, which is an important property in some applications. We formally presented the two types of validity in this section and discussed the limitations and ability of CP algorithms to achieve these goals.

To discuss the validity of conformal prediction sets, we use the same terminologies, notations, and definitions that were used in [43] in this section. For a given miscoverage rate ϵ, we would like the prediction intervals to be correct $100(1-\epsilon)\%$ of the time. Under the online setting, we will look at a sequence of successive predictions and we would

expect that on average $100(1-\epsilon)\%$ of these prediction intervals cover the true value. More formally, for a data set $\omega = ((x_1, y_1), (x_2, y_2), ...)$, we record the error of the nth prediction trial made by the conformal predictor Γ at significance level ϵ as follow:

$$err_n^\epsilon(\Gamma, \omega) := \begin{cases} 1 & \text{if } y_n \notin \Gamma^\epsilon(x_1, y_1, .., x_{n-1}, y_{n-1}, x_n) \\ 0 & \text{otherwise,} \end{cases}$$

and denote the total number of errors made in the first n predictions by $Err_n^\epsilon(\Gamma, \omega) := \sum_{i=1}^n err_i^\epsilon(\Gamma, \omega)$.

Definition 2. *If ω is generated from an exchangeable probability P on Z^∞, then the numbers $err_i^\epsilon(\Gamma, \omega)$ for i=1,2,... are realized values of random variables $err_i^\epsilon(\Gamma, P)$. A confidence predictor Γ is exactly valid if for each $\epsilon \in (0, 1)$, $err_1^\epsilon(\Gamma, P)$,$err_2^\epsilon(\Gamma, P)$,... is a sequence of independent Bernoulli random variables with parameter equals ϵ.*

A confidence predictor is *conservatively valid* if $err_n^\epsilon(\Gamma, P)$ is dominated in distribution by a sequence of independent Bernoulli random variables with parameter ϵ.

Definition 3. *The confidence predictor is asymptotically exact if for any exchangeable probability P on Z^∞ and any given significance level ϵ, $\lim_{n\to\infty} \frac{Err_n^\epsilon(\Gamma,P)}{n} = \epsilon$ with probability 1.*

Similarly, a confidence predictor is *asymptotically conservative* if for any exchangeable probability distribution P on Z^∞ and any significance level ϵ,

$$\limsup_{n\to\infty} \frac{Err_n^\epsilon(\Gamma, P)}{n} \leq \epsilon,$$

with probability 1. However, it is impossible to have an exact conformal predictor [43, Theorem 2.1]. Thus, the author introduces the definition of conservative validity which requires the sequence of $err_i^\epsilon(\Gamma, \omega)$ to be dominated in distribution by a sequence of independent Bernoulli random variables with parameter ϵ. To obtain an exactly valid conformal predictor, the author introduces randomized values $\tau_1, \tau_2, ...$ that are independently

and randomly drawn from uniform [0,1] distribution to modify the conformal predictor. This new predictor is called a smoothed confidence predictor which is different from the conformal predictor by treating borderline cases (i.e. $\alpha_i = \alpha_n$) more carefully. The p-value of a smoothed conformal predictor is defined as

$$p_y := \frac{|i = 1, ..., n : \alpha_1 > \alpha_n| + \tau_n |i = 1, .., n : \alpha_i = \alpha_n|}{n}. \tag{3.6}$$

The prediction set includes all values of $y \in Y$ such that $p_y > \epsilon$. In this way, we obtained an exactly valid predictor.

For a data sequence ω drawn from an exchangeable distribution, we denote the conformal predictor by Γ and the corresponding smoothed conformal predictor which uses the same significance level ϵ and the same nonconformity score by Γ'. Then for each $y \in Y$, we will have $p'_y < p_y$ and thus $\Gamma^{\epsilon'}(\omega) \subseteq \Gamma^{\epsilon}(\omega)$ and $err_n \leq err'_n$. If the smoothed conformal predictor is exactly valid then the corresponding conformal predictor which uses the same nonconformity score and significance level is conservatively valid by definition. Thus, it suffices to show that smoothed conformal predictors are exactly valid. Intuitively, since the examples are drawn from an exchangeable distribution, their nonconformity scores are also exchangeable. Note that p-value p_y for the new example $z = (x_n, y)$ is determined by the rank of α_n among all values of $\alpha_1, ..., \alpha_n$ which would be equally likely to take any value in $\{1,...,n\}$. Thus, p_y of a smooth conformal predictor is uniformly distributed in [0,1] and err_n^{ϵ} are Bernoulli random variables that equal 1 with probability ϵ. Also, random variables err_n^{ϵ} are independent [43, Prop 1]. Finally, we can conclude that at least (1-ϵ) of the predictions given by a conformal predictor covers the true value. [43, Theorem 8.2] presents a fully rigorous proof of the validity of smooth conformal predictors. Moreover, it is obvious that an exact or conservative confidence predictor is also asymptotically exact or conservative respectively by the law of large numbers [43].

Efficiency

Once the validity of prediction interval is guaranteed, efficiency becomes the most important property we would like conformal predictors to optimize. In other words, we want the smallest prediction interval for a given significance level. Although in the previous subsection, we observed that conformal predictor is powerful as it always produces valid prediction intervals only requiring exchangeable assumption and continuity on the nonconformity measure, we should note that a poor underlying prediction algorithm will greatly reduce the efficiency of conformal prediction. Some articles are devoted to formally defining the efficiency of conformal prediction algorithms. For example, in [21] the authors proposed one natural notion to assess the efficiency of prediction set by defining it as the prediction result's closedness to the oracle prediction set $C^{or} := argmin_C \Lambda(C)$ where C ranges over the measurable subsets of Z such that $Q(C) \geq 1 - \epsilon$ where Q represents the data generating distribution on Z and Λ represents the Lebesgue measure on $\mathbb{R}^d$. The closedness of a prediction set Γ^ϵ and C^{or} is defined as $\Lambda(\Gamma^\epsilon \triangle C^{or})$.

3.3 Calibration Methods

From the previous section, we see that full conformal prediction utilizes all information from the whole dataset to make future predictions. Although such design is the most statistically efficient one and enables elegant theoretical proofs for its properties, it suffers from remarkably high computational cost which forbids its usage in many situations. Thus, it is of great motivation to improve FCP by sampling the training dataset into proper training sets and calibration sets. In other words, how to split the limited data available to us and create prediction sets to achieve a good balance among validity, efficiency, and computation cost become an interesting research topic. In this section, several strategies for data splitting will be discussed.

3.3.1 Split Conformal Prediction

While being very straightforward from a mathematical point of view, full conformal predictors are very computationally intensive, and, a lot of literature was aiming to address this issue. In particular, inductive conformal predictors (ICP), also referred to as split conformal predictors are the first and most popular ones to be proposed as an important computationally efficient alternative method. As implied by the names of these two methods, the ideas behind ICP and TCP are rooted in two important concepts in machine learning: transductive learning and inductive learning [38]. The two concepts delineate the pathways through which algorithms learn and generalize from data. Inductive learning operates on the principle of generalization from specific instances to broader rules, applicable to unseen data. In contrast, transductive learning focuses on making predictions for a specific set of unseen data, emphasizing direct inference for these instances rather than a generalized rule applicable across all potential data points. Consistent with these fundamental concepts, full conformal prediction creates confidence measures with the most statistical efficiency while suffering great computation costs when the training dataset is large [43]. Conversely, inductive conformal prediction algorithms embrace the ethos of inductive learning by utilizing a two-phase process—initial calibration on a subset of data to establish prediction confidence levels, followed by the application of these calibrated models to generalize across unseen data, thus providing a faster algorithm [34].

Inductive conformal prediction was first proposed in 2002 [30] for regression problems. To be consistent with the previous section, we introduce the general split conformal prediction algorithm under the online setting. Given the set $(z_1, ..., z_{n-1})$ of n-1 observed examples, where $z_i = (x_i, y_i)$ with x_i being the object and y_i being the label. To implement an ICP for prediction of new example z_n with knowledge of its object x_n, first, we need to define a finite or infinite sequence of *update trial* $m_1, m_2, ...$ which are positive integers in ascending order. Define strangeness measure $\{A_n\}_{n=1}^{\infty}$ for ICP as follows:

$$(\alpha_1, ..., \alpha_n) = A_n((x_1, y_1), ..., (x_n, y_n)),$$

where the nonconformity scores are defined by

$$\alpha_i := \Delta(y_i, D_{\wr(x_1,y_1),...,(x_n,y_n)\int}(x_i))$$

or

$$\alpha_i := \Delta(y_i, D_{\wr(x_1,y_1),...,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),...(x_n,y_n)\int}(x_i)),$$

where $D_{\wr(x_1,y_1),...,(x_n,y_n)\int} : \mathbf{X} \to \hat{\mathbf{Y}}$ is an inductive algorithm that maps a bag of examples $\wr(x_1, y_1), ..., (x_n, y_n)\int$ to a function which is the decision rule, $\Delta : \mathbf{Y} \times \hat{\mathbf{Y}} \to \mathbb{R}$ is the discrepancy measure that measures the discrepancy between the predicted label $\hat{y}_i$ and true label y_i. Note that $\hat{\mathbf{Y}}$ may not equal $\mathbf{Y}$ as for some rare cases the predicted label contains additional information. The inductive conformal prediction sets determined by the update trial $(m_1, m_2, ...)$ and nonconformity measure $\{A_n\}_{n=1}^{\infty}$ can be obtained as follows:

- **For** $n \leq m_1$**,** $\Gamma^{\epsilon}(x_1, y_1, ..., x_{n-1}, y_{n-1}, x_n)$ is found using a fixed transductive conformal predictor.

- **For** $n > m_1$**, ,** find k such that $m_k < n \leq m_{k+1}$ and set

$$\Gamma^{\epsilon}(x_1, y_1, ..., x_{n-1}, y_{n-1}, x_n) := \{y \in \mathbf{Y} : \frac{|\{j = m_k + 1, ..., n : \alpha_j \geq \alpha_n\}|}{n - m_k}\},$$

  where the nonconformity scores α_j are defined by

$$\alpha_j := A_{m_k+1}(\wr(x_1, y_1), \ldots, (x_{m_k}, y_{m_k})\int, (x_j, y_j)), \text{for } j = m_k + 1, \ldots, n - 1,$$

$$\alpha_n := A_{m_k+1}(\wr(x_1, y_1), \ldots, (x_{m_k}, y_{m_k})\int, (x_n, y)).$$

Note that the most important type of ICP which randomly splits the data once and uses $(z_1, ..., z_{m_1})$ as a training set, $(z_{m_1+1}, ..., z_{n-1})$ as calibration set can be expressed by setting the update trial as $(m_1, \infty, \infty, ...)$. We can also define a randomized or smoothed inductive conformal predictor similarly as the randomized transductive conformal predictor

by introducing a random component into the definition of p-values which handles the borderline cases in a more refined way. The prediction set of smoothed ICPs is defined as follows:

$$\Gamma^{\varepsilon}((x_1,\tau_1,y_1),\ldots,(x_n,\tau_n)) := \left\{ y \in Y : \frac{|\{j : \alpha_j > \alpha_n\}| + \tau_n |\{j : \alpha_j = \alpha_n\}|}{n - m_k} > \varepsilon \right\},$$

where $j = m_k + 1, ..., n$ and $\tau_n \in [0, 1]$ are the random numbers. One example of ICP with one update trial member (one even split) applicable to the regression problem is provided below.

Algorithm 2 Split Conformal Prediction [25]

Input: Data $(X_i, Y_i), i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$, regression algorithm A, object x.
Output: Prediction interval $C_{\text{split}}(x)$.
Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets I_1, I_2.
$\hat{\mu} = A(\{(X_i, Y_i) : i \in I_1\})$.
$R_i = |Y_i - \hat{\mu}(X_i)|, i \in I_2$.
$d =$ the kth smallest value in $\{R_i : i \in I_2\}$, where $k = \lceil (n/2 + 1)(1 - \alpha) \rceil$.
$C_{\text{split}}(x) = [\hat{\mu}(x) - d, \hat{\mu}(x) + d],$ for all $x \in \mathbb{R}^d$.

Validity of ICPs

The biggest advantage that makes ICP an extremely popular choice among all conformal predictors is that ICP also has an automatic guaranteed coverage rate. Actually, the inductive conformal predictor can be viewed as a special case of the full conformal predictor. For example, for regression problem, consider a trivial algorithm A that returns the same fixed pre-fitted function $\hat{\mu}$ regardless of the input data. Thus, the validity guarantee of ICPs is just a special case from the general theorem 8.2 in [43].

Theorem 2. *All ICPs are conservatively valid. All smoothed ICPs are exactly valid.*

If we further assume residuals obtained from the calibration set have a continuous joint distribution, then [20, Theorem 2] shows that the unconditional coverage of ICPs is

also bounded above similar to the TCPs.

$$P(Y_{n+1} \in C_{\text{split}}(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2}$$

Asymptotic validity is automatically guaranteed [43].

While unconditional coverage guarantee is automatically achieved under the exchangeable assumption, the realization of conditional coverage in most cases requires modifications to the basic algorithm. We will examine several conditional validity criteria on ICP with only one data split in this section. Firstly, we examine if ICP holds predictive coverage after conditioning on the training data set $D_n = (z_1, .., z_n)$. For training conditional coverage, [41] establishes a PAC-type 2 parameter definition to formalize this property. Define the miscoverage rate as a function of D_n

$$\alpha_P(D_n) = \mathbb{P}_P\{Y_{n+1} \notin \hat{C}_n(X_{n+1})|D_n\}.$$

Theorem 1 only ensures the unconditional coverage rate:

$$\mathbb{P}_{P^{n+1}}\{Y_{n+1} \in C_{\text{split}}(X_{n+1})\} \geq 1 - \alpha \leftrightarrow \mathbb{E}_{P^n}[\alpha_P(D_n)] \leq \alpha.$$

For the guarantee of training conditional coverage rate we need to show

$$\mathbb{P}_{P^n}\{\alpha_p(D_n) > \alpha + o(1)\} \leq o(1).$$

Vovk provided proof of the training-conditional coverage through Hoeffding inequality [41, prop 2a].

Theorem 3. *Consider the split conformal method defined with sample size $n = n_0 + n_1$, where $n_0 \geq 1$ many data points are used for training the fitted model $\hat{\mu}_{n_0}$ (with an arbitrary algorithm) while the remaining $n_1 \geq 1$ data points are used as the holdout (calibration) set. Then, for any*

distribution P and any $\delta \in (0, 0.5]$,

$$\mathbb{P}_{P^n}\left(\alpha_P(D_n) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2n_1}}\right) > 1 - \delta.$$

The probability that a training set results in a significantly higher training-conditional miscoverage rate than the nominal rate is vanishing small for split conformal prediction. Note that this result holds for both ICP and smoothed ICP. By choosing $\alpha' := \alpha - \sqrt{\frac{log(1/\delta)}{2n_1}}$ while performing ICP, we would obtain a slightly more conservative prediction interval that satisfies

$$\mathbb{P}_{P^n}\{\alpha_P(D_n) \leq \alpha\} \geq 1 - \delta.$$

This process is called a probably approximately correct (PAC) guarantee.

For label conditional validity, [34] proposed conditional ICP which modifies the split conformal prediction to guarantee label conditional validity for classification problems. Suppose $\wr z_1, ..., z_m \wr$ is the proper training set and $\wr z_{m+1}, ..., z_l \wr$ is the calibration set. Define an inductive m-taxonomy as a measurable function $K : \mathbf{Z}^m \times \mathbf{Z} \to \mathbf{K}$, where $\mathbf{K}$ is a measurable space. Usually the category $K((z_1, \ldots, z_m), z)$ of an example z is a kind of classification of z, which may depend on the proper training set $(z_1, \ldots, z_m)$. The conditional inductive conformal predictor (conditional ICP) corresponding to K and an inductive conformity measure A is defined as the set predictor

$$\Gamma^\epsilon(\wr z_1, \ldots, z_l \wr, x) := \{y : p^y > \epsilon\},$$

where the p-values p^y are now defined by

$$p^y := \frac{|\{i = m+1, \ldots, l \mid \kappa_i = \kappa^y \& \alpha_i \leq \alpha^y\}| + 1}{|\{i = m+1, \ldots, l \mid \kappa_i = \kappa^y\}| + 1}, \tag{3.7}$$

the categories κ are defined by

$$\kappa_i := K((z_1, \ldots, z_m), z_i), \quad i = m+1, \ldots, l, \quad \kappa^y := K((z_1, \ldots, z_m), (x, y)), \tag{3.8}$$

and the conformity scores α are defined as

$$\alpha_i := A((z_1, ..., z_m), z_i), i = m+1, ..., l \text{ , and } \alpha^y := ((z_1, ..., z_m), (x, y)).$$

A label-conditional ICP with the inductive taxonomy K produces prediction sets with guaranteed label conditional coverage rate [34, prop 3]

$$P\{Y_{l+1} \in \Gamma^\epsilon(\wr z_1, \ldots, z_l \int, X_{l+1}) | K((z_1, \ldots, z_m), z_{l+1})\} \geq 1 - \epsilon.$$

The method is suitable for cases where it is important to achieve coverage within each category of labels.

Finally, for object-conditional coverage, unfortunately, it is impossible to achieve a distribution-free object-conditional coverage guarantee as follows

$$P\{Y_{l+1} \in \Gamma^\epsilon(\wr z_1, \ldots, z_l \int, X_{l+1}) | X_{l+1} = x\} \geq 1 - \epsilon$$

as there may be a discontinuity at X=x for the distribution P. However, [25] shows that ICPs satisfy *local validity* which is an original concept invented as a measure of coverage that lies between marginal and conditional validity. Also, the authors proposed that it is possible to construct asymptotic efficient and asymptotic conditional valid inductive conformal predictors. [15] explores more possibilities on the gap between conditional and unconditional validity for ICPs as well as for FCPs. To obtain exact conditional coverage, one needs to add some kind of smoothness assumption on the distribution P such as the smoothness condition proposed in [12].

Computation Cost and Prediction efficiency

Compared to TCP which requires re-training of the underlying prediction algorithm at each prediction with full observed data, ICP achieves a great reduction in computation cost as it only requires re-training of the prediction algorithm at each update trial. The computation cost of ICP is always cheaper than the computation cost of a corresponding TCP. Thus, one advantage of ICP is that we can combine it with computationally heavy estimators [14]. Exceptions exist in some special cases where computation tricks can be applied to TCPs to reduce computation cost such as [24] which chooses Lasso to construct nonconformity measures.

While validity is taken for granted in conformal framework, efficiency is related to the accuracy of the underlying algorithm. Achieving computational efficiency does not come for free. A drawback of inductive conformal predictors is their lack of prediction efficiency. We waste the calibration set when developing the prediction rule, and we sacrifice the proper training set when computing the p-values. One way to solve this disadvantage is Cross-conformal prediction [42], a hybrid of the methods of inductive conformal prediction and cross-validation. Other CP methods providing tradeoff between computation efficiency and prediction efficiency such as the Jackknife method [41] will be introduced in the following sections.

3.3.2 Aggregated Conformal Predictors

While a single split enables better computation efficiency and preserves validity property with minimal assumptions, the method suffers from information loss. The problem can become more serious when the dataset is small because small proper training set would result in an inaccurate underlying prediction algorithm. To overcome such issues, extensions of ICPs have been developed among which an important group of conformal predictors are named aggregated conformal predictors. These conformal predictors were invented based on the concept of ensemble learning in machine learning by dividing

the training data into multiple proper training sets and calibration sets. Formally, denote the training set by $\zeta = (z_1, ..., z_n) \in Z^n$ and suppose we sample K subsets (folds) ζ_k^t for $k = 1, ..., K$ as proper training sets and ζ_k^c as the corresponding calibration sets by using any sampling strategy that satisfying ζ_k^c is exchangeable with ζ for all $k \in \{1, ..., K\}$. Nonconformity scores for each example can be calculated from each (ζ_k^t, ζ_k^c) pair and for each fold we obtain a p-value for test example $z_{n+1} = (x_{x+1}, y)$. Apparently, there are two fundamental problems for this group of methods: which sampling strategy to choose and how to combine the multiple fitted algorithms or p-values obtained from each pair of proper training and calibration sets to get one prediction result in the end. In this section, we will first introduce some typical conformal predictors in this category and discuss the overall behavior with generalized aggregated conformal predictors.

Cross-Conformal Predictor (CCP)

One of the most prevalent conformal predictors under this category is the cross-conformal predictor introduced by Vovk [42]. It merges inductive conformal prediction with the idea of cross-validation to provide valid and efficient predictions for each individual model as well as for their combinations. Similar to the widely accepted cross-validation method, CCP first partitions the training set ζ into K non-empty subsets (folds) $\zeta_k^c, k = 1, ..., K$, where $K \in \{2, 3, ...\}$ is a parameter of this algorithm and $\zeta = U_{k=1}^K \zeta_k^c$. Denote the corresponding proper training sets as $\zeta_k^t := \zeta \setminus \zeta_k^c$. For every pair (ζ_k^t, ζ_k^c) where $k \in \{1, ..., K\}$ and each potential label $y \in Y$ of x_{n+1} calculate the nonconformity scores in the same way as an inductive conformal predictor. For each example $z_i \in \zeta_k^c$ and each provisional example (x_{n+1}, y), nonconformity scores are defined as follows

$$\alpha_{i,k} := A(\zeta_k^t, z_i), z_i \in \zeta_k^c; \alpha_k^y := A(\zeta_k^t, z_i), (x_{n+1}, y)).$$

The p-value for $y \in Y$ is defined by

$$p^y := \frac{\sum_{k=1}^{K} |\{z_i \in \zeta_k^c | \alpha_{i,k} \geq \alpha_k^y\}| + 1}{|\zeta| + 1},$$

which is essentially the average of the p-value for each fold p_k^k defined as follows

$$p_k^y := \frac{|\{z_i \in \zeta_k^c : \alpha_{i,k} \geq \alpha_K^y\}| + 1}{|\zeta_k^c| + 1}.$$

Note that when the K folds have equivalent sizes we have

$$p^y = \bar{p}^y + \frac{K-1}{|\zeta|+1}(\bar{p}^y - 1),$$

where $\bar{p}^y := \frac{1}{K}\sum_{k=1}^{K} p_k^y$ and $p^y = \bar{p}^y$ for $K < |\zeta|$. Usually, for a cross-validation method, one is suggested to take $K \in \{5, 10\}$ however CCP is not exactly a cross-validation and it remains an open problem for researchers to decide what value of K to take in practice [42]. The biggest difference between CCP and ICP is the former uses the entire training set ζ for calibration by adopting a cross-validation sampling strategy. In this way, CCP achieves better statistical efficiency while sacrificing theoretical guarantee on the automatic validity and some computation efficiency as we need to train the underlying model K times using CCP. Moreover, CCP obviously has better p-value stability compared to ICP as ICP only uses one random split of the whole training set ζ which could lead to very different p-values when applied to the same ζ multiple times.

Jackknife Method

One extreme case of the cross-conformal predictor is the Jackknife method or leave-one-out conformal predictor (LOOCP). LOOCP is equivalent to a cross-conformal predictor taking K equals the size of the training set n (i.e. split the training data into n subsets each containing one unique example from the training set). The nonconformity measure for LOOCP can be interpreted as leave-one-out residuals under regression setting and the

Algorithm 3 Cross-Conformal Predictors

Input: Training set $Z = \{(X_i, Y_i) : i = 1, \ldots, n\}$, significance level $\epsilon \in (0, 1)$, inductive nonconformity measure A, number of folds K.
Output: Prediction set $C_{n,\epsilon}(x)$ for a new object x.
Partition the training set into K non-empty subsets $Z_k^c, k = 1, \ldots, K$ and denote corresponding proper training set by Z_{-k}^t.
for each $k \in \{1, \ldots, K\}$ **do**
  for each potential label $y \in Y$ of x **do**
    Compute nonconformity scores for examples in each calibration set Z_k^c and (x, y) by
    $\alpha_i^k := A(Z_{-k}^t, z_i), z_i \in Z_k^c$,
    $\alpha_y^k := A(Z_{-k}^t, (x, y))$.
  end for
end for
Combine the conformity scores to compute p-values:
$p_y := \frac{1}{n+1} \sum_{k=1}^{K} \big(|\{z_i \in Z_k^c : \alpha_i^k \leq \alpha_y^k\}| + 1\big)$.
return $C_{n,\epsilon}(x) := \{y | p_y > \varepsilon\}$.

conformal prediction set in this case can be expressed as below

$$\hat{C}_{n,\alpha}^{\text{Jackknife}} = [\hat{q}_{n,\alpha}^{-}\{\hat{\mu}(X_{n+1}) - R_i^{\text{LOO}}\}, \hat{q}_{n,\alpha}^{+}\{\hat{\mu}(X_{n+1}) + R_i^{\text{LOO}}\}],$$

where $\hat{\mu}$ is the underlying regression algorithm trained on the training set ζ, $R_i^{\text{LOO}} := |\hat{\mu}_{-i}(X_i) - Y_i|$are the leave-one-out residuals for each $z_i \in \zeta$, $\hat{q}_{n,\alpha}^{+}$ and $\hat{q}_{n,\alpha}^{-}$ denote respectively the $1 - \alpha$ and α quantile defined as below

$$\hat{q}_{n,\alpha}^{+}\{v_i\} = \text{the } \lceil (1-\alpha)(n+1) \rceil \text{-th smallest value of } v_1, ..., v_n,$$

$$\hat{q}_{n,\alpha}^{-}\{v_i\} = \text{the } \lfloor \alpha(n+1) \rfloor \text{-th smallest value of } v_1, ..., v_n.$$

Jackknife method is more efficient than ICP as it provides smaller prediction sets [20]. However, LOOCP has only finite sample in-sample coverage property [20] and no automatic out-of-sample validity guarantee. Also, it requires K times training of the underlying algorithm which can be computationally costly. The predictive accuracy of the jackknife under assumptions of algorithm stability is explored by [35] for the linear regres-

sion setting. Hence, while the full and split conformal intervals are valid under minimal assumptions, the same is not true for the Jackknife ones.

Jackknife+ and CV+

It is noteworthy that a modification to the Jackknife method as well as to CCP with a rigorous theoretical guarantee on out-of-sample coverage guarantee was proposed recently in [5]. Using the same notations as above Jackknife+ prediction interval is defined as

$$\hat{C}_{n,\alpha}^{\text{Jackknife+}} = [\hat{q}_{n,\alpha}^{-}\{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\}, \hat{q}_{n,\alpha}^{+}\{\hat{\mu}_{-i}(X_{n+1}) + R_i^{\text{LOO}}\}].$$

Jackknife+ is able to achieve a better coverage rate than the standard Jackknife method when the underlying regression algorithm is not stable. Moreover, [5] proved that this modified version achieves a theoretical guarantee of $1 - 2\alpha$ coverage rate with minimal assumption.

Algorithm 4 Jackknife+ Method [5]

Input: Training data $\{(X_i, Y_i)\}_{i=1}^n$, regression algorithm $\mathcal{A}$, test point X_{n+1}, miscoverage rate α.
Output: Predictive confidence interval $\hat{C}_{n,\alpha}(X_{n+1})$ for Y_{n+1}.
for $i = 1$ to n **do**
  Fit regression algorithm$\hat{\mu}_{-i} = \mathcal{A}\left(\{(X_j, Y_j)\}_{j \neq i}\right)$.
  Calculate leave-one-out residuals $R_i^{\text{LOO}} = |Y_i - \hat{\mu}_{-i}(X_i)|$.
end for
Return: Prediction interval

$$\hat{C}_{n,\alpha}(X_{n+1}) = \left[\hat{q}_{n,\alpha}^{-}\left\{\hat{\mu}_{-i}(X_{n+1}) - R_i^{\text{LOO}}\right\}, \hat{q}_{n,\alpha}^{+}\left\{\hat{\mu}_{-i}(X_{n+1}) + R_i^{\text{LOO}}\right\}\right].$$

The author also proposed another variant of CP named the Jackknife-minmax method which produces prediction interval

$$\hat{C}_{n,\alpha}^{\text{jackknife-mm}} = [\min_{i=1,..,n} \hat{\mu}_{-i}(X_{n+1} - \hat{q}_{n,\alpha}^{+}\{R_i^{\text{LOO}}\}), \max_{i=1,..,n} \hat{\mu}_{-i}(X_{n+1} + \hat{q}_{n,\alpha}^{+}\{R_i^{\text{LOO}}\})].$$

Although this predictor is proven to have automatic validity, it behaves overly conservatively in practice. Similarly, CV+ is adapted from CCP with the conformal prediction interval defined as

$$\hat{C}^{\text{CV+}}_{n,\alpha,K} = [\hat{q}^{-}_{n,\alpha}\{\hat{\mu}_{\zeta_{-k}}(X_{n+1}) - R^{\text{CV}}_i\}, \hat{q}^{+}_{n,\alpha}\{\hat{\mu}_{\zeta_{-k}}(X_{n+1}) + R^{\text{CV}}_i\}],$$

where $\hat{\mu}_{\zeta_{-k}}$ denotes the underlying regression function estimator fitted onto the data $\zeta \setminus \zeta_k$ and $R^{CV+}_i = |Y_i - \hat{\mu}_{\zeta_{-k_{(i)}}}(X_i)|$ with $k_{(i)}$ represents which fold among the total K folds include example i. Again, Jackknife+ can be viewed as a special case of CV+. [5] proved a lower bound for CV+ coverage as

$$P\{Y_{n+1} \in \hat{C}^{\text{CV+}}_{n,\alpha,K} \geq 1 - 2\alpha - \frac{1 - K/n}{K+1}\},$$

which provides a meaningful bound for the case when K is large. Together with the result from [45], it was shown that the coverage is essentially $1 - 2\alpha$ for any K and the excess noncoverage is at most $\sqrt{2/n}$ uniformly over any choice of K.

Bootstrap Conformal Predictor (BCP)

A modification of CCP based on a bootstrap sampling strategy is called the bootstrap conformal predictor (BCP) proposed by Vovk in 2015 [42]. In other words, this method generates the proper training sets by taking K bootstrapped samples $Z_1, ..., Z_k$ and the complements of these set $Z^c_1, ..., Z^c_k$ where $Z^c_i = Z \setminus Z_i$ for $i \in \{1, 2, ..., k\}$become the corresponding calibration sets with Z denoting the whole training set. The nonconformity measure within each sample is defined the same way as before

$$\alpha_{i,k} := A(Z_k, z_i); \alpha^y_k := A(Z_k, (x_{n+1}, y)).$$

The aggregated p-value is defined as

$$p^y := \frac{\sum_{k=1}^{K} |\{z_i \notin Z_k : \alpha_{i,k} \leq \alpha_k^y\}| + T/|Z|}{T + T/|Z|},$$

where $T := \sum_{k=1}^{K}(n - |Z_k|)$ is the total size of the calibration sets. The randomized BCP can be obtained in a similar way as before. Empirical result [42] shows that BCP is well-calibrated as CCP, however, it does not outperform ICP in informational efficiency while K is relatively small. Informational efficiency can be improved by increasing K while this action also increases computational cost.

Generalized Aggregated Conformal Predictor

The several examples of aggregated conformal predictors introduced above can be generalized as follows. Firstly, we can use any sampling strategy to create k pairs of proper training and calibration sets. Then, for each sample, we obtain a p-value $p_{k,m}^y$in the same way as a full conformal predictor. Finally, the aggregated p-value with respect to the provisional label y is the mean of these p-values. The validity of aggregated conformal predictors is not automatic, and it has been shown that they can not be proved to be exactly valid.

3.3.3 Out-of-bag Conformal Predictor

Out-of-bag conformal predictor (OOBCP) was initially studied by [22] which adopts random forest as the base prediction algorithm under regression setting and can be a potential competitor for ICPs. Unlike ICPs, this new method does not need to sacrifice any available data for training the underlying model through calibrating on out-of-bag instances. Multiple early empirical studies show the advantages of using out-of-bag examples as calibration sets over existing techniques. [27] generalizes out-of-bag (OOB) calibration strategy names this type of predictor out-of-bag conformal predictors and further compares its properties versus different types of ACPs that we have introduced before.

Similar to BCP, the first step of OOBCP is taking K bootstrap samples ζ_k^t of size n from the original training sets and it requires each bootstrap sample to include approximately two-thirds of the unique patterns in the total training set ζ. Denote the classifier induced from each proper training set ζ_k^t as h_k and the OOB property indicator as

$$O_k^i = 1 \text{ if } z_i \in \zeta_k^t \text{ and } 0 \text{ , otherwise.}$$

Nonconformity score of each $z_i \in \zeta$ is defined as the average score of all bags that z_i is an out-of-bag example

$$\alpha_i := \frac{1}{\sum_{k=1}^{K} O_k^i} \sum_{k=1}^{K} O_k^i A(\zeta_k^t, (x_i, y_i))$$

For a test example (x_{n+1}, y), the nonconformity score is obtained as follows

$$\alpha^y := \frac{1}{\sum_{k=1}^{K} O_k^r} \sum_{k=1}^{K} O_k^r A(\zeta_k^t, (x_{n+1}, y))$$

, where r is randomly chosen from $\{1, ..., K\}$. Finally, the p-value for this potential label y is defined similarly to before

$$p_{n+1}^y := \frac{|\{z_i \in \zeta \setminus \{z_r\} : \alpha_i \geq \alpha^y\}| + 1}{|\zeta \setminus \{z_r\}| + 1}$$

or in a smoothed manner defined in the same way as before. The procedure can be interpreted as deleting a random example from the training set and letting the test example take its place. So far, there is a gap in theoretical evaluations of OOBCPs on their properties. Although the coverage guarantee is not automatic for OOBCPs, they perform well in practice. Their p-value stability is similar to that of an ICP, and they have shown better efficiency than BCP.

3.4 Conditional Valid Conformal Predictor

Although most conformal predictors automatically guarantee the marginal coverage rate for future prediction, satisfying the average coverage rate alone can be insufficient and even misleading in certain cases. For example, suppose we are dealing with a binary classification problem, we may get a super high coverage rate for one class and an extremely low coverage rate for the other class while still getting the marginal miscoverage rate below the specified significance level. In this case, the marginal coverage rate will be meaningless for interpretation. Thus, in reality, guarantee of within-category coverage and conditional validity is important for prediction algorithms to consider. In this section, we will introduce several variants of conformal predictors that are constructed to possess this nice property.

3.4.1 Training-conditional Validity

In this section, we discuss the training-conditional validity of different conformal predictors. This property ensures that most draws of training data produce reliable predictions for future test examples. Following [41], current literature studies the coverage probability conditioning on the training set by using a PAC (Probably Approximately Correct) form definition. Let $D_n = (Z_1, ..., Z_n)$ denote the training data set, the miscoverage rate of a conformal predictor conditioning on its training data can be expressed as follows

$$\alpha_P(D_n) = P\{Y_{n+1} \notin \hat{C}_n(X_{n+1})|D_n\},$$

where the probability is only with respect to the test point (X_{n+1}, Y_{n+1}) drawn from the distribution of examples P. A conformal predictor is (approximate) training-conditional valid if it satisfies the following equation

$$P\{\alpha_P(D_n) > \alpha + o(1)\} \leq o(1).$$

In other words, the probability that a training set results in a significantly higher training-conditional miscoverage rate than the nominal rate is vanishingly small. ICP [41] is one training-conditional valid example satisfying the following inequality

$$P\{\alpha_P(D_n) \leq \alpha + \sqrt{\frac{log(1/\delta)}{2n_1}}\} \geq 1 - \delta,$$

where n_1 is the size of the holdout or calibration set, and $\delta \in (0, 0.5]$ is arbitrary. Also, [8] shows that the CV+ method also achieves this property with the target nominal rate set to 2α which is also the proven marginal miscoverage rate

$$P\{\alpha_P(D_n) \leq 2\alpha + \sqrt{\frac{2log(K/\delta)}{m}}\} \geq 1 - \delta.$$

Whereas FCP and Jackknife+ method do not hold such properties without addressing further assumptions on the distribution of training data or on the nonconformity score function.

3.4.2 Label-conditional Validity

Another important property that is often desirable in practice is the guarantee of coverage while conditioning on important groups of examples. For example, in classification, we may want to achieve more control of coverage rate within each category. Thus, there is a strong motivation for developing conformal predictors with label-conditional validity which allows learners to control the set-prediction analogs of false positive and false negative rates.

Mondrian Conformal Predictor

Mondrian conformal predictor is a typical example of a label-conditional valid conformal predictor and it was first proposed by Vovk in 2003 [44]. The method obtains its name from the inspiration of Mondrian paintings. This method is commonly used for achieving

conditional validity within an important category of examples. For regression problem, [9] proposes mondrian conformal regressors which quantify uncertainty at the instance-level and overcome the problems with existing normalized conformal methods.

Other examples under this category include conditional ICPs [41] which applies to both regression and classification settings. Adaptive prediction sets (APS) and regularized adaptive prediction sets (RAPS) are developed to achieve class-conditional validity specific to classification problems which will be introduced in the next chapter.

3.4.3 Object-conditional Validity

In most cases, while talking about conditional coverage, we mainly concerned with the probability of covering true label conditioning on the object. Unfortunately, precise object-conditional validity was shown impossible to achieve for distribution-free predictors in a nontrivial way [41] [26]. This result was originally discovered and proved in section A.2. of [26]. Here, we presented the proof in the case of classification in proposition 4 of [41] and the proof of the regression case follows similar idea.

Theorem 4. *[41]Suppose X is a separable metric space equipped with the Borel σ-algebra. Let $\epsilon \in (0, 1)$. Suppose that a set predictor C has $1 - \epsilon$ object conditional validity. In the case of regression, we have, for all P and P_X-almost all P_X-non-atoms x,*

$$P^n(\hat{C}_n(x) = \infty) \geq 1 - \epsilon.$$

In the case of classification, we have, for all P, all $y \in Y$, and P_X-almost all P_X-non-atoms x,

$$P^n(y \in \hat{C}_n(x)) \geq 1 - \epsilon.$$

Proof. For classification case, suppose there exists a measurable subset E of P_X-non-atoms $x \in X$ such that $P_X(E) > 0$ and $P^n(y \in \hat{C}_n(x)) < 1 - \epsilon$. Shrink E such that $P_X(E) > 0$ still holds and there exists $\delta > 0$ such that, for all $x \in E$, we have $P^n(y \in \hat{C}_n(x)) \leq 1 - \epsilon - \delta$.

Denote the total variation distance between P and Q by $V(P,Q) := sup_A|P(A) - Q(A)|$. Then, follows the idea in section 2.4 of [37], we have

$$V(P^n, Q^n) \leq \sqrt{2}\sqrt{1-(1-V(P,Q))^n}.$$

Then, shrink E again such that $P_X(E) > 0$ and

$$\sqrt{2}\sqrt{1-(1-P_X(E))^n} \leq \delta/2.$$

Define probability distribution Q on Z in the way that $Q(A \times B) = P(A \times B)$ for all measurable $A \subseteq (X \setminus E)$ and all $B \subseteq Y$ and that $Q(A \times \{y\}) = P_X(A)$ for all measurable $A \subseteq E$. But then for each $x \in E$ we have

$$Q^n(y \in \hat{C}(Z_1, ..., Z_n, x)) \leq 1 - \epsilon - \delta/2$$

$$\Rightarrow Q^n(y \in \hat{C}(Z_1, ..., Z_n, X_{n+1}) \text{ and } X_{n+1} \in E) \leq (1 - \epsilon - \delta/2)Q_X(E),$$

which contradicts the assumption that predictor C is object conditionally valid.

□

Acknowledging the impossibility of exact object conditional validity, researchers have been devoted to exploring the gap between unconditional and label-conditional coverage. One branch of studies such as [20] and [11] focuses on studying the asymptotic label-conditional coverage. Another line of research was devoted to developing approximations or relaxations of exact label-conditional validity such as [15].

3.5 Relaxing Assumptions

From previous sections, although CP is a powerful method with no restriction on the distribution of training examples, we must guarantee that the test and training data are

exchangeable since proofs of their properties were built on this assumption. However, exchangeable data is not always achievable in practice. In recent years, researchers have shown increasing interest in developing extensions of conformal prediction framework [6] to conditions breaking the exchangeable assumption such as working in a time series scenario. In this section, several important variants will be introduced.

3.5.1 Covariate Shift

One of the first successful attempts in this research direction is the work by Tibashirani and his collaborators in 2019 [36]. They investigate the feasibility of conformal inference in the case when the distribution of test data differs from that of the training data with the conditional distribution of $Y|X$ remaining the same across both training and test data. Formally, the distributions of training and test data can be expressed as follows

$$(X_i, Y_i) \sim P = P_X \times P_{Y|X}, i = 1, ..., n,$$

$$(X_{n+1}, Y_{n+1}) \sim \tilde{P} = \tilde{P}_X \times P_{Y|X}, \text{ independently.}$$

Under such a setting, the change in distribution comes from covariate. This assumption was referred to as *covariate shift* in literature. [36] suggested to use a weighted methodology. By introducing weights defined from the likelihood ratio $w(x) = \frac{d\widetilde{P}(x)}{dP(x)}$ for $x \in X$ as follows

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)}, i = 1, ..., n \text{ and } p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}.$$

In this way, more weights were put on calibration points which are more likely under the distribution of test points. The weighted conformal prediction gives the result

$$\hat{C}_{n,\alpha}(x) = \{y \in \mathbb{R} : \alpha_{n+1}^{(x,y)} \leq Quantile(1-\alpha; \sum_{i=1}^n p_i^w(x)\delta_{\alpha_i} + p_{n+1}^w(x)\delta_\infty),$$

where α_i are residual-based nonconformity scores and $Quantile(1-\alpha; \frac{1}{n}\sum_{i=1}^{n}\delta_{\alpha_i})$ is the level $1-\alpha$ quantile of the empirical distribution of values α_i's with δ_{α_i} denoting the point mass at α_i. Also, this method covers the original conformal prediction as a special case because when all data are exchangeable, the weights will all equal $\frac{1}{n+1}$ which gives an original conformal prediction.

More generally, the weighted method is applicable in the setting that data is under-weighted exchangeable distribution [36, Theorem 2]. However, in practice, most applications of the weighted conformal method require extensive computation costs. Only in special cases of covariate shift such that the likelihood ratio is known or can be estimated from unlabeled data, does the weighted method become feasible in reality [36].

3.5.2 Distributional Shift

Distributional shift depicts another type of change in the distribution of training examples and test examples. This setting is more complex than the covariate shift case because it allows for changes in distribution for both label and object such as in the case of time series where data distribution can change gradually over time. Suppose the calibration data and test data are drawn independently from different distributions denoted by P_i and P_{test} respectively.

Weighted Approach

Following the work introduced in the previous section, [6] further improved the weighted method to be applicable to the distributional shift setting in the following way. For covariate shift settings, quantiles are likelihood-weighted in order to address the change in covariate distribution. In this case, they introduced a custom-weighted conformal prediction method that uses arbitrary fixed weights. Their method upweights points in the training set that are more representative of test distribution and downweights the rest to get closer to the desired coverage guarantee. For example, if each data Z_i occurs at time i, we may choose $w_1 \leq, ..., \leq w_n$ to emphasize the importance of more recent data

points. If all weights equal one, then this method becomes identical to the normal version of conformal prediction.

Let $w_i \in [0,1], i = 1, ..., n$ be fixed and arbitrary weights, and define

$$\tilde{w}_i = \frac{w_i}{w_1 + w_2 + ... + w_n + 1}, i = 1, ..., n, \text{ and } \tilde{w}_{n+1} = \frac{1}{w_1 + w_2 + ... + w_n + 1}.$$

The prediction result following a full conformal prediction design would be

$$\hat{C}_{n,\alpha}(X_{n+1}) = \{y \in Y : \alpha_{n+1}^{(X_{n+1},y)} \leq Quantile(1-\alpha; \sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{\alpha_i})\}.$$

This weighted method also can address the issue of adopting nonsymmetric underlying prediction algorithms [6]. While this method is more robust to violation of the regular assumptions of conformal prediction by giving users the freedom to customize weights, the design of a proper weight strategy can be a hard problem in practice.

Adaptive Approach

Another promising approach that makes conformal prediction framework more robust was proposed in 2021 [16]. The core of this method, adaptive confidence inference (ACI), is not only the idea of conformal prediction. It follows the construction of conformal quantile regression proposed in [32] which combines conformal prediction with quantile regression. Also, we should be aware that ACI was proposed under an online setting. While it makes predictions sequentially, ACI adjusted the nominal miscoverage rate α_t at each step to maintain the realized miscoverage rate denoted by $M_t(\alpha)$ at the desired level α.

Formally, unlike CQR which uses only one score function and quantile function, ACI defines score functions $S_t(\cdot)$ and quantile function $\hat{Q}(\cdot)$ that are changing over time to adjust for shift in distribution. For each prediction $\hat{C}_t(\alpha) := \{y : S_t(X_t, Y_t) \leq \hat{Q}_t(1-\alpha)\}$,

define the realized miscoverage rate as

$$M_t(\alpha) := P(S_t(X_t, Y_t) > \hat{Q}_t(1-\alpha)),$$

where the probability is over the test point (X_t, Y_t) and the data used to fit $S_t(\cdot)$ and $\hat{Q}(\cdot)$. Then, we can define the adaptive miscoverage rate as

$$\alpha_t^* := sup\{\beta \in [0,1] : M_t(\beta) \leq \alpha\}.$$

In practice, such calibration is obtained by online updating that is intuitively straightforward as follows

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - err_t),$$

where $\gamma > 0$ is a fixed step size. If the previous prediction covers the true value, then we slightly increase the miscoverage rate and vice versa. Alternatively, we can also make adjustments according to a sequence of past predictions

$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \sum_{s=1}^{t} w_s err_s),$$

where $\{w_s\}_{1 \leq s \leq t} \subseteq [0,1]$ is a sequence of increasing weights with $\sum_{s=1}^{t} w_s = 1$. ACI succeeds in improving the robustness of conformal inference. In practice, the choice of step-size γ can be a problem in application. Recently, several alternative works have been proposed independently to avoid this obstacle such as in [17], [46] and [7].

3.6 Summary

Chapter 3 provides an overview of various conformal prediction algorithms. The chapter is structured to guide the reader from fundamental concepts of conformal prediction framework to improvements and enhancements on the basic CP framework. Extensions

of CP are designed to improve the conformal predictor's validity, efficiency, and robustness under different conditions. Following the introduction of the basic terminologies and procedures of conformal prediction based on Full Conformal Prediction (FCP), we discussed two important evaluation criteria: validity and efficiency for conformal predictors with a focus on the trade-offs between validity and efficiency.

The rest of this chapter introduces multiple variants of conformal predictors that address the limitations of the original framework. First, it discusses how different calibration strategies improve the computation efficiency of original conformal prediction framework. It also points out the limitation faced by some of these conformal predictors' on achieving a theoretical guarantee on coverage rate. Furthermore, conditional valid conformal predictors were brought to attention as they are important in application to guarantee useful coverage guarantee across different categories. Lastly, we present several inspiring improvements on the conformal framework proposed in recent years that allow the method to work for nonexchangeable data. We presented how the researchers achieved this goal by adopting either weighted nonconformity scores or an adaptive miscoverage level.

Chapter 4

Logits-based Nonconformity Score Applicable to Image Classification Problems

This chapter focuses specifically on conformal classifiers designed to quantify the uncertainty of neural network algorithms that address image classification problems. Within this specific scope, we mainly introduce two of the most popular methods Adaptive Prediction Set (APS) [33] and Regularized Adaptive Prediction Set (RAPS) [3] under this category. Then modifies them to use logits-based nonconformity scores and refer to the two new methods as APS-Logits and RAPS-Logits. Finally, we conducted empirical comparison of the four algorithms' performance on image classification applications.

4.1 Motivation

Neural networks have revolutionized the field of image classification, a critical task in computer vision with applications ranging from medical diagnostics to autonomous driving. Neural networks, particularly deep learning models such as Convolutional Neural Networks (CNNs), have performed exceptionally in identifying and categorizing images

into predefined classes. This success is largely attributed to their ability to automatically learn and extract hierarchical features from raw image data, significantly outperforming traditional machine learning methods. Despite their success, neural networks often provide point estimates without quantifying the uncertainty of their predictions. This limitation can be critical in high-stakes applications where understanding the confidence of a model's prediction is as important as the prediction itself. To address this challenge, conformal prediction has emerged as a promising technique that complements neural network models by providing a measure of confidence alongside each prediction.

While the design of nonconformity score of conformal classifiers mainly relies on the 'probability' output of deep learning algorithms, there exist empirical results [18] [39] [47] suggest that using logits instead of 'probabilities' such as the softmax output is better in practice. Thus, it become an interesting topic to explore how substituting softmax output with logits in the application of conformal classifiers influences the prediction results. In the following subsections, we examines the performances of changing existing probability-based conformal algorithms to logits-based methods by first introducing the design of APS and RAPS algorithms and then presented an application of image classification.

4.1.1 Adaptive Prediction Sets

The Adaptive Prediction Set algorithm is designed to construct prediction sets with valid and adaptive coverage for multi-class classification problems [33]. Given a dataset $\{(X_i, Y_i)\}_{i=1}^n$ with features $X_i \in \mathbb{R}^p$ and discrete labels $Y_i \in \{1, 2, \ldots, C\}$, the APS algorithm seeks to form a prediction set $\hat{C}_{n,\alpha}(X_{n+1}) \subseteq \{1, 2, \ldots, C\}$ for a new data point (X_{n+1}, Y_{n+1}) such that the coverage probability $P[Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})] \geq 1 - \alpha$ holds marginally regardless of the accuracy of the underlying black-box algorithm.

Let $\pi_y(x) = P[Y = y|X = x]$ denote the conditional probability for each $y \in Y$, and denote the order statistics for $\pi_y(x)$ as $\pi_{(1)}(x) \geq \pi_{(1)}(x) \geq ... \geq \pi_{(C)}(x)$. Define $L(x; \pi, \tau)$ as

the generalized conditional quantile function:

$$L(x;\pi,\tau) = min\{c \in \{1,...,C\} : \pi_{(1)}(x) + \pi_{(2)}(x) + ... + \pi_{(c)}(x) \geq \tau\}.$$

The APS algorithm defines a generalized inverse quantile function to calculate a conformity score as follows

$$E(x,y,u;\hat{\pi}) = \min\{\tau \in [0,1] : y \in S(x,u;\hat{\pi},\tau)\},$$

where

$$S(x,u;\pi,\tau) = \begin{cases} \text{'y' indices of the } L(x;\pi,\tau)-1 \text{ largest } \pi_y(x), & \text{if } u \leq V(x;\pi,\tau), \\ \text{'y' indices of the } L(x;\pi,\tau) \text{ largest } \pi_y(x), & \text{otherwise} \end{cases}$$

is a prediction set formed based on estimated class probabilities $\hat{\pi}_y(x)$ obtained from the underlying black-box algorithm with

$$V(x;\pi,\tau) = \frac{1}{\pi_{L(x;\pi,\tau)}(x)} \left[\sum_{c=1}^{L(x;\pi,\tau)} \pi_c(x) - \tau \right].$$

The algorithm can be carried out with not only the split conformal calibration scheme but also the Jackknife+ or CV+ calibration which have been introduced in the previous chapter. Using the split-conformal calibration as an example, the APS algorithm trains a black-box classifier on a subset of data I_1 to generate estimated probabilities $\hat{\pi}$, and then using hold-out data I_2 to calibrate the conformity scores. This results in prediction sets that adapt to the underlying data distribution, providing strong empirical performance in terms of both marginal and approximate conditional coverage. Formally, the prediction set for a new observation X_{n+1} is given by

$$\hat{C}_{n,\alpha}(X_{n+1}) = \left\{ y \in \{1,2,\ldots,C\} : E(X_{n+1},y,U_{n+1};\hat{\pi}) \leq \hat{Q}_{1-\alpha} \right\},$$

where $\hat{Q}_{1-\alpha}$ is the estimated $1-\alpha$ quantile of the nonconformity scores computed on the hold-out set I_1.

4.1.2 Regularized Adaptive Prediction Sets

The Regularized Adaptive Prediction Sets algorithm builds on the APS method and can provide reliable prediction sets with better prediction efficiency [3]. While APS aims to construct prediction sets that achieve a specified marginal coverage level by using hold-out samples to determine a conformity score threshold, it often results in large prediction sets due to noisy probability estimates, especially for classes with low predicted probabilities. In contrast, RAPS incorporates a regularization term that penalizes the inclusion of unlikely classes based on their rank, leading to smaller and more stable prediction sets without sacrificing coverage. This regularization addresses inefficiency of APS method, where the order of classes with low probabilities can significantly impact the size of the prediction set.

Formally, the RAPS algorithm defines a nonconformity score for each class $y \in \{1, 2, \ldots, C\}$ using the formula

$$E(x, y, u; \hat{\pi}) = \sum_{y'=1}^{C} \hat{\pi}_{y'}(x) \mathbb{I}_{\{\hat{\pi}_{y'}(x) > \hat{\pi}_y(x)\}} + \hat{\pi}_y(x) \cdot u + \lambda \cdot (\text{rank}(y) - k_{\text{reg}})^+,$$

where $\hat{\pi}_y(x)$ is the estimated probability of class y, $u \sim \text{Uniform}(0, 1)$ is a random variable to break ties, λ is a regularization parameter, and k_{reg} is a parameter determining which labels should be penalized, and rank(y) represents the ranking of probability of label y among all the labels based on the estimated probabilities.

To obtain the threshold τ, we start with a set-valued function $C(X, u, \tau)$, which maps a feature vector X and a uniform random variable $u \in [0, 1]$ to a subset of the possible labels. The function $C(X, u, \tau)$ is indexed by the parameter τ, where larger τ corresponds to a larger prediction set size. The goal is to find the smallest value of τ that ensures the prediction set covers the true label Y with a probability of at least $1 - \alpha$. Thus, we

calibrate τ on a calibration set I_1. τ is chosen such that the proportion of calibration examples (X_i, Y_i) where $Y_i \in C(X_i, U_i, \tau)$ is at least $1 - \alpha$. We estimate τ by the following formula

$$\hat{\tau} = \inf \left\{ \tau : \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{E(X_i, Y_i, u; \hat{\pi}) \leq \tau)\}} \geq \frac{\lceil (1-\alpha)(1+n) \rceil}{n} \right\}$$

Choosing $\tau = \hat{\tau}$ ensures that the prediction sets produced on new data will have the desired coverage probability $1 - \alpha$.

The prediction set $\hat{C}(X_{n+1})$ for a new observation X_{n+1} is constructed by including classes in increasing order of their scores until the threshold τ, determined from a hold-out set, is reached:

$$\hat{C}(X_{n+1}) = \{ y \in \{1, 2, \ldots, C\} : E(X_{n+1}, y, U_{n+1}; \hat{\pi}) \leq \hat{\tau} \}.$$

The algorithm ensures that the coverage probability $P(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$ holds marginally, and the regularization helps reduce the impact of noisy probability estimates, resulting in more efficient and reliable prediction sets.

With a similar aim of quantifying uncertainty, another emerging stream of research named out of distribution detection which is popular in recent years shares many similarities with the topic of this thesis: conformal prediction. Inspired by the trend of designing out-of-distribution scores with raw logits, we propose two potential designs of logit-based nonconformity scores in the following section.

4.2 Method

In both APS and RAPS, the estimated probabilities of neural network classifiers are typically obtained using the softmax function applied to the raw logits. However, our method introduces a novel approach by directly using the raw logits to construct the conformity scores aiming to avoid potential distortions introduced by the softmax transformation.

This approach leverages the raw logits to construct prediction sets that maintain the same desired coverage properties while potentially providing better performance in terms of prediction set size and stability.

Formally, let $z_y(x)$ denote the raw logit output of a neural network for class $y \in \{1, 2, \ldots, C\}$ given input $x \in \mathbb{R}^p$. Instead of using the softmax-transformed probabilities $\hat{\pi}_y(x)$, we directly utilize the raw logits to define our conformity score. The nonconformity score for class y is given by

$$E(x, y, u; z) = \sum_{y'=1}^{C} z_{y'}(x)\mathbb{I}_{\{z_{y'}(x) > z_y(x)\}} + z_y(x) \cdot u + \lambda \cdot (\text{rank}(y) - k_{\text{reg}})^+,$$

where $u \sim \text{Uniform}(0, 1)$ is a random variable to break ties, λ is a regularization parameter, and k_{reg} serves the same purpose as before, and rank(y) represents the ranking of $z_y(x)$.

The prediction set $\hat{C}(X_{n+1})$ for a new observation X_{n+1} is then constructed by first estimating the threshold τ based on the calibration data set in the same way as in RAPS method and then including all classes giving a nonconformity score not greater than the estimated threshold $\hat{\tau}$:

$$\hat{C}(X_{n+1}) = \{y \in \{1, 2, \ldots, C\} : E(X_{n+1}, y, U_{n+1}; z) \leq \hat{\tau}\}.$$

By avoiding the softmax transformation, our method directly utilizes the model's raw outputs, potentially capturing more precise information about the classifier's confidence and relationships among classes. In the following section, we examine the empirical performance of our methods against APS and RAPS on handling image classification tasks.

4.3 Experiment

Our experiment was carried out following the design of experiment 2 of [3]. The performance of the RAPS, APS and our methods was evaluated on the ImageNet validation

Model	Top-1	Top-5	APS	APS-Logits	RAPS	RAPS-Logits
ResNet152	0.782	0.940	0.935	0.965	0.906	0.899
ResNet101	0.773	0.935	0.935	0.965	0.900	0.898
ResNet50	0.716	0.928	0.930	0.961	0.904	0.896
ResNet18	0.697	0.891	0.922	0.952	0.901	0.899
DenseNet161	0.770	0.935	0.931	0.964	0.908	0.896
ResNeXt101	0.792	0.945	0.937	0.969	0.906	0.899
VGG16	0.716	0.904	0.927	0.957	0.904	0.897
Inception-v3	0.695	0.886	0.924	0.968	0.901	0.899
ShuffleNet-v2	0.693	0.884	0.927	0.957	0.900	0.902

Table 4.1: Comparison of four algorithms' empirical coverage rate on ImageNet data set with 9 different pre-trained base classifiers.

data set. The objective was to compare the coverage and efficiency of the prediction sets generated by four conformal prediction methods. We utilized nine standard, pre-trained ImageNet classifiers from the torchvision repository, including models such as ResNet152, ResNet50, and DenseNet161, among others. The ImageNet data set contains 50000 images in total. For each trial, we randomly sampled 10000 images where 3000 of them were used to select parameter λ and the reset 7000 of them were used to calibrate the threshold τ. The rest 40000 images in the dataset were used as testing examples. This sampling procedure was repeated 100 times to generate the results shown in the table below.

For the two methods building on estimated probabilities, all classifiers were calibrated using temperature scaling, a form of Platt scaling, to adjust the predicted probabilities and improve the accuracy of the prediction sets. Following calibration, we applied the APS and RAPS methods to the evaluation set, recording both the coverage and the size of the prediction sets. The median-of-means over the 100 trials was used to report the results for both coverage and prediction set size. The objective of the experiment was to determine the ability of each method to achieve the desired coverage while minimizing the size of the prediction sets.

Model	Top-1	Top-5	APS	APS-Logits	RAPS	RAPS-Logits
ResNet152	0.782	0.940	9.895	9.681	2.103	2.539
ResNet101	0.773	0.935	10.376	10.326	2.250	2.714
ResNet50	0.716	0.928	11.067	10.516	2.496	2.933
ResNet18	0.697	0.891	15.101	15.652	4.261	5.309
DenseNet161	0.770	0.935	10.801	10.498	2.362	2.697
ResNeXt101	0.792	0.945	19.218	10.830	1.983	2.415
VGG16	0.716	0.904	13.289	13.144	3.493	4.088
Inception-v3	0.695	0.886	85.900	44.540	5.115	6.430
ShuffleNet-v2	0.693	0.884	30.054	93.158	4.837	6.184

Table 4.2: Comparison of the four algorithms' prediction sets sizes on ImageNet data set with 9 different pre-trained base classifiers.

4.4 Discussion

From the results in Tables 1 and 2, we know that all methods in this experiment successfully achieved the desired coverage rate across various pre-trained models on the ImageNet validation dataset. This consistency in coverage indicates that each method is capable of generating prediction sets that reliably include the true label, which is one of the most important properties of conformal prediction algorithms. Among the four methods, RAPS consistently produced the smallest prediction sets, as shown in Table 2. For example, with the ResNet152 model, RAPS generated a median set size of 2.103, which is significantly smaller than the median set size produced by APS (9.895). For the efficiency of the two methods proposed in this section, our logits-based variant of APS (i.e. APS-Logits) generates smaller prediction sets compared to the standard APS method. For instance, with the DenseNet161 model, APS-Logits produced a median set size of 10.498, whereas APS produced a slightly larger median set size of 10.801. This suggests that incorporating logits into the design of nonconformity scores could enhance the efficiency of the prediction sets for conformal algorithms. However, when examining the RAPS versus RAPS-Logits, the results indicate that the introduction of logits did not lead to a similar efficiency gain. In some cases, RAPS-Logits even produced slightly larger prediction sets compared to RAPS. For example, with the ResNet50 model, the median set size

for RAPS-Logits was 2.933, whereas for RAPS, it was slightly smaller at 2.496. The reason that RAPS-Logits seem to be less competitive versus RAPS potentially comes from the difficulty of choosing regularization parameter λ. Since raw logits are numerically unstable in general, it is harder to find suitable regularization parameters when our logits-based method tries to adopt the same regularization technique as a probability-based method.

Overall, the experiment demonstrates that under certain circumstances, logit-based methods can improve the efficiency of prediction sets while maintaining the desired coverage. However, further theoretical explorations on logits-based methods' properties are needed such as evaluation of the conditional coverage guarantee of APS-Logits and RAPS-Logits. Moreover, it is also an interesting topic to explore how to design suitable regularization terms for nonconformity scores based directly on raw logits, which may offer a more competitive conformal prediction algorithm compared to RAPS. It is important to note that our methods represent relatively naive attempts to incorporate logits into the design of nonconformity scores. We modified existing methods to create logits-based variants rather than developing entirely new nonconformity scores grounded in the original properties of logits. These initial attempts show the potential of logits-based nonconformity scores. There exists considerable room for further innovation. More creative and powerful methods that were originally designed around the use of logits could potentially offer even greater improvements in efficiency while maintaining the desired coverage.

Chapter 5

Conclusion

This thesis has explored the conformal prediction framework, emphasizing its ability to provide robust uncertainty quantification in predictive modeling. Through a combination of theoretical analysis and practical experimentation, we have highlighted both the strengths and limitations of this approach.

In Chapter three, we introduced the foundational concepts of conformal prediction, exploring how to design nonconformity scores for different conformal predictors. Multiple calibration strategies applicable to conformal prediction algorithms were presented with comparison of their validity and efficiency. We also presented the challenge of achieving object-conditional validity in a distribution-free setting for most conformal predictors. Lastly, this chapter discussed how to extend the conformal framework beyond the exchangeable setting. We introduced various modifications to conformal framework that were designed to relax exchangeable assumptions, enabling the application of conformal prediction to more complex data scenarios, such as covariate shift and distributional shift.

In Chapter four, we attempted to evaluate the possibility of using logits to design nonconformity scores. We adapted two promising conformal approaches: APS and RAPS, which are designed to quantify the uncertainty for classification problems. We examined on the performance the logits-based version of these two methods: APS-Logits and

RAPS-Logits by conducting experiments on ImageNet dataset. By comparing the prediction sets' coverage and sizes, we found that APS-Logits had better efficiency than APS while maintaining the desired coverage rate. However, RAPS-Logits did not outperform RAPS. While the results suggest using raw logits versus estimated probability can be promising in improvement of conformal algorithm's efficiency in some cases, it remains a challenge to design logits-based nonconformity with suitable regularization term.

In conclusion, conformal prediction represents a valuable tool in statistical learning. It offers a flexible and reliable means of uncertainty quantification. While challenges remain, such as in adapting the framework to non-exchangeable data, ongoing research continues to push the boundaries of what is possible. Future work may focus on exploring its integration with other uncertainty quantification methods such as out-of-distribution detection.

Bibliography

[1] ALEKSANDROVA, M., AND CHERTOV, O. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications* (08–10 Sep 2021), L. Carlsson, Z. Luo, G. Cherubin, and K. An Nguyen, Eds., vol. 152 of *Proceedings of Machine Learning Research*, PMLR, pp. 151–170.

[2] ANGELOPOULOS, A. N., AND BATES, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

[3] ANGELOPOULOS, A. N., BATES, S., MALIK, J., AND JORDAN, M. I. Uncertainty sets for image classifiers using conformal prediction. *ArXiv abs/2009.14193* (2020).

[4] BALASUBRAMANIAN, V., HO, S.-S., AND VOVK, V. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*, 1st ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2014.

[5] BARBER, R. F., CANDÈS, E. J., RAMDAS, A., AND TIBSHIRANI, R. J. Predictive inference with the jackknife+. *The Annals of Statistics* (2019).

[6] BARBER, R. F., CANDES, E. J., RAMDAS, A., AND TIBSHIRANI, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics 51*, 2 (2023), 816–845.

[7] BASTANI, O., GUPTA, V., JUNG, C., NOAROV, G., RAMALINGAM, R., AND ROTH, A. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems 35* (2022), 29362–29373.

[8] BIAN, M., AND BARBER, R. F. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics* (2022).

[9] BOSTRÖM, H., AND JOHANSSON, U. Mondrian conformal regressors. In *Conformal and Probabilistic Prediction and Applications* (2020), PMLR, pp. 114–133.

[10] CHEN, W., CHUN, K.-J., AND BARBER, R. F. Discretized conformal prediction for efficient distribution-free inference. *Stat 7*, 1 (2018), e173. e173 sta4.173.

[11] CHERNOZHUKOV, V., WÜTHRICH, K., AND YINCHU, Z. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory* (2018), PMLR, pp. 732–749.

[12] CHERNOZHUKOV, V., WÜTHRICH, K., AND ZHU, Y. Distributional conformal prediction. *Proceedings of the National Academy of Sciences 118*, 48 (2021), e2107794118.

[13] CORTES, C., AND VAPNIK, V. N. Support-vector networks. *Machine Learning 20* (1995), 273–297.

[14] FONTANA, M., ZENI, G., AND VANTINI, S. Conformal prediction: a unified review of theory and new challenges. *Bernoulli 29*, 1 (2023), 1–23.

[15] FOYGEL BARBER, R., CANDES, E. J., RAMDAS, A., AND TIBSHIRANI, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA 10*, 2 (2021), 455–482.

[16] GIBBS, I., AND CANDES, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems 34* (2021), 1660–1672.

[17] GIBBS, I., AND CANDÈS, E. J. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research 25*, 162 (2024), 1–36.

[18] HENDRYCKS, D., BASART, S., MAZEIKA, M., ZOU, A., KWON, J., MOSTAJABI, M., STEINHARDT, J., AND SONG, D. Scaling out-of-distribution detection for real-world settings, 2022.

[19] HEWITT, E., AND SAVAGE, L. J. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society 80*, 2 (1955), 470–501.

[20] JING LEI, MAX G'SELL, A. R. R. J. T., AND WASSERMAN, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association 113*, 523 (2018), 1094–1111.

[21] JING LEI, J. R., AND WASSERMAN, L. Distribution-free prediction sets. *Journal of the American Statistical Association 108*, 501 (2013), 278–287. PMID: 25237208.

[22] JOHANSSON, U., BOSTRÖM, H., LÖFSTRÖM, T., AND LINUSSON, H. Regression conformal prediction with random forests. *Machine learning 97* (2014), 155–176.

[23] JOHANSSON, U., LINUSSON, H., LÖFSTRÖM, T., AND BOSTRÖM, H. Model-agnostic nonconformity functions for conformal classification. *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), 2072–2079.

[24] LEI, J. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika 106*, 4 (2019), 749–764.

[25] LEI, J., AND WASSERMAN, L. Distribution free prediction bands. *arXiv preprint arXiv:1203.5422* (2012).

[26] LEI, J., AND WASSERMAN, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology 76*, 1 (2014), 71–96.

[27] LINUSSON, H., JOHANSSON, U., AND BOSTRÖM, H. Efficient conformal predictor ensembles. *Neurocomputing 397* (2020), 266–278.

[28] MONTGOMERY, D. C., PECK, E. A., AND VINING, G. G. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[29] PAPADOPOULOS, H., AND HARALAMBOUS, H. 2011 special issue: Reliable prediction intervals with regression neural networks. *Neural Netw. 24*, 8 (oct 2011), 842–851.

[30] PAPADOPOULOS, H., PROEDROU, K., VOVK, V., AND GAMMERMAN, A. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13* (2002), Springer, pp. 345–356.

[31] PAPADOPOULOS, H., VOVK, V., AND GAMMERMAN, A. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research 40* (Apr. 2011), 815–840.

[32] ROMANO, Y., PATTERSON, E., AND CANDES, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems* (2019), H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc.

[33] ROMANO, Y., SESIA, M., AND CANDÈS, E. J. Classification with valid and adaptive coverage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2020), NIPS '20, Curran Associates Inc.

[34] SHAFER, G., AND VOVK, V. A tutorial on conformal prediction. *J. Mach. Learn. Res. 9* (June 2008), 371–421.

[35] STEINBERGER, L., AND LEEB, H. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801* (2016).

[36] TIBSHIRANI, R. J., FOYGEL BARBER, R., CANDES, E., AND RAMDAS, A. Conformal prediction under covariate shift. *Advances in neural information processing systems 32* (2019).

[37] TSYBAKOV, A. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

[38] VALIANT, L. G. A theory of the learnable. *Communications of the ACM 27*, 11 (1984), 1134–1142.

[39] VAZE, S., HAN, K., VEDALDI, A., AND ZISSERMAN, A. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations* (2022).

[40] VOVK, V. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* (2002), pp. 187–196.

[41] VOVK, V. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning* (Singapore Management University, Singapore, 04–06 Nov 2012), S. C. H. Hoi and W. Buntine, Eds., vol. 25 of *Proceedings of Machine Learning Research*, PMLR, pp. 475–490.

[42] VOVK, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence 74* (2015), 9–28.

[43] VOVK, V., GAMMERMAN, A., AND SHAFER, G. *Algorithmic learning in a random world*, vol. 29. Springer, 2005.

[44] VOVK, V., LINDSAY, D., NOURETDINOV, I., AND GAMMERMAN, A. Mondrian confidence machine. *Technical Report* (2003).

[45] VOVK, V., AND WANG, R. Combining p-values via averaging. *Biometrika 107*, 4 (2020), 791–808.

[46] ZAFFRAN, M., FÉRON, O., GOUDE, Y., JOSSE, J., AND DIEULEVEUT, A. Adaptive conformal predictions for time series. In *International Conference on Machine Learning* (2022), PMLR, pp. 25834–25866.

[47] ZHANG, Z., AND XIANG, X. Decoupling maxlogit for out-of-distribution detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 3388–3397.