UNIVERSITY OF CAMBRIDGE

Part III Mathematics 2024-25

New Advances in Conformal Prediction

Shreyas Pandit

May 2025

Contents

1 Introduction

Quantifying the uncertainty in the predictions produced by predictive models is important to enable them to be confidently deployed in real-world settings. Uncertainty in predictions is typically quantified through a prediction set, a set containing the true value of the response with high probability. For example, it is well-known that one can construct prediction sets containing the true response with any desired probability in the normal linear model. In Bayesian statistics, the posterior predictive distribution can be used to construct prediction sets. An important limitation of these approaches to uncertainty quantification is that they make strong distributional assumptions on the data-generating process.

Conformal prediction is a framework for uncertainty quantification developed by [VGS] that constructs prediction sets whose validity holds in finite-samples and does not depend on the exact distribution of the data. The primary appeal of conformal prediction is that it can be combined with any base predictive model to generate valid prediction sets for this base model. This presents conformal prediction as a promising approach to uncertainty quantification for modern machine learning methods, which continue to be highly successful at predictive tasks, and has led to conformal prediction becoming an increasingly prominent topic of research in both statistics and machine learning.

The task we will focus on in this essay is as follows. Suppose we are given predictor-response pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and a base predictive model. Suppose (X_{n+1}, Y_{n+1}) is a test point, where X_{n+1} is given and we wish to predict the unknown response Y_{n+1}. Conformal prediction uses the data points $(X_1, Y_1), \ldots, (X_n, Y_n)$, X_{n+1} and the base model to construct a prediction set $C(X_{n+1})$ such that

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha, \tag{1.1}$$

where $\alpha \in (0, 1)$ is a pre-specified miscoverage level. A result of the above form is referred to as a *coverage guarantee*.

We begin Section 2 by introducing the key mathematical components of conformal prediction. In Section 2.1, we will introduce the notion of exchangeability, which underpins the validity of conformal prediction. Section 2.2 will detail the *full conformal prediction* procedure and prove that it achieves the coverage guarantee (1.1). In Section 2.3, we will study the *split conformal prediction* procedure, which is a more computationally efficient version of conformal prediction, and prove that it also achieves the coverage guarantee (1.1). To conclude Section 2, Section 2.4 will illustrate specific instances of split conformal prediction through numerical experiments.

In Section 3, we explore the theory behind three different extensions to the standard conformal prediction framework. Whilst exchangeability is fundamental in achieving the coverage guarantees in Section 2, the method presented in Section 3.1 provides a modified coverage guarantee under violations of exchangeability. A specific kind of violation of exchangeability is distribution shift, and a variant of conformal prediction for this setting is presented in Section 3.2. Finally, Section 3.3 returns to the exchangeable setting and presents a method that extends the class of guarantees of conformal prediction.

1.1 Notation

Throughout the essay, we use the following notation.

For any positive integer n, we define $[n] \coloneqq \{1, \ldots, n\}$ and S_n to be the set of permutations on n elements. If $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and $\sigma \in S_n$, then $\sigma(v) \coloneqq (v_{\sigma(1)}, \ldots, v_{\sigma(n)})$. We write $\mathcal{X}$ and $\mathcal{Y}$ to denote the (measurable) spaces of predictors and responses, respectively.

1.2 Numerical Experiments and Code

We implemented the code for the numerical experiments and figures using Python. We used the package quantile-forest [Joh24], which implements the method in [MR06]. The code is provided in a Github repository.

2 Foundations of Conformal Prediction

In this section, we introduce the key mathematical tools for conformal prediction and explain how it constructs a prediction set that achieves the coverage guarantee (1.1).

2.1 Quantiles and exchangeability

In the following definitions, we introduce the notions of quantiles, empirical cumulative distribution functions and exchangeability, which are fundamental to conformal prediction. The results in this subsection are stated as facts in [ABB24] (chapter 2.4), and we provide our own proofs for these. The proof of Lemma 2.2 follows the proof of Lemma 1 in [RPC19].

Definition 2.1 (Exchangeability)**.** The random variables $Z_1, ..., Z_n$ are exchangeable if for all $\sigma \in S_n$, we have that

$$(Z_1, \ldots, Z_n) \overset{d}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}),$$

where $\overset{d}{=}$ means equal in distribution. Equivalently, $Z_1, ..., Z_n$ are exchangeable if for any measurable set A and for any $\sigma \in S_n$, we have that

$$\mathbb{P}\left((Z_1, \ldots, Z_n) \in A\right) = \mathbb{P}\left((Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}) \in A\right).$$

Remark 2.1. Note that if $Z_1, \ldots, Z_n$ are exchangeable random variables taking values in $\mathcal{Z}$, then they must be identically distributed. Indeed, for any measurable set $A \in \mathcal{Z}$ and $i \in [n]$, we have that

$$\begin{aligned}
\mathbb{P}\left(Z_i \in A\right) &= \mathbb{P}\left((Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1} \ldots, Z_n) \in \mathcal{Z} \times \cdots \times \mathcal{Z} \times A \times \mathcal{Z} \times \ldots \times \mathcal{Z}\right) \\
&= \mathbb{P}\left((Z_i, \ldots, Z_{i-1}, Z_1, Z_{i+1} \ldots, Z_n) \in \mathcal{Z} \times \cdots \times \mathcal{Z} \times A \times \mathcal{Z} \times \ldots \times \mathcal{Z}\right) \\
&= \mathbb{P}\left(Z_1 \in A\right),
\end{aligned}$$

where we use exchangeability to obtain the second equality. However, exchangeable random variables need not be independent. Indeed, if $Z_1, \ldots, Z_n$ are sampled without replacement from the set $[n]$, then they are exchangeable, since any particular realisation has probability $\frac{1}{n!}$, but $Z_1, \ldots, Z_n$ are certainly not independent. Therefore, exchangeability is a weaker condition than being i.i.d.

Remark 2.2. Another consequence of exchangeability is as follows. Suppose the real-valued random variables $Z_1, \ldots, Z_n$ are almost surely distinct and exchangeable. Taking $A = \{(z_1, \ldots, z_n) \in \mathbb{R}^n : z_1 < z_2 < \cdots < z_n\}$, we have that for any $\sigma \in S_n$,

$$\mathbb{P}\left(Z_1 < \cdots < Z_n\right) = \mathbb{P}\left((Z_1, \ldots, Z_n) \in A\right) = \mathbb{P}\left((Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}) \in A\right) = \mathbb{P}\left(Z_{\sigma(1)} < \cdots < Z_{\sigma(n)}\right).$$

This means that $Z_1, \ldots, Z_n$ are equally likely to appear in any given order.

Definition 2.2. Let $w = (w_1, \ldots, w_n) \in [0, 1]^n$ satisfy $\sum_{i=1}^n w_i = 1$, and let $z \in \mathbb{R}^n$.

(i) The *weighted empirical cumulative distribution function* of z with respect to the *weights* w is the function $\hat{F}_z^w : (-\infty, \infty] \to [0,1]$ given by

$$\hat{F}_z^w(x) \coloneqq \sum_{i=1}^{n} w_i \mathbb{1}\{z_i \leq x\},$$

for $x \in \mathbb{R}$. We define $\hat{F}_z^w(\infty) = 1$.

(ii) The *empirical cumulative distribution function* of z is the function $\hat{F}_z : (-\infty, \infty] \to [0,1]$ given by

$$\hat{F}_z(x) \coloneqq \hat{F}^{(1/n,\ldots,1/n)}(x),$$

for $x \in (-\infty, \infty]$.

(iii) The *weighted quantile function* $\hat{Q}_z^w : (0,1] \to (-\infty, \infty]$ of z with respect to the weights w is defined by

$$\hat{Q}_z^w(\beta) \coloneqq \inf\left\{x \in (-\infty, \infty] : \hat{F}_z^w(x) \geq \beta\right\},$$

for $\beta \in (0,1]$.

(iv) The *quantile function* $\hat{Q}_z : (0,1] \to (-\infty, \infty]$ of z is defined by

$$\hat{Q}_z(\beta) \coloneqq \hat{Q}_z^{(1/n,\ldots,1/n)}(\beta),$$

for $\beta \in (0,1]$.

(v) The k^{th} *order statistic* of z is defined by

$$z_{(k)} \coloneqq \hat{Q}_z(k/n).$$

The weighted versions of the above quantities will play an important role in Section 3. Therefore, some of the results in this section are proven for the weighted quantities, even if we only require the case where all the weights are equal.

Lemma 2.1. *Let $z \in \mathbb{R}^n$, $w = (w_1, \ldots, w_n) \in [0,1]^n$ satisfy $\sum_{i=1}^n w_i = 1$ and $\beta \in (0,1)$.*

(i) We have that $\hat{F}_z^w\left(\hat{Q}_z^w(\beta)\right) \geq \beta$.

(ii) If the components of z are distinct, then $\hat{F}_z\left(\hat{Q}_z(\beta)\right) = \frac{\lceil n\beta \rceil}{n}$.

Proof. (i) First note that by Definition 2.2, it is clear that $\hat{F}_z^w$ is right-continuous. By the definition of the infimum, there exists a sequence $(x_m)_m$ of real numbers satisfying $\hat{F}_z^w(x_m) \geq \beta$ for all m and $x_m \downarrow \hat{Q}_z^w(\beta)$ as $m \to \infty$. By the right continuity of $\hat{F}_z^w$, it follows that $\hat{F}_z^w\left(\hat{Q}_z^w(\beta)\right) \geq \beta$.

(ii) We observe that for all $x \in \mathbb{R}$, we have that $\hat{F}_z(x) \in \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\}$. Since the components of $z = (z_1, \ldots, z_n)$ are distinct, the value of $\hat{F}_z$ increases by $\frac{1}{n}$ at each z_i, for $i \in [n]$. Therefore

$$\hat{F}_z\left(\hat{Q}_z(\beta)\right) = \frac{1}{n} \inf\{k \in \{0\} \cup [n] : \ k/n \geq \beta\} = \frac{\lceil \beta n \rceil}{n}.$$

□

Remark 2.3. Another consequence of the fact that $\hat{F}_z(x) \in \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\}$ is that for any $\beta \in (0,1)$ and $z \in \mathbb{R}^n$,

$$\hat{Q}_z(\beta) = \hat{Q}_z\left(\frac{\lceil \beta n \rceil}{n}\right) = z_{(\lceil \beta n \rceil)}.$$

This follows from the fact that $\hat{F}_z(x) \geq \beta \iff \hat{F}_z(x) \geq \frac{\lceil \beta n \rceil}{n}$ for any $x \in \mathbb{R}$.

We now prove a lemma which underpins the proof of the coverage guarantee in Section 2.2 and links all of the quantities defined above.

Lemma 2.2. *If the random variables $Z_1, \ldots, Z_n$ are exchangeable, then for any $i \in [n]$ and $\beta \in (0,1)$, we have*

$$\mathbb{P}\left(Z_i \leq \hat{Q}_Z(\beta)\right) \geq \beta,$$

where $Z \coloneqq (Z_1, \ldots, Z_n)$. If, moreover, $Z_1, \ldots, Z_n$ are almost surely distinct, then

$$\mathbb{P}\left(Z_i \leq \hat{Q}_Z(\beta)\right) = \frac{\lceil \beta n \rceil}{n}.$$

Proof. Fix $\beta \in (0,1)$. We first claim that the exchangeability of $Z_1, \ldots, Z_n$ implies that

$$\mathbb{P}\left(Z_j \leq \hat{Q}_Z(\beta)\right) = \mathbb{P}\left(Z_1 \leq \hat{Q}_Z(\beta)\right),$$

for any $j \in [n]$. Fix $j \in [n]$ and define $S = \left\{y \in \mathbb{R}^n : y_j \leq \hat{Q}_y(\beta)\right\}$. Define $\tau \in S_n$ to be the transposition exchanging 1 and j. We have that

$$\begin{aligned}
\mathbb{P}\left(Z_j \leq \hat{Q}_Z(\beta)\right) &= \mathbb{P}\left((Z_1, \ldots, Z_n) \in S\right) \\
&= \mathbb{P}\left((Z_{\tau(1)}, \ldots, Z_{\tau(n)}) \in S\right) \\
&= \mathbb{P}\left(Z_1 \leq \hat{Q}_Z(\beta)\right),
\end{aligned}$$

where the second equality follows from exchangeability. This proves our claim.

To complete the proof, we use the deterministic result from Lemma 2.1. By the claim shown above, we have that for any $i \in [n]$,

$$\mathbb{P}\left(Z_i \leq \hat{Q}_Z(\beta)\right) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}\left(Z_j \leq \hat{Q}_Z(\beta)\right) = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\left\{Z_j \leq \hat{Q}_Z(\beta)\right\}\right] = \mathbb{E}\left[\hat{F}_Z(\hat{Q}_Z(\beta))\right].$$

By Lemma 2.1, we have that $\hat{F}_Z(\hat{Q}_Z(\beta)) \geq \beta$. Moreover, if $Z_1, \ldots, Z_n$ are almost surely distinct, then $\hat{F}_Z(\hat{Q}_Z(\beta)) = \frac{\lceil \beta n \rceil}{n}$ almost surely. Taking expectations completes the proof of the lemma. □

Remark 2.4. In the case where $Z_1, \ldots, Z_n$ are almost surely distinct, we can obtain the above result using a simpler argument. Fix $i \in [n]$. As discussed in Remark 2.2, each of the $n!$ orderings of $Z_1, \ldots, Z_n$ are equally likely. For any $k \in [n]$, since there are $(n-1)!$ orderings where Z_i is the k^{th} smallest element of Z, we deduce that the probability that Z_i is the k^{th} smallest element of Z is $1/n$. Summing over k ranging from 1 to $\lceil \beta n \rceil$ and using Remark 2.3 gives that

$$\mathbb{P}\left(Z_i \leq \hat{Q}_Z(\beta)\right) = \frac{\lceil \beta n \rceil}{n}.$$

A standard justification for the validity of conformal prediction ([Lei+18; AB21]) involves noting that ranks of exchangeable random variables are uniformly distributed. This relies on the argument given above. Our presentation in Section 2.2 will not take this approach for two reasons. Firstly, it does not deal with the case when there are ties. Secondly, the extensions to conformal prediction presented in Section 3 are more easily expressed using the notation of empirical cumulative distribution functions and quantiles.

2.2 Full conformal prediction

In this subsection, we present the full conformal prediction method. The main theorem of this subsection, Theorem 2.1, will provide a way to construct a prediction set with a valid coverage guarantee as in (1.1). The presentation of the material in this subsection and the proof of Theorem 2.1 is inspired by [ABB24]. In the rest of this essay, we denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

We first introduce the notion of a *conformity score*. A conformity score is a function

$$s : \mathcal{Z} \times \cup_{j \geq 1} \mathcal{Z}^j \to \mathbb{R}.$$

Whilst the conformity score can be an arbitary function, we now explain the intuition behind the conformity score and how it is typically constructed. The first argument of the conformity score represents a test point and the second argument a training dataset. The conformity score measures the discrepancy between the test point and a model fitted using the training dataset; a large, positive value of the conformity score indicates that the test point "conforms" poorly with the fitted model. In particular, we note that computing the conformity score involves fitting a model trained on the data in the second argument. We will also refer to $s(z; D)$ as the conformity score of z with respect to the data D.

Example 1. Consider the regression setting where $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}$ for some positive integer p. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{Z}$ and each $(X_i, Y_i) \sim P$ for $i \in [n]$ and some distribution P. Suppose $(X, Y) \sim P$ is independent of $((X_i, Y_i))_{i=1}^n$ and $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$ is an estimate of the regression function $x \mapsto \mathbb{E}(Y|X = x)$ trained on $((X_i, Y_i))_{i=1}^n$.

An important example of a conformity score in this case is the *absolute residual score* given by

$$s((x, y); ((X_1, Y_1), \ldots, (X_n, Y_n))) = |y - \hat{\mu}(x)|.$$

Example 2. Consider the classification setting where $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = [K]$ for some positive integers p and K. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{Z}$ and each $(X_i, Y_i) \sim P$ for $i \in [n]$ and some distribution P. Suppose $(X, Y) \sim P$ is independent of $((X_i, Y_i))_{i=1}^n$ and $\hat{p}_k : \mathcal{X} \to [0, 1]$ is an estimate of the conditional class probability $x \mapsto \mathbb{P}(Y = k \mid X = x)$ trained on $((X_i, Y_i))_{i=1}^n$. The *high-probability score* [ABB24] is given by

$$s((x, y); (X_1, Y_1), \ldots, (X_n, Y_n)) = -\hat{p}_y(x).$$

We now introduce the notion of a symmetric conformity score, which is one that does not depend on the order in which the training data points are provided.

Definition 2.3. A conformity score s is *symmetric* if for any $z \in \mathcal{Z}$, positive integer j and $D \in \mathcal{Z}^j$, we have that

$$s(z; D) = s(z; \sigma(D)),$$

for all $\sigma \in S_j$.

Before proceeding to the formal statement of Theorem 2.1, we give a brief informal description of how the prediction set will be constructed. We will consider the conformity score of each (X_i, Y_i) with respect to $((X_i, Y_i))_{i=1}^{n+1}$. We will show that if s is symmetric, then these conformity scores are exchangeable. By Lemma 2.2, it will follow that the probability that the conformity score of the test point (X_{n+1}, Y_{n+1}) lies in the bottom $1-\alpha$ fraction of all the conformity scores is at least $1-\alpha$. This is the key idea used to construct the prediction set. However, since Y_{n+1} is unknown, we will instead consider the test point (X_{n+1}, y) and include those $y \in \mathcal{Y}$ in the prediction set, whose conformity score falls in the bottom $1-\alpha$ fraction of the conformity scores.

We now introduce the notation used in Theorem 2.1. Write

$$D = ((X_i, Y_i))_{i=1}^{n+1}, \quad \text{and} \quad D^y = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)),$$

for any $y \in \mathcal{Y}$. Write

$$S_i = s((X_i, Y_i); D), \quad S_i^y = s((X_i, Y_i); D^y),$$

for $i \in [n]$ and

$$S_{n+1}^y = s((X_{n+1}, y); D^y).$$

Additionally, let

$$S^y = (S_1^y, \ldots, S_{n+1}^y).$$

Theorem 2.1. *Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ are exchangeable and s is a symmetric conformity score. Define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \quad S_{n+1}^y \leq \hat{Q}_{S^y}(1-\alpha) \right\}. \tag{2.1}$$

Then we have that

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

If, moreover, the scores $S_1, \ldots, S_{n+1}$ are almost surely distinct, then we have that

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

Proof. We begin by showing that $S_1, \ldots, S_{n+1}$ are exchangeable, as mentioned in the discussion above. Making the dependence on D explicit, note that $S_i = s((X_i, Y_i); D)$ and $S_{\sigma(i)} = s((X_{\sigma(i)}, Y_{\sigma(i)}); D)$ for any $i \in [n+1]$ and $\sigma \in S_{n+1}$. Define also the function $h : \mathcal{Z}^{n+1} \to \mathbb{R}^{n+1}$ by

$$h : z \mapsto (s(z_1; z), \ldots, s(z_{n+1}; z)),$$

for any $z = (z_1, \ldots, z_{n+1}) \in \mathcal{Z}^{n+1}$. Then we have that for any measurable set A,

$$\begin{aligned}
\mathbb{P}\left(\left(S_{\sigma(1)}, \ldots, S_{\sigma(n+1)}\right) \in A\right) &= \mathbb{P}\left(\left(s((X_{\sigma(1)}, Y_{\sigma(1)}); D), \ldots, s((X_{\sigma(n+1)}, Y_{\sigma(n+1)}); D)\right) \in A\right) \\
&= \mathbb{P}\left(\left(s((X_{\sigma(1)}, Y_{\sigma(1)}); \sigma(D)), \ldots, s((X_{\sigma(n+1)}, Y_{\sigma(n+1)}); \sigma(D))\right) \in A\right) \\
&= \mathbb{P}\left(h(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) \in A\right),
\end{aligned}$$

where the second equality follows from the symmetry of the score function. By the exchangeability of $Z_1, \ldots, Z_{n+1}$, we have that

$$\begin{aligned}
\mathbb{P}\left(h(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) \in A\right) &= \mathbb{P}\left((Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) \in h^{-1}(A)\right) \\
&= \mathbb{P}\left((Z_1, \ldots, Z_{n+1}) \in h^{-1}(A)\right) \\
&= \mathbb{P}\left(h(Z_1, \ldots, Z_{n+1}) \in A\right) \\
&= \mathbb{P}\left((S_1, \ldots, S_{n+1}) \in A\right),
\end{aligned}$$

which shows that

$$\mathbb{P}\left(\left(S_{\sigma(1)},\ldots,S_{\sigma(n+1)}\right)\in A\right)=\mathbb{P}\left((S_1,\ldots,S_{n+1})\in A\right).$$

Therefore, $S_1,\ldots,S_{n+1}$ are exchangeable.

Finally, denoting $S\coloneqq(S_1,\ldots,S_{n+1})$, note that since $D^{Y_{n+1}}=D$, we have that

$$Y_{n+1}\in C(X_{n+1})\iff S_{n+1}\leq\hat{Q}_S(1-\alpha)$$

by the definition of $C(X_{n+1})$. Therefore, Lemma 2.2 implies that

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

and that if $S_1,\ldots,S_{n+1}$ are almost surely distinct, then

$$\mathbb{P}\left(Y_{n+1}\in C(X_{n+1})\right)=\frac{\lceil(1-\alpha)(n+1)\rceil}{n+1}\leq 1-\alpha+\frac{1}{n+1}.$$

□

We now present a lemma that we will use to derive a slightly different characterisation of the prediction set (2.1). The proof is inspired by [RPC19] (Appendix A, Lemma 2), but we present it using empirical cumulative distribution functions and extend it to weighted quantiles.

Lemma 2.3. *Let $z\in\mathbb{R}^n$, $y\in\mathbb{R}$ and $w\in[0,1]^n$ satisfy $\sum_{i=1}^n w_i=1$. We have that for any $\beta\in(0,1)$,*

$$y\leq\hat{Q}^w_{(z,y)}(\beta)\iff y\leq\hat{Q}^w_{(z,\infty)}(\beta).$$

Proof. First note that for any $x\in(-\infty,\infty]$,

$$\hat{F}^w_{(z,\infty)}(x)=\sum_{i=1}^n w_i\mathbb{1}\left\{x\geq z_i\right\}+w_{n+1}\mathbb{1}\left\{x\geq\infty\right\}\leq\sum_{i=1}^n w_i\mathbb{1}\left\{x\geq z_i\right\}+w_{n+1}\mathbb{1}\left\{x\geq y\right\}=\hat{F}^w_{(z,y)}(x).$$

Therefore, $\hat{F}^w_{(z,y)}(\hat{Q}^w_{(z,\infty)}(\beta))\geq\hat{F}^w_{(z,\infty)}(\hat{Q}^w_{(z,\infty)}(\beta))\geq\beta$, which implies that

$$\hat{Q}^w_{(z,\infty)}(\beta)\geq\hat{Q}^w_{(z,y)}(\beta).\tag{2.2}$$

This proves the forward direction of the lemma. Now suppose that $y>\hat{Q}^w_{(z,y)}(\beta)$. Then we have that

$$\begin{aligned}\hat{F}^w_{(z,\infty)}(\hat{Q}^w_{(z,y)}(\beta))&=\sum_{i=1}^n w_i\mathbb{1}\left\{\hat{Q}^w_{(z,y)}(\beta)\geq z_i\right\}+w_{n+1}\mathbb{1}\left\{\hat{Q}^w_{(z,y)}(\beta)\geq\infty\right\}\\&=\hat{F}^w_{(z,y)}(\hat{Q}^w_{(z,y)}(\beta))+w_{n+1}\left(\mathbb{1}\left\{\hat{Q}^w_{(z,y)}(\beta)\geq\infty\right\}-\mathbb{1}\left\{\hat{Q}^w_{(z,y)}(\beta)\geq y\right\}\right)\\&=\hat{F}^w_{(z,y)}(\hat{Q}^w_{(z,y)}(\beta))\geq\beta.\end{aligned}$$

This implies that

$$\hat{Q}^w_{(z,y)}(\beta)\geq\hat{Q}^w_{(z,\infty)}(\beta),$$

which, together with (2.2), also implies

$$\hat{Q}^w_{(z,y)}(\beta)=\hat{Q}^w_{(z,\infty)}(\beta).$$

Thus, $y>\hat{Q}^w_{(z,\infty)}(\beta)$, which completes the proof of the lemma. □

The following lemma explains how to compute the quantity $\hat{Q}_{(z,\infty)}(\beta)$ appearing in Lemma 2.3. This lemma is a standard fact used in conformal prediction (e.g., see [Lei+18]) and we give our own proof of it.

Lemma 2.4. *Let $z \in \mathbb{R}^n$ and $\beta \in (0,1)$. Then we have that*

$$\hat{Q}_{(z,\infty)}(\beta) = \begin{cases} \hat{Q}_z\left(\frac{\lceil \beta(n+1) \rceil}{n}\right) & \text{if } \frac{\lceil \beta(n+1) \rceil}{n} \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

Proof. First note that the condition $\frac{\lceil \beta(n+1) \rceil}{n} \leq 1$ is equivalent to $\beta \leq \frac{n}{n+1}$. If $\beta > \frac{n}{n+1}$, then $\hat{Q}_{(z,\infty)}(\beta) = \infty$ since $\frac{1}{n+1}\sum_{i=1}^n \mathbb{1}\{x \geq z_i\} \leq \frac{n}{n+1}$ if $x < \infty$. Now observe that for any $x < \infty$, we have that

$$\hat{F}_{(z,\infty)}(x) = \frac{1}{n+1}\sum_{i=1}^n \mathbb{1}\{x \geq z_i\} + \frac{1}{n+1}\mathbb{1}\{x \geq \infty\} = \frac{n}{n+1}\hat{F}_z(x).$$

If $\beta \leq \frac{n}{n+1}$, we have that

$$\frac{\beta(n+1)}{n} \leq \frac{n+1}{n}\hat{F}_{(z,\infty)}(\hat{Q}_{(z,\infty)}(\beta)) = \hat{F}_z(\hat{Q}_{(z,\infty)}(\beta)),$$

which implies that

$$\hat{Q}_{(z,\infty)}(\beta) \geq \hat{Q}_z\left(\frac{\beta(n+1)}{n}\right) = \hat{Q}_z\left(\frac{\lceil \beta(n+1) \rceil}{n}\right),$$

by Remark 2.3. We also have that

$$\hat{F}_{(z,\infty)}\left(\hat{Q}_z\left(\frac{\lceil \beta(n+1) \rceil}{n}\right)\right) = \frac{n}{n+1}\hat{F}_z\left(\hat{Q}_z\left(\frac{\lceil \beta(n+1) \rceil}{n}\right)\right) = \frac{n}{n+1}\frac{\lceil \beta(n+1) \rceil}{n} \geq \beta.$$

This shows that

$$\hat{Q}_z\left(\frac{\lceil \beta(n+1) \rceil}{n}\right) = \hat{Q}_{(z,\infty)}(\beta).$$

□

The key implication of Lemma 2.4 is that we can reformulate the inequality in (2.1) in a way that only the left-hand-side depends on S_{n+1}^y. This will be important in Section 2.3 and is recorded in Theorem 2.2 below, which is an immediate consequence of Lemma 2.3.

Theorem 2.2. *Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ are exchangeable and s is a symmetric conformity score. Define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \quad S_{n+1}^y \leq \hat{Q}_{(S_1^y, \ldots, S_n^y, \infty)}(1-\alpha) \right\}.$$

Then we have that

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

If, moreover, the scores $S_1, \ldots, S_{n+1}$ are almost surely distinct, then we have that

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

We make two remarks on the above results. Firstly, we note that in the case $\frac{\lceil (1-\alpha)(n+1) \rceil}{n} > 1$, we have that $C(X_{n+1}) = \mathcal{Y}$ since $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)} = \infty$. However, since $\frac{\lceil (1-\alpha)(n+1) \rceil}{n} > 1$ is equivalent to $\alpha < \frac{1}{n+1}$, we can ignore this situation provided α and n are not too small. Secondly, we note that if $z \in \mathbb{R}^n$, then

$$\hat{Q}_z \left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n} \right) = z_{(\lceil (1-\alpha)(n+1) \rceil)},$$

by Remark 2.3. This means that we can calculate $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}$ by simply finding the $\lceil (1-\alpha)(n+1) \rceil$ smallest element in the list $(\hat{S}_1, \ldots, \hat{S}_n)$.

Theorem 2.2 and the above remarks lead to Algorithm 1 given below.

Algorithm 1 Full conformal prediction algorithm

Input: Data $((X_i, Y_i))_{i=1}^n$; test predictor X_{n+1}; miscoverage level $\alpha \in (0,1)$, conformity score s.
  Initialise $C \leftarrow \emptyset$
  if $\lceil (1-\alpha)(n+1) \rceil > n$ **then**
    $C \leftarrow \mathcal{Y}$
  else
    for $y \in \mathcal{Y}$ **do**
      Compute $S_i^y = s((X_i, Y_i); D^y)$ for $i \in [n]$.
      Compute $S_{n+1}^y = s((X_{n+1}, y); D^y)$.
      Compute $\hat{Q}$ as the $\lceil (1-\alpha)(n+1) \rceil$ element in the list $S_1^y, \ldots, S_n^y$.
      if $S_{n+1}^y \leq \hat{Q}$ **then**
        $C \leftarrow C \cup \{y\}$.
      end if
    end for
  end if
Output: C

It is important to note that since computing a conformity score involves fitting a model, implementing Algorithm 1 would require fitting a model for each $y \in \mathcal{Y}$. Therefore, Algorithm 1 is, in general, computationally expensive to implement. The method presented in Section 2.3 will resolve this issue. Furthermore, note that if $\mathcal{Y}$ is not discrete, we could not, in general, implement Algorithm 1 exactly since it iterates over $y \in \mathcal{Y}$. However, if $\mathcal{Y}$ is discrete (as in the classification setting), then this is possible.

2.3 Split conformal prediction

In this subsection, we will present the split conformal prediction algorithm. We will show that it is, in fact, a special case of full conformal prediction, and so it satisfies the coverage guarantee from Theorem 2.2. We will also compare full and split conformal prediction, discussing their respective advantages and disadvantages. The presentation of the material in this subsection is inspired by [ABB24; AB21; Tib24a]. The proof of Theorem 2.3 follows the ideas in section 3.4 of [ABB24].

In the context of split conformal prediction, we refer to the data $D_{\text{cal}} \coloneqq ((X_i, Y_i))_{i=1}^n$ as the *calibration data*. We assume that we are given a function $\hat{s} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that depends on a *proper training set* $D_{\text{tr}} \in \cup_{j \geq 1} \mathcal{Z}^j$ that is independent of the calibration data. To construct the prediction set, we will only require access to $\hat{s}$ and not D_{tr}. For split conformal prediction, we refer to $\hat{s}$ as a conformity score; the proof of Theorem 2.3 below will explain why this is consistent with our notion of a conformity score

from Section 2.2. Split conformal prediction uses the calibration data, $\hat{s}$ and X_{n+1} to form a prediction set for Y_{n+1}.

We use the following notation. Let $\hat{S}_i = \hat{s}(X_i, Y_i)$ for $i \in [n]$ and let $\hat{S}^y_{n+1} = \hat{s}(X_{n+1}, y)$ for any $y \in \mathcal{Y}$.

Theorem 2.3. *Suppose $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ are exchangeable. Define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ \hat{S}^y_{n+1} \leq \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1-\alpha) \right\}.$$

Then we have that

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

If, moreover, the scores $\hat{S}_1, \ldots, \hat{S}_{n+1}$ are almost surely distinct, then we have that

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

Proof. We work conditional on D_{tr}, so we may treat $\hat{s}$ as a fixed function. The key observation is that since Theorem 2.2 holds for any symmetric conformity score, we may choose the conformity score to be independent of its second argument. Define the conformity score $s(z; \tilde{D}) = \hat{s}(z)$ for all $z \in \mathcal{Z}$ and $\tilde{D} \in \cup_{j \geq 1} \mathcal{Z}^j$. Since this is independent of $\tilde{D}$, it is certainly symmetric. In the notation of Theorem 2.2, we then have that

$$S_i = S^y_i = \hat{S}_i \quad \text{and} \quad S^y_{n+1} = \hat{s}(X_{n+1}, y),$$

for all $y \in \mathcal{Y}$. Therefore, the prediction set in Theorem 2.2 takes exactly the form stated above, so we have that

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1}) \mid D_{\text{tr}}\right) \geq 1 - \alpha.$$

If $\hat{S}_1, \ldots, \hat{S}_n$ are almost surely distinct, then

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1}) \mid D_{\text{tr}}\right) \leq 1 - \alpha + \frac{1}{n+1}.$$

The result follows by taking the expectation over D_{tr}. □

Algorithm 2 Split conformal prediction algorithm

Input: Calibration $((X_i, Y_i))_{i=1}^n$; test predictor X_{n+1}; miscoverage level $\alpha \in (0, 1)$, conformity score $\hat{s}$.
  Initialise $C \leftarrow \emptyset$
  if $\lceil (1-\alpha)(n+1) \rceil > n$ **then**
    $C \leftarrow \mathcal{Y}$
  else
    Compute $\hat{S}_i = \hat{s}(X_i, Y_i)$ for each $i \in [n]$.
    for $y \in \mathcal{Y}$ **do**
      Compute $\hat{s}(X_{n+1}, y)$.
      Compute $\hat{Q}$ as the $\lceil (1-\alpha)(n+1) \rceil$ smallest element in the list $\hat{S}_1, \ldots, \hat{S}_n$.
      if $\hat{s}(X_{n+1}, y) \leq \hat{Q}$ **then**
        $C \leftarrow C \cup \{y\}$.
      end if
    end for
  end if
Output: C

Example 3. Consider the regression setting as in Example 1. Suppose $\hat{\mu}$ is an estimate of the regression function obtained from a proper training set D_{tr}. In the case of split conformal prediction, we refer to the conformity score

$$\hat{s}(x, y) = |y - \hat{\mu}(x)|$$

as the *absolute residual score*. In this case, the prediction set is an interval centered at $\hat{\mu}(x)$ given by

$$\begin{aligned} C(x) &= \left\{ y \in \mathbb{R} : \, |y - \hat{\mu}(x)| \leq \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha) \right\} \\ &= \left[\hat{\mu}(X_{n+1}) - \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha), \; \hat{\mu}(x) + \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha) \right]. \end{aligned} \tag{2.3}$$

It is important to note that this simplified form of the prediction set is a consequence of Lemma 2.3 and the split conformal prediction procedure. Specifically, this form of the prediction set arises from the fact that neither the estimated regression function, nor the quantile used in the definition of the prediction set depend on y. The former is a consequence of the split conformal prediction procedure, which ensures $\hat{\mu}$ only depends on D_{tr}. The latter is a consequence of both the split conformal prediction algorithm, which ensures $S_i^y = \hat{s}(X_i, Y_i)$ is independent of y, and Lemma 2.3, which ensures the quantile used in the prediction set depends only on D_{cal}.

An important interpretation of the above results is as follows. We have shown that given both an arbitary pretrained model and a notion of error in the form of a conformity score $\hat{s}$, we can apply split conformal prediction to obtain prediction sets with valid coverage. Note that, in practice, if $\hat{s}$ is not given, we may split the training set into independent proper training and calibration sets, and obtain the conformity score $\hat{s}$ by fitting a model to the proper training set.

Split conformal prediction has a clear advantage in terms of computational efficiency compared to full conformal prediction and is, therefore, typically used to implement conformal prediction in practice. Indeed, split conformal prediction only requires fitting the model once to obtain $\hat{s}$, whereas in full conformal predicition, the model must be fitted for each $y \in \mathcal{Y}$ - as discussed at the end of Section 2.2. Algorithm 2 suggests that - in general - we must still check whether each $y \in \mathcal{Y}$ is included in the prediction set. However, we see in Example 3 and will see in Section 2.4 that several choices of $\hat{s}$ in the regression setting lead to prediction intervals, which can be efficiently described by their endpoints. A further advantage of using split conformal prediction is that it imposes no restrictions on D_{tr} and the fitting procedure used to obtain $\hat{s}$, whereas full conformal prediction requires the conformity score to be symmetric. However, we note that full conformal prediction uses all of the data to fit the model, whereas in split conformal prediction, only the proper training set is used to fit the model.

We have shown that split conformal prediction is valid for any conformity score $\hat{s}$. In particular, split conformal prediction remains valid even if $\hat{s}$ is obtained using a model that is poorly fitted to D_{tr}. This raises the question of how the fitting procedure used to obtain $\hat{s}$ affects the prediction set. We now present a numerical experiment illustrating that a better fitting procedure is more desirable as it typically yields narrower prediction intervals.

We use simulated data generated as follows:

$$\begin{aligned} X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} \text{Uniform}(-5, 5) \\ Y_i | X_i &\overset{\text{indep.}}{\sim} \mathcal{N}\left(\mu(X_i), \; 0.5^2\right), \end{aligned}$$

for each i, where

$$\mu(x) = \frac{1}{1 + x^2} + \frac{2}{1 + (x - 3)^2}.$$

Given a test dataset D_{test} of data points in $\mathcal{Z}$, the *empirical coverage* is given by

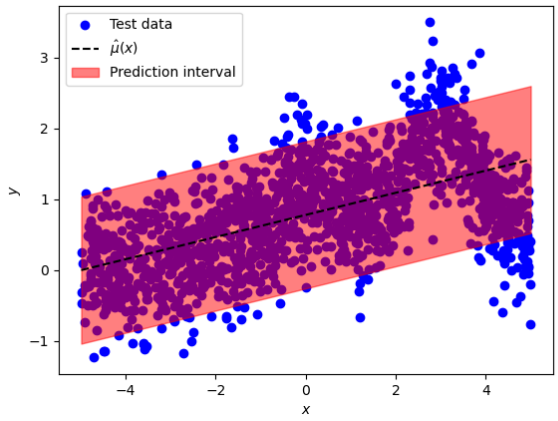
$$\frac{1}{|D_{\text{test}}|} \sum_{(X,Y) \in D_{\text{test}}} \mathbb{1}\{Y \in C(X)\}. \tag{2.4}$$

We set $\alpha = 0.1$ and generate i.i.d. proper training, calibration and test datasets (as described above) with $1000, 1500, 1500$ data points, respectively. We train regression models on the proper training set using both linear regression and random forests. We then compute the conformity scores on the calibration set with the absolute residual score and generate split conformal prediction intervals for each data point in the test dataset. We calculate both the empirical coverage (as in (2.4)) and the average length of the prediction intervals for the test dataset. We repeat this over 2000 independent draws of the calibration and test datasets. The code is provided in the file `2_3_example.py`.

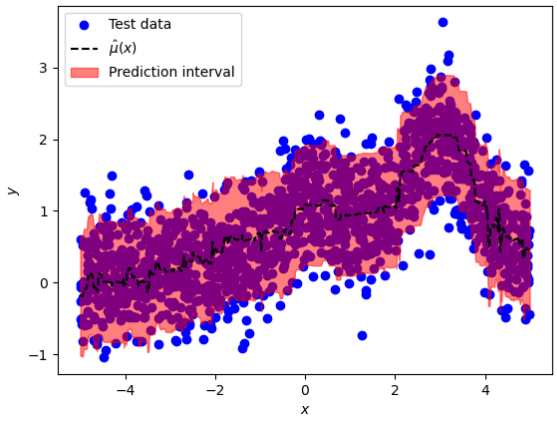
The first row of Figure 1 plots the test dataset, the fitted regression function and the split conformal prediction interval for a single draw of the calibration and test datasets. The second row plots a histogram of the conformity scores for the same instance of the calibration and test datasets. The third row plots a histogram of the average lengths of the prediction intervals obtained in each of the 2000 draws of the calibration and test datasets.

We estimate the coverage by averaging the empirical coverages over the 2000 draws of the calibration and test datasets. For this numerical experiment, we find that the estimated coverage is 0.89979 and 0.90046 for the linear regression and random forest models, respectively. This is expected by Theorem 2.3 since the data is i.i.d and thus exchangeable. Although both methods provide the desired coverage, it is clear from Figure 1 that the linear regression model underfits the data and the random forests model provides a much improved fit. This is reflected in the conformal prediction intervals by the fact that the distribution of the average length of the prediction intervals in the random forest model is concentrated around smaller values than in the linear regression model, as seen in the third row of Figure 1. Mathematically, we can explain this by considering (2.3). Due to the improved fit of the random forests model, the distribution of the absolute residual scores of the calibration data points is more skewed towards zero. Therefore, the quantile $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)$ is lower for the random forests model, and so the prediction intervals are narrower. This is confirmed by the second row of plots in Figure 1.

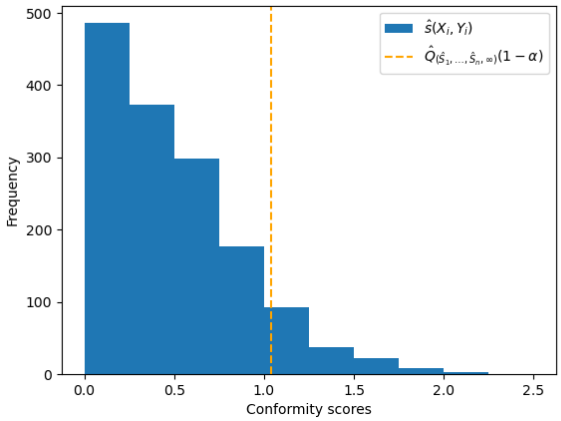
Overall, this numerical experiment highlights that whilst any model fitting procedure can be used to obtain coverage, it is still desirable to use a procedure that fits the data well, as the resulting conformal prediction intervals are on average narrower.



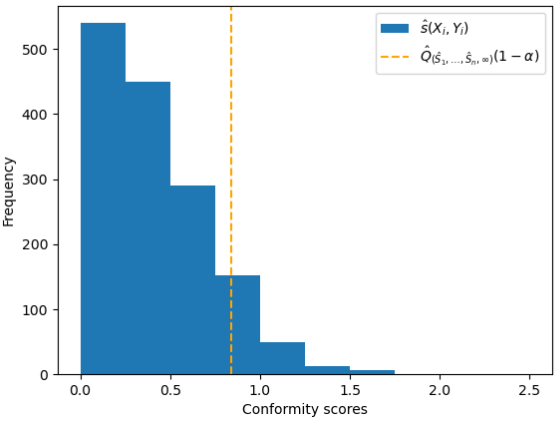
(a) Linear regression



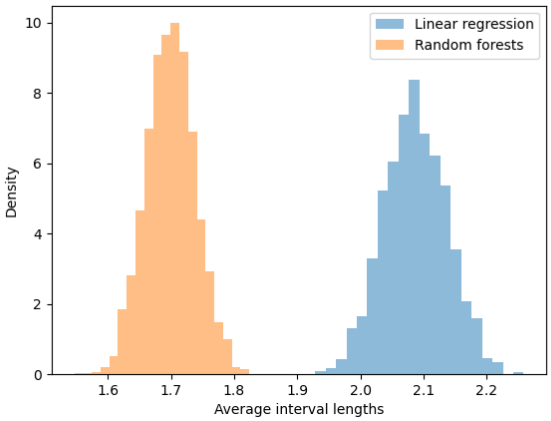
(b) Random forests



(c) Linear regression



(d) Random forests



(e) Distribution of average prediction interval lengths over 2000 independent draws of the calibration and test datasets.

Figure 1: First row: plot of the split conformal prediction intervals. Second row: histogram of the conformity scores. Third row: distribution of the average prediction interval length.

2.3.1 Training-conditional coverage

An important observation regarding the coverage guarantee of conformal predicition in both Theorem 2.1 and Theorem 2.3 is that it provides *marginal coverage*. This refers to the fact that the probability $\mathbb{P}(Y_{n+1} \in C(X_{n+1}))$ is marginalising over $((X_i, Y_i))_{i=1}^{n+1}$. One reason why this may be an issue, is that it does not guarantee coverage for any given value of the test predictor X_{n+1}, i.e. it does not guarantee *test-conditional coverage*:

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

This means that conformal prediction could overcover in some regions of $\mathcal{X}$ and undercover in other regions, whilst still achieving at least $1 - \alpha$ coverage on average. Lack of test-conditional coverage is a significant disadvantage of conformal prediction, but we will not discuss it further in this essay.

Instead, we will consider the *training-conditional* coverage [ABB24; Tib24a]

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1}) \mid (X_1, Y_1), \ldots, (X_n, Y_n))$$

in this subsection. Note that this is a random quantity since it is a function of $((X_i, Y_i))_{i=1}^{n}$. The presentation of this subsection is inspired by [AB21; ABB24; Tib24a].

In the case of split conformal prediction, it is possible to exactly derive the distribution of this quantity. We provide our own proof of this result, following the steps outlined in [Tib24b].

Lemma 2.5 ([Tib24a])**.** *Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are i.i.d. and suppose that (conditional on D_{tr}) $\hat{S}_i$ has a continuous cumulative distribution function F. Then*

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1}) \,|\, D_{\text{cal}}) \sim \text{Beta}(k_\alpha, n + 1 - k_\alpha),$$

where $k_\alpha = \lceil (1 - \alpha)(n + 1) \rceil$.

Proof. Throughout, we work conditional on D_{tr}, so that $\hat{s}$ is fixed. We first claim that if $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$ with order statistics $U_{(1)} \leq \cdots \leq U_{(n)}$, then for any $k \in [n]$,

$$U_{(k)} \sim \text{Beta}(k, n + 1 - k).$$

Note that for any $x \in \mathbb{R}$, we have that

$$\mathbb{P}\left(U_{(k)} \leq x\right) = \sum_{r=k}^{n} \binom{n}{r} x^r (1 - x)^{n-r}$$

since $U_{(k)} \leq x$ if and only if at least k of the random variables $U_1, \ldots, U_n$ are less than or equal to x. Note also that $\mathbb{P}(U_i \leq x) = x$ for any $i \in [n]$. Therefore, if $g_r(z) = z^r (1 - z)^{n-r}$, then we have that

$$\begin{aligned}
\lim_{\epsilon \to 0} \frac{\mathbb{P}\left(x \leq U_{(k)} \leq x + \epsilon\right)}{\epsilon} &= \sum_{r=k}^{n} \lim_{\epsilon \to 0} \frac{g_r(x + \epsilon) - g_r(x)}{\epsilon} \\
&= \sum_{r=k}^{n} \binom{n}{r} g_r'(x) \\
&= \sum_{r=k}^{n} \binom{n}{r} r x^{r-1} (1 - x)^{n-r} - \sum_{r=k}^{n} \binom{n}{r} (n - r) x^r (1 - x)^{n-r-1} \\
&= n \sum_{r=k-1}^{n-1} \binom{n-1}{r} r x^{r-1} (1 - x)^{n-r} - n \sum_{r=k}^{n-1} \binom{n-1}{r} (n - r) x^r (1 - x)^{n-r-1} \\
&= n \binom{n-1}{k-1} x^{k-1} (1 - x)^{n-k},
\end{aligned}$$

which is the density of a $\text{Beta}(k, n+1-k)$ distribution.

Let F be the cumulative distribution function of $\hat{S}_1, \ldots, \hat{S}_{n+1}$. It is a standard result (see Appendix, Section A) that $F(\hat{S}_1), \ldots, \hat{F}(S_{n+1}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$. Therefore, for any $k \in [n]$

$$\mathbb{P}\left(\hat{S}_{n+1} \leq \hat{S}_{(k)} \,|\, D_{\text{cal}}\right) = F(\hat{S}_{(k)}) \overset{\text{d}}{=} U_{(k)} \sim \text{Beta}(k, n+1-k)$$

since F is increasing. Finally, we note that

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1}) \,|\, D_{\text{cal}}\right) = \mathbb{P}\left(\hat{S}_{n+1} \leq \hat{S}_{(k_\alpha)} \,|\, D_{\text{cal}}\right),$$

and so the result follows. □

A further question is as follows: what if we fix the calibration set and compute the empirical coverage on a finite number of test points?

As above, suppose D_{cal} consists of i.i.d. calibration data points, and suppose also that $\tilde{Z}_1, \ldots, \tilde{Z}_{n_{\text{test}}} \in \mathcal{Z}$ are i.i.d. test points. Writing $\tilde{Z}_j = (\tilde{X}_j, \tilde{Y}_j)$, we then have that

$$\frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \mathbb{1}\left\{\tilde{Y}_j \in C(\tilde{X}_j)\right\} \mid D_{\text{cal}} \sim \frac{1}{n_{\text{test}}} \text{Bin}(n_{\text{test}}, \mu)$$
$$\mu \sim \text{Beta}(k_\alpha, n+1-k_\alpha)$$

by Lemma 2.5, where Bin denotes the binomial distribution. This means that

$$\frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \mathbb{1}\left\{\tilde{Y}_j \in C(\tilde{X}_j)\right\} \sim \frac{1}{n_{\text{test}}} \text{BetaBin}(n_{\text{test}}, k_\alpha, n+1-k_\alpha),$$

where BetaBin is the beta-binomial distribution. Therefore, the empirical coverage (2.4) is a single draw from this distribution. This result is also discussed in [AB21; Tib24a]. In Figure 2, we plot a histogram of empirical coverages from the numerical example in Section 2.3 and overlay the beta-binomial probability mass function, illustrating the above result in this example. The code used to generate Figure 2 is provided in the file `training_conditional.py`.

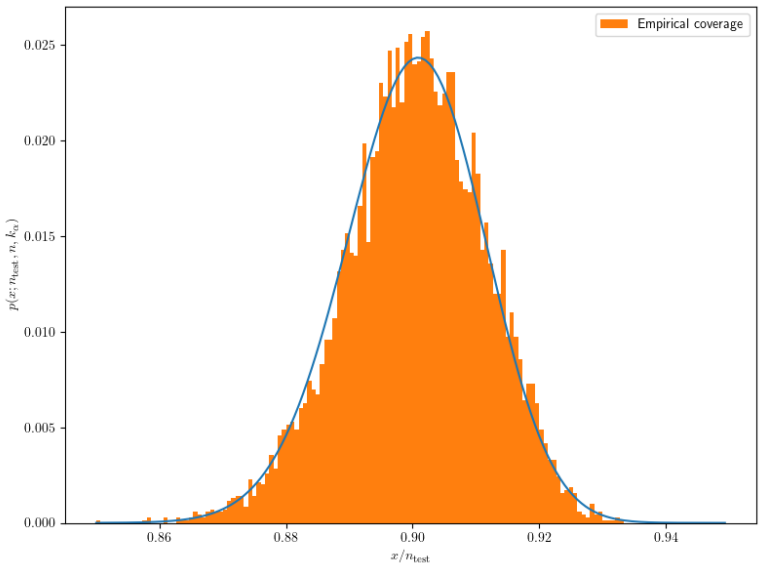


Figure 2: Histogram of 7000 samples of the empirical coverage (generated for the example in Section 2.3) and the beta-binomial probability mass function overlayed.

2.4 Choice of conformity score

In Section 2.3, we showed that any conformity score can be used to construct the split conformal prediction set (2.3). However, it is not immediately clear from (2.3) how the choice of conformity score affects the prediction set. We gain some insight into this from Example 3, where we see that for split conformal prediction with the absolute residual score, the conformity score influences the width of the prediction interval through the quantile $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)$. In this subsection, we further explore how the choice of conformity score affects the properties of the resulting prediction set. In addition to the absolute residual score, we consider two further examples of conformity scores in the regression setting and compare their empirical performance through numerical experiments. Throughout this subsection, we use split conformal prediction.

We work in the regression setting as in Example 3 with $\mathcal{X} = \mathbb{R}$. Consider the absolute residual score and its corresponding predicition interval (2.3). We observe that a consequence of using the absolute residual score is that the prediction interval has a constant width for all $x \in \mathbb{R}$. If the data generating process is heteroscedastic, i.e. $\text{Var}(Y|X = x)$ is not constant in x, then the prediction interval (2.3) does not accurately capture the uncertainty in Y given $X = x$.The two conformity scores we present aim to make the prediction interval adaptive to heteroscedasticity.

2.4.1 Locally Weighted Residual Score

If $\hat{\mu}$ is an estimate of the regression function $\mu : x \mapsto \mathbb{E}(Y|X = x)$ and $\hat{\sigma}$ is an estimate of the *conditional mean absolute deviation* $x \mapsto \mathbb{E}(|Y - \mu(X)| \mid X = x)$, then the *locally weighted score* is the conformity score given by

$$\hat{s}(x, y) = \frac{|y - \hat{\mu}(x)|}{\hat{\sigma}(x)}.$$

The corresponding prediction set is given by

$$C(x) = \left[\hat{\mu}(x) - \hat{\sigma}(x)\,\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha),\ \hat{\mu}(x) + \hat{\sigma}(x)\,\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)\right],$$

using the notation of Section 2.3.

This conformity score was originally introduced by [Lei+18] and aims to account for heteroscedasticity by scaling the width of the interval in (2.3) by $\hat{\sigma}(x)$ for each $x \in \mathcal{X}$. In practice, $\hat{\sigma}$ can be estimated by first regressing Y_i onto X_i for $(X_i, Y_i) \in D_{\text{tr}}$ to obtain $\hat{\mu}$ and then regressing $|Y_i - \hat{\mu}(X_i)|$ onto X_i for $(X_i, Y_i) \in D_{\text{tr}}$.

2.4.2 Conformalised Quantile Regression

A second approach to generate prediction intervals that are adaptive to heteroscedasticity is to estimate the conditional quantile function

$$q_\tau(x) = \inf\left\{z \in \mathbb{R} : \mathbb{P}\left(Y \leq z | X = x\right) \geq \tau\right\}, \quad \tau \in (0,1)$$

directly. This is motivated by noting that

$$\mathbb{P}\left(Y \in [q_{\alpha/2}(X), q_{1-\alpha/2}(X)] \mid X\right) = 1 - \alpha,$$

i.e. the interval $[q_{\alpha/2}(X), q_{1-\alpha/2}(X)]$ has exact $(1-\alpha)-$level coverage conditional on X. The approach of estimating $q_\tau(x)$ is referred to as *quantile regression*. In this essay, we do not discuss methods for constructing quantile regression estimators. However, we note the following important fact, which is also discussed in [Koe05].

Define the $\tau-$*pinball loss* by $\ell(y, y') = \rho_\tau(y - y')$, where

$$\rho_\tau(u) = u(\tau - \mathbb{1}\{u < 0\}) = \begin{cases} u\tau & \text{if } u \geq 0, \\ u(\tau - 1) & \text{otherwise.} \end{cases}$$

Lemma 2.6. *Let U be a real-valued random variable with density f and strictly increasing cumulative distribution function F. Then for all $\tau \in (0,1)$, we have that*

$$F^{-1}(\tau) = \operatorname*{argmin}_{t \in \mathbb{R}} \mathbb{E}\left(\ell(Y, t)\right)$$

Proof. We have that

$$\begin{aligned} \mathbb{E}\left(\ell(Y,t)\right) = \mathbb{E}\left(\rho_\tau(Y-t)\right) &= \int_{-\infty}^{\infty} \rho_\tau(y-t) f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{t} (\tau - 1)(y-t) f(y)\,\mathrm{d}y + \int_{t}^{\infty} \tau(y-t) f(y)\,\mathrm{d}y. \end{aligned}$$

Equating the derivative of the above expression with respect to t to 0 gives

$$\int_{-\infty}^{t} (1-\tau) f(y)\,\mathrm{d}y - \int_{t}^{\infty} \tau f(y)\,\mathrm{d}y = (1-\tau)F(t) - \tau(1 - F(t)) = 0,$$

which is equivalent to

$$t = F^{-1}(\tau).$$

The second derivative is equal to $f(t)$, so $F^{-1}(\tau)$ is indeed a minimiser. □

Therefore, in the same way that training a model with respect to the least-squares loss gives an estimate for the conditional mean, training a model with respect to the $\tau-$pinball loss gives an estimate of the τ^{th} quantile of the conditional distribution $Y|X$.

Using a quantile regression procedure, we may obtain estimates $\hat{q}_{\alpha/2}(x)$ and $\hat{q}_{1-\alpha/2}(x)$ for $q_{\alpha/2}(x)$ and $q_{1-\alpha/2}(x)$, respectively. However, the interval $[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$ is not guaranteed to have $(1-\alpha)-$level coverage. *Conformalised quantile regression* [RPC19] calibrates this interval using conformal prediction to provide it with a coverage guarantee as in Theorem 2.3.

After obtaining estimates $\hat{q}_{\alpha/2}(x)$ and $\hat{q}_{1-\alpha/2}(x)$ from the proper training set, conformalised quantile regression applies split conformal prediction with the conformity score

$$\hat{s}(x, y) = \max\left\{\hat{q}_{\alpha/2}(x) - y,\, y - \hat{q}_{1-\alpha/2}(x)\right\}.$$

This results in the conformal prediction interval

$$\begin{aligned}&\left\{y \in \mathbb{R} : \hat{q}_{\alpha/2}(X_{n+1}) - y \leq \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha) \quad \text{and} \quad y - \hat{q}_{1-\alpha/2}(X_{n+1}) \leq \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)\right\} \\ &= \left[\hat{q}_{\alpha/2}(X_{n+1}) - \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha), \hat{q}_{1-\alpha/2}(X_{n+1}) + \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)\right].\end{aligned}$$

To intuitively understand this conformity score, we note the following points, inspired by the discussion in [RPC19]. If $y < \hat{q}_{\alpha/2}(x)$, i.e. y is below the predicted lower quantile, then $\hat{s}(x, y) = |y - \hat{q}_{\alpha/2}(x)|$ is the absolute error compared to the predicted lower quantile. Similarly, if $y > \hat{q}_{1-\alpha/2}(x)$, then $\hat{s}(x, y) = |y - \hat{q}_{1-\alpha/2}(x)|$. In both of these cases, $\hat{s}(x, y) > 0$, indicating the interval $[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$ has failed to cover y, and its value measures the magnitude of the error in the fitted model with respect to (x, y). If $\hat{q}_{\alpha/2}(x) < y < \hat{q}_{1-\alpha/2}(x)$, i.e. the interval $[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$ covers y, then $\hat{s}(x, y) < 0$ and

$$\hat{s}(x, y) = \min\left\{y - \hat{q}_{\alpha/2}(x),\, \hat{q}_{1-\alpha/2}(x) - y\right\},$$

which may be interpreted as the smaller of the two 'margins' by which the interval $[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$ covers y. We see from the above that if the quantile $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)$ is positive (i.e. quantile regression tends to undercover), then conformalised quantile regression widens the interval. Similarly if $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)$ is negative (quantile regression tends to overcover), then conformalised quantile regression narrows the interval.

2.4.3 Numerical Experiments

We now present a numerical experiment designed to highlight that the locally weighted score and conformalised quantile regression are more adaptive to heteroscedasticity. We consider two data generating processes. Setting 1 generates i.i.d. data points with homoscedastic noise, and setting 2 generates i.i.d data points with heteroscedastic noise.

(i) **Setting 1:**

$$\begin{aligned}X_1, X_2, \ldots &\overset{\text{i.i.d}}{\sim} \text{Uniform}(-5, 5) \\ \epsilon_1, \epsilon_2, \ldots &\overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1) \\ Y_i &= 1 - X_i + 2\epsilon_i.\end{aligned}$$

for all $i \in [n]$.

(ii) **Setting 2:**

$$X_1, X_2, \ldots \overset{\text{i.i.d}}{\sim} \text{Uniform}(-5, 5)$$
$$\epsilon_1, \epsilon_2, \ldots \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$$
$$Y_i = 1 - X_i + \frac{1}{2}\left(|X_i| + 2\right)\left(\sin(2X_i) + \frac{3}{2}\right)\epsilon_i$$

for all $i \in [n]$.

In each setting, we set $\alpha = 0.1$ and generate i.i.d. proper training, calibration and test datasets with $1000, 1500, 1500$ data points, respectively. We then generate split conformal prediction intervals (using the calibration dataset) for each data point in the test dataset and calculate both the empirical coverage (as in (2.4)) and the average length of the prediction intervals for the test dataset. We repeat this over 2000 independent draws of the calibration and test datasets. We repeat this experiment using the absolute residual score, the locally weighted residual score, conformalised quantile regression and quantile regression. The code is provided in the files `2_4_setting1.py` and `2_4_setting_2.py`.

In Table 1, we record the average empirical coverage and the average length of the prediction intervals, averaged over the 2000 draws. In Figure 5, we plot a histogram of the average length of the prediction intervals obtained in each of the 2000 draws. This estimates the distribution of

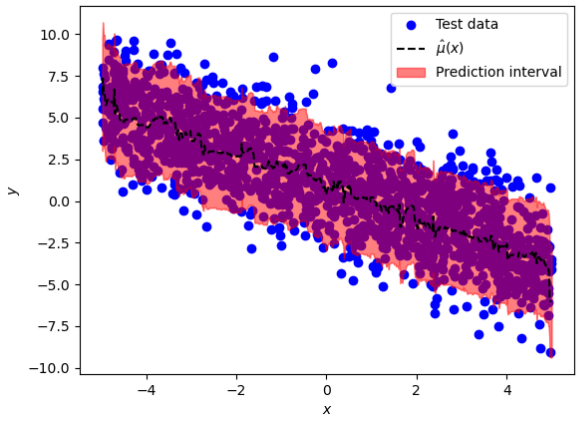
$$\mathbb{E}\left[|C(X_{n+1}; D_{\text{cal}})| \mid D_{\text{cal}}\right],$$

where, in the above display, we make the dependence of the prediction interval on the calibration data D_{cal} explicit, and $|\cdot|$ denotes the length of the prediction interval. Figure 3 and Figure 4 plot the prediction intervals obtained using the three different conformity scores, where each plot refers to a single draw of the calibration and test datasets.

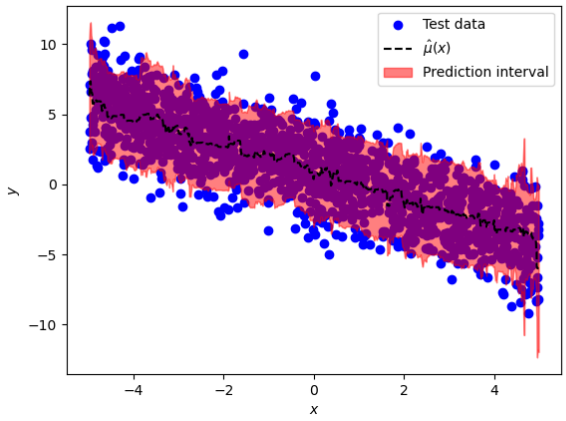
From Table 1, we observe that all three conformity scores achieve the target coverage in both settings, as guaranteed by Theorem 2.3 since all the data points are i.i.d., and thus exchangeable. However, quantile regression by itself does not give valid coverage. In setting 1, Figure 5 shows that the three conformity scores give similar distributions for the average length of the prediction intervals. However, in setting 2, we see in Figure 5 that the locally weighted score and conformalised quantile regression tend to give much narrower prediction intervals as compared to the absolute residual score. It is clear from Figure 4 that the width of the prediction intervals $C(x)$ obtained using the locally weighted residual score and conformalised quantile regression vary with x to account for the heteroscedasticity. In contrast, the intervals obtained using the absolute residual score have constant width, leading to them overcovering in some regions and undercovering in others.

	Absolute residual		Locally weighted		Conformalised quantile regression		Quantile regression	
	Average coverage	Average length	Average coverage	Average length	Average coverage	Average length	Average coverage	Average length
Setting 1	0.90022	6.6991	0.90008	6.7815	0.90028	6.8889	0.86825	6.3046
Setting 2	0.90024	13.146	0.90012	11.305	0.90015	11.595	0.87957	10.814

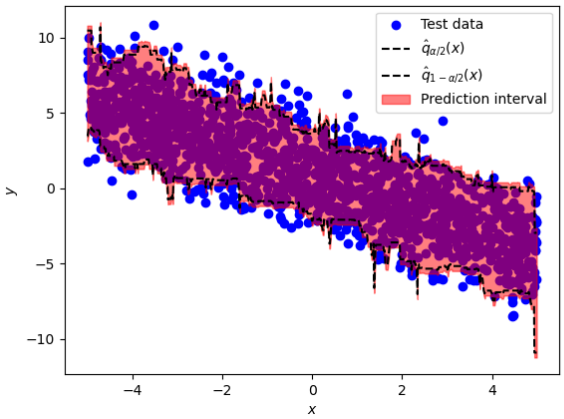
Table 1: Average coverage and average length of the conformal prediction intervals on the test dataset



(a) Absolute residual score

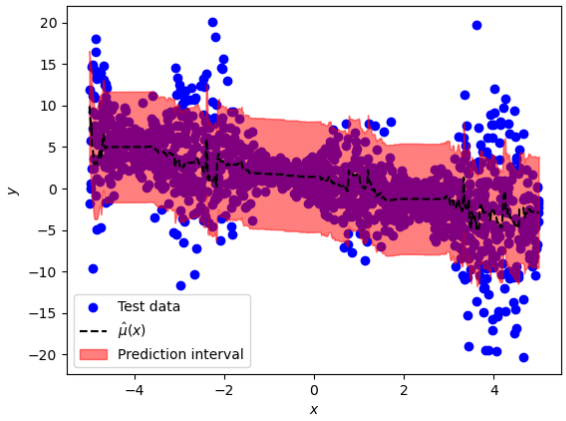


(b) Locally weighted score

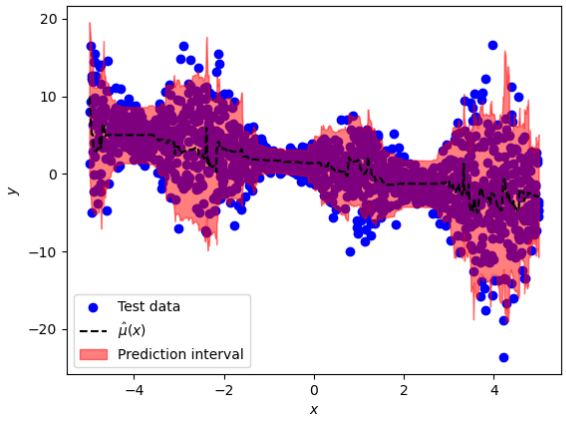


(c) Conformalised quantile regression

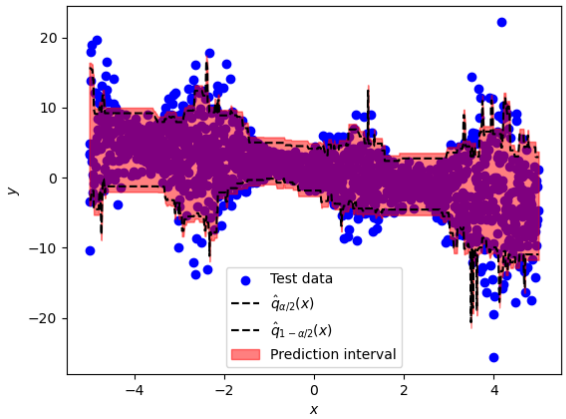
Figure 3: Conformal prediction intervals in setting 1.



(a) Absolute residual score

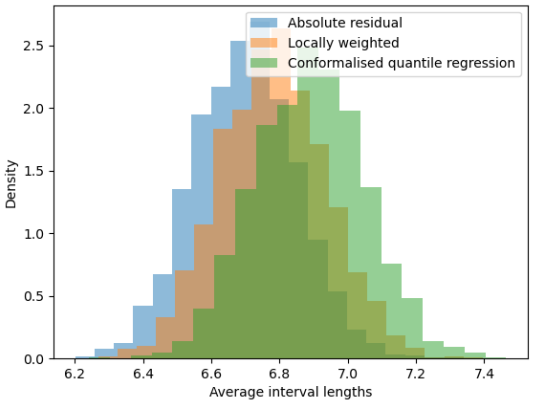


(b) Locally weighted score

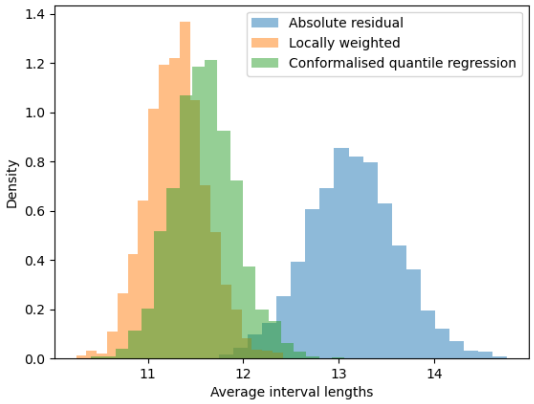


(c) Conformalised quantile regression

Figure 4: Conformal prediction intervals in setting 2.



(a) Setting 1



(b) Setting 2

Figure 5: Distribution of average prediction interval lengths from 2000 independent draws of calibration and test datasets.

3 Extensions of Conformal Prediction

In this section, we explore three theoretical extensions to the conformal prediction framework presented in Section 2. As the proof of Theorem 2.1 demonstrates, exchangeability is a fundamental assumption for the validity of conformal prediction. In many real-world applications, exchangeability may not hold. A key example of this is distribution shift, a setting where the test data has a different distribution to the training data. In Section 3.1, we consider the case where exchangeability does not hold and present a conformal prediction procedure developed by [Bar+23] that provides a coverage guarantee in this setting. The specific case of distribution shift is subsequently analysed in Section 3.2. To conclude this section, we return to the setting of exchangeability but consider an extension of conformal prediction, developed by [Ang+24], that imparts a more general class of guarantees to conformal prediction sets.

3.1 Nonexchangeable Conformal Prediction

In this subsection, we present the *nonexchangeable conformal prediction* (NexCP) method developed in [Bar+23]. Full conformal prediction relies on two main assumptions: the exchangeability of the data and the symmetry of the conformity score. The key contribution of [Bar+23], presented in Theorem 3.1, is deriving a modification of the full conformal prediction procedure that has a coverage guarantee when both of these assumptions are violated.

The key theoretical insight to extend conformal prediction beyond exchangeability, that is common to both the NexCP method and the method presented in Section 3.2, is to introduce weights. We will compare the NexCP method and the method in Section 3.2 at the end of Section 3.2. We will now use the weighted versions of the quantities defined in Definition 2.2. The NexCP method assigns fixed weights $w_1, \dots, w_n \in [0, \infty)$ to the data points $(X_1, Y_1), \dots, (X_n, Y_n)$, respectively. The corresponding normalised weights are defined by

$$\tilde{w}_i = \frac{w_i}{1 + \sum_{j=1}^{n} w_j} \quad \text{and} \quad \tilde{w}_{n+1} = \frac{1}{1 + \sum_{j=1}^{n} w_j} \tag{3.1}$$

for all $i \in [n]$.

Before we present the main theorem of this subsection, we introduce the required notation. Given data points $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$, $y \in \mathcal{Y}$ and $k \in [n+1]$, we define

$$D = ((X_i, Y_i))_{i=1}^{n+1}, \quad D^y = ((X_i, Y_i)_{i=1}^{n}, (X_{n+1}, y)), \qquad D^{(k)} = \pi_k(D), \quad \text{and} \quad D^{y,(k)} = \pi_k(D^y).$$

where $\pi_k \in S_{n+1}$ is the transposition exchanging k and $n+1$. Note that $D^{(k)}$ is simply D with (X_k, Y_k) and (X_{n+1}, Y_{n+1}) swapped.

Let s be a conformity score. As mentioned above, we will not assume that s is symmetric. Therefore, the model fitting procedure within the conformity score may take the order of the data points into account. We define

$$S = (s((X_i, Y_i); D))_{i=1}^{n+1}, \quad \text{and} \quad S^{(k)} = (s((X_{\pi_k(i)}, Y_{\pi_k(i)}); D^{(k)}))_{i=1}^{n+1}.$$

We also define

$$S_i^{y,(k)} = \begin{cases} s((X_i, Y_i); D^{y,(k)}) & \text{if } i = 1, \dots, n \\ s((X_{n+1}, y); D^{y,(k)}) & \text{if } i = n+1, \end{cases} \quad \text{and} \quad S^{y,(k)} = (S_i^{y,(k)})_{i=1}^{n+1}.$$

We define K to be random variable taking values in $[n+1]$ such that

$$\mathbb{P}(K = k) = \tilde{w}_k \tag{3.2}$$

for all $k \in [n+1]$, and we take K and D to be independent.

We also recall the definition of the total variation distance.

Definition 3.1. Let U and V be random variables. The *total variation distance* between U and V is defined as

$$\mathrm{d_{TV}}(U, V) = \sup_A |\mathbb{P}(U \in A) - \mathbb{P}(V \in A)|,$$

where the supremum is taken over all measurable sets A.

We now prove the main theorem of this subsection which provides the predicition set and coverage guarantee for the NexCP method. We follow the proof of Theorem 2 in [Bar+23].

Theorem 3.1. *Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ be a sequence of data points and s be a conformity score. Let $w_1, \ldots, w_{n+1} \in [0, \infty)$ be fixed real numbers and define $\tilde{w}_i$ according to (3.1) for all $i \in [n+1]$. Let K be a random variable as in (3.2) that is independent of D. Define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ S_{n+1}^{y,(K)} \leq \hat{Q}^{\tilde{w}}_{S^{y,(K)}}(1-\alpha) \right\}. \tag{3.3}$$

Then we have that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha - \sum_{k=1}^{n} \tilde{w}_k \mathrm{d_{TV}}(S, S^{(k)}).$$

Proof. If $Y_{n+1} \notin C(X_{n+1})$, then

$$S_{n+1}^{Y_{n+1},(K)} > \hat{Q}^{\tilde{w}}_{S_{n+1}^{Y_{n+1},(K)}}(1-\alpha).$$

This implies that

$$S_{n+1}^{Y_{n+1},(K)} > \hat{Q}^{\tilde{w}}_{\left(S_1^{Y_{n+1},(K)}, \ldots, S_n^{Y_{n+1},(K)}, \infty\right)}(1-\alpha)$$

by Lemma 2.3. We now claim that

$$\hat{Q}^{\tilde{w}}_{\left(S_1^{Y_{n+1},(K)}, \ldots, S_n^{Y_{n+1},(K)}, \infty\right)}(1-\alpha) \geq \hat{Q}^{\tilde{w}}_{S^{(K)}}(1-\alpha).$$

If $K = n+1$, then this follows by (2.2) in the proof of Lemma 2.3. If $K \leq n$, we have that for any $x \in \mathbb{R}$,

$$\begin{aligned}
\hat{F}^{\tilde{w}}_{S^{(K)}}(x) &= \sum_{i=1}^{n+1} \tilde{w}_i \mathbb{1}\left\{ x \geq S_{\pi_K(i)}^{Y_{n+1},(K)} \right\} \\
&= \sum_{\substack{i=1 \\ i \neq K}}^{n} \tilde{w}_i \mathbb{1}\left\{ x \geq S_i^{Y_{n+1},(K)} \right\} + \tilde{w}_K \mathbb{1}\left\{ x \geq S_{n+1}^{Y_{n+1},(K)} \right\} + \tilde{w}_{n+1} \mathbb{1}\left\{ x \geq S_K^{Y_{n+1},(K)} \right\} \\
&= \hat{F}^{\tilde{w}}_{\left(S_1^{Y_{n+1},(K)}, \ldots, S_n^{Y_{n+1},(K)}, \infty\right)}(x) + \tilde{w}_K \mathbb{1}\left\{ x \geq S_{n+1}^{Y_{n+1},(K)} \right\} + \tilde{w}_{n+1} \mathbb{1}\left\{ x \geq S_K^{Y_{n+1},(K)} \right\} \\
&\qquad - \tilde{w}_K \mathbb{1}\left\{ x \geq S_K^{Y_{n+1},(K)} \right\} - \tilde{w}_{n+1} \mathbb{1}\{ x \geq \infty \} \\
&= \hat{F}^{\tilde{w}}_{\left(S_1^{Y_{n+1},(K)}, \ldots, S_n^{Y_{n+1},(K)}, \infty\right)}(x) + \tilde{w}_K \left(\mathbb{1}\left\{ x \geq S_{n+1}^{Y_{n+1},(K)} \right\} - \mathbb{1}\{ x \geq \infty \} \right) \\
&\qquad + (\tilde{w}_{n+1} - \tilde{w}_K) \left(\mathbb{1}\left\{ x \geq S_K^{Y_{n+1},(K)} \right\} - \mathbb{1}\{ x \geq \infty \} \right) \\
&\geq \hat{F}^{\tilde{w}}_{\left(S_1^{Y_{n+1},(K)}, \ldots, S_n^{Y_{n+1},(K)}, \infty\right)}(x).
\end{aligned}$$

Therefore, we have that

$$\hat{F}^{\tilde{w}}_{S^{(K)}}\left(\hat{Q}^{\tilde{w}}_{\left(S_1^{Y_{n+1},(K)},\ldots,S_n^{Y_{n+1},(K)},\infty\right)}(1-\alpha)\right) \geq 1-\alpha,$$

by Lemma 2.1 which shows the claim above.

So far, we have shown that

$$Y_{n+1} \notin C(X_{n+1}) \implies S_{n+1}^{Y_{n+1},(K)} > \hat{Q}^{\tilde{w}}_{S^{(K)}}(1-\alpha).$$

Noting that $S_{n+1}^{Y_{n+1},(K)} = S_K^{(K)}$, this implies that

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq \mathbb{P}\left(S_K^{(K)} \leq \hat{Q}^{\tilde{w}}_{S^{(K)}}(1-\alpha)\right).$$

Thus, we have that

$$\begin{aligned}
\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) &\geq \mathbb{P}\left(S_K^{(K)} \leq \hat{Q}^{\tilde{w}}_{S^{(K)}}(1-\alpha)\right) \\
&= \sum_{k=1}^{n+1} \mathbb{P}\left(S_k^{(k)} \leq \hat{Q}^{\tilde{w}}_{S^{(k)}}(1-\alpha),\, K=k\right) \\
&= \sum_{k=1}^{n+1} \tilde{w}_k \mathbb{P}\left(S_k^{(k)} \leq \hat{Q}^{\tilde{w}}_{S^{(k)}}(1-\alpha)\right) \\
&= \sum_{k=1}^{n+1} \tilde{w}_k \mathbb{P}\left(S_k \leq \hat{Q}^{\tilde{w}}_{S}(1-\alpha)\right) \\
&\qquad + \sum_{k=1}^{n+1} \tilde{w}_k \left[\mathbb{P}\left(S_k^{(k)} \leq \hat{Q}^{\tilde{w}}_{S^{(k)}}(1-\alpha)\right) - \mathbb{P}\left(S_k \leq \hat{Q}^{\tilde{w}}_{S}(1-\alpha)\right)\right] \\
&\geq \mathbb{E}\left[\sum_{k=1}^{n+1} \tilde{w}_k \mathbb{1}\left\{S_k \leq \hat{Q}^{\tilde{w}}_{S}(1-\alpha)\right\}\right] - \sum_{k=1}^{n+1} \tilde{w}_k \, \mathrm{d_{TV}}(S, S^{(k)}) \\
&= \mathbb{E}\left[\hat{F}^{\tilde{w}}_{S}(\hat{Q}^{\tilde{w}}_{S}(1-\alpha))\right] - \sum_{k=1}^{n+1} \tilde{w}_k \, \mathrm{d_{TV}}(S, S^{(k)}) \\
&\geq 1-\alpha - \sum_{k=1}^{n+1} \tilde{w}_k \, \mathrm{d_{TV}}(S, S^{(k)}),
\end{aligned}$$

where the third line follows from the independence of K and D, the fifth line follows from Definition 3.1 and the final inequality follows from Lemma 2.1. □

The corresponding algorithm, referred to as *nonexchangeable full conformal predicition* is stated below.

Algorithm 3 Nonexchangeable full conformal prediction algorithm

Input: Calibration $((X_i, Y_i))_{i=1}^n$; test predictor X_{n+1}; miscoverage level $\alpha \in (0,1)$, conformity score $\hat{s}$.
  Initialise $C \leftarrow \emptyset$
  Draw K from $[n+1]$ according to (3.2).
  for $y \in \mathcal{Y}$ **do**
    Compute $S_i^{y,(K)}$ for $i \in [n+1]$
    Compute $\hat{Q}^{\tilde{w}}_{S^{y,(K)}}(1-\alpha)$
    if $S_{n+1}^{y,(K)} \leq \hat{Q}^{\tilde{w}}_{S^{y,(K)}}$ **then**
      $C \leftarrow C \cup \{y\}$
    end if
  end for
Output: C

In the same way that split conformal prediction is shown to be a special case of full conformal prediction in Section 2, we may derive a corresponding result for NexCP, referred to as *nonexchangeable split conformal predicition*. We use the notation from Section 2.3, denoting $\hat{s} : \mathcal{X} \to \mathcal{Y}$ to be the conformity score and D_{tr} the proper training set. We also write $\hat{S}_i = \hat{s}(X_i, Y_i)$ for all $i \in [n]$ and $\hat{S}_{n+1}^y = \hat{s}(X_{n+1}, y)$.

Corollary 3.1. *Suppose $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ are data points. Define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ \hat{S}_{n+1}^y \leq \hat{Q}^{\tilde{w}}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1-\alpha) \right\}.$$

Then we have that

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha - \sum_{k=1}^{n} \tilde{w}_k \mathrm{d}_{\text{TV}}(S, S^{(k)}).$$

Proof. This follows from Theorem 3.1 and Lemma 2.3 in exactly the same way as in the proof of Theorem 2.3 by taking $s((x,y); \tilde{D}) = \hat{s}(x,y)$ for $(x,y) \in \mathcal{Z}$ and $\tilde{D} = \cup_{j \geq 1} \mathcal{Z}^j$. □

We now discuss the implications of Theorem 3.1. As mentioned at the beginning of this subsection, the coverage guarantee in Theorem 3.1 makes no assumption on the distribution of the data or on the conformity score. The quantity

$$\sum_{k=1}^{n} \tilde{w}_k \mathrm{d}_{\text{TV}}(S, S^{(k)}).$$

may be interpreted as the loss in coverage that occurs due to not assuming exchangeability or that the conformity score is symmetric. As in [Bar+23], we can further understand the result by highlighting some important special cases. We record these are corollaries below.

Corollary 3.2. *Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ and s be a symmetric conformity score. Let $C(X_{n+1})$ be the prediction set (2.1). Then we have that*

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha - \frac{1}{n+1} \sum_{k=1}^{n} \mathrm{d}_{\text{TV}}(S, S^{(k)}).$$

Proof. Since s is symmetric, we have that $S_i^{y,(K)} = s((X_i, Y_i); D^{y,(K)}) = s((X_i, Y_i); D^y) = S_i^y$ for $i \in [n]$ and similarly, $S_{n+1}^{y,(K)} = S_{n+1}^y$. Moreover, if we take $w_i = 1$ for each $i \in [n]$, then the prediction set (3.3) coincides with (2.1), so the result follows from Theorem 3.1. □

Corollary 3.3. *Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ be exchangeable and s be a conformity score. With $\tilde{w}$ and K as in Theorem 3.1, define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \; S_{n+1}^{y,(K)} \leq \hat{Q}^{\tilde{w}}_{S^{y,(K)}}(1-\alpha) \right\}.$$

Then we have that

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

Proof. If the data is exchangeable, then $S \stackrel{d}{=} S^{(k)}$ for any $k \in [n]$, so

$$\sum_{k=1}^{n} \tilde{w}_k \mathrm{d_{TV}}(S, S^{(k)}) = 0,$$

and the result follows from Theorem 3.1. □

An interpretation of Corollary 3.2 is that if we apply the standard full conformal prediction algorithm Algorithm 1 to nonexchangeable data, then

$$\frac{1}{n+1} \sum_{k=1}^{n} \mathrm{d_{TV}}(S, S^{(k)})$$

is the loss in marginal coverage compared to the $1-\alpha$ level in the exchangeable case. If we interpret the quantity in the above display as quantifying the extent to which exchangeability is violated, then Corollary 3.2 shows that if the violation of exchangeability is small, then the coverage loss from applying standard conformal prediction is small too.

As mentioned in [Bar+23], Corollary 3.3 demonstrates that using the NexCP method does not lead to a loss of coverage in the case that the data is exchangeable. In fact, taking $w_i = 1$ for each $i \in [n]$, Corollary 3.3 shows precisely how the full conformal prediction set must be modified in order that the coverage guarantee holds for conformity scores that are not necessarily symmetric. We note that extending full conformal prediction to include non-symmetric conformity scores is useful theoretically since it enables fitting algorithms that do not treat the data symmetrically (e.g. weighted regression) in full conformal prediction. However, since full conformal prediction is rarely used in practice, and split conformal prediction is valid with any conformity score function, the extension to nonsymmetric scores may have limited practical applicability.

Overall, the primary strengths of the NexCP method is that it quantifies the loss of coverage when exchangeability and the symmetry of the score function fail to hold and that it provides extensions of the results in Section 2 in the exchangeable case. The main limitation of this method is that it provides no way of choosing the weights. From Theorem 3.1, we see that the result is only meaningful if the weights can be suitably chosen to ensure that the loss of coverage is small. Whilst the authors of [Bar+23] leave the problem of choosing the weights open, they show ([Bar+23] Lemma 1) that if $(X_1, Y_1) \ldots, (X_{n+1}, Y_{n+1})$ are independent, then

$$\sum_{k=1}^{n} \tilde{w}_k \mathrm{d_{TV}}(S, S^{(k)}) \leq 2 \sum_{k=1}^{n} \tilde{w}_k \mathrm{d_{TV}}((X_i, Y_i), (X_{n+1}, Y_{n+1})).$$

This supports the intuition that assigning a higher weight to points whose distribution is similar to the test point yields a smaller loss in coverage.

3.2 Distribution Shift

Having considered a general violation of exchangeability in Section 3.1, in this subsection, we focus on distribution shift. Specifically, denoting $Z_i = (X_i, Y_i) \in \mathcal{Z}$ for $i \in [n+1]$, we will consider the case where the calibration data satisfies $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ for a distribution P and the test point $Z_{n+1} \sim Q$ for some other distribution Q. Note that $Z_1, \ldots, Z_{n+1}$ are not exchangeable since they are not identically distributed. The main theorem of this subsection, Theorem 3.2, demonstrates that a modified conformal prediction procedure achieves $(1-\alpha)-$level coverage in this setting. We will refer to this procedure as *weighted conformal prediction*.

The main ideas for the procedure we discuss were developed in [Tib+19; PR21]. Our presentation - inspired by [ABB24] - will incorporate some ideas from the more recent works [Bar+24; Tan23], which develops theory to frame the concepts of [Tib+19; PR21] more generally.

We first introduce a notion that generalises exchangeability, referred to as *weighted exchangeability* [Tib+19; Bar+24; Tan23]. In the following, $\mathcal{U}$ denotes a separable complete metric space, $\mathcal{B}(\mathcal{U})$ denotes the Borel $\sigma-$algebra on $\mathcal{U}$ and Λ denotes the set of measurable functions from $\mathcal{U}$ to $(0, \infty)$. The conditions on $\mathcal{U}$ are required due to measure-theoretic results ensuring the existence of regular conditional distributions, which we do not discuss further here (see Appendix A.1 [Bar+24]).

Definition 3.2 ([Bar+24; Tan23])**.** (i) A probability measure Q on $\mathcal{U}^n$ is *exchangeable* if for all $A_1, \ldots, A_n \in \mathcal{B}(\mathcal{U})$,

$$Q(A_1 \times \cdots \times A_n) = Q(A_{\sigma(1)} \times \cdots \times A_{\sigma(n)})$$

for all $\sigma \in S_n$.

(ii) Given $\lambda = (\lambda_1, \ldots, \lambda_n) \in \Lambda^n$, a probability measure Q on $\mathcal{U}^n$ is called $\lambda-$*weighted exchangeable* if the measure $\bar{Q}$ defined as

$$\bar{Q}(B) = \int_B \frac{\mathrm{d}Q(x_1, \ldots, x_n)}{\lambda_1(x_1) \cdots \lambda_n(x_n)}, \quad \text{for } B \in \mathcal{B}(\mathcal{U}^n),$$

is exchangeable.

This relates to our standard notion of exchangeability by noting that $U_1, \ldots, U_n$ are exchangeable according to Definition 2.1 if and only if $\mathbb{P} \circ U^{-1}$ is exchangeable according to Definition 3.2, where $U = (U_1, \ldots, U_n)$. Moreover, note that Q is exchangeable if and only if Q is $\lambda-$weighted exchangeable for $\lambda_1, \ldots, \lambda_n \equiv 1$. We also note that the notion of weighted exchangeability above generalises the definition in [Tib+19], which we record below.

Definition 3.3 ([Tib+19])**.** Suppose $U_1, \ldots, U_n$ are continuous real-valued random variables with joint density f. They are said to be $\lambda-$*weighted-exchangeable* if

$$f(u_1, \ldots, u_n) = g(u_1, \ldots, u_n) \prod_{i=1}^{n} \lambda(u_i),$$

for some g satisfying $g(u_1, \ldots, u_n) = g(u_{\sigma(1)}, \ldots, u_{\sigma(n)})$ for all $u_1, \ldots, u_n \in \mathbb{R}$ and $\sigma \in S_n$.

Specifically, $U_1, \ldots, U_n$ are $\lambda-$weighted-exchangeable according to Definition 3.3 if and only if $\mathbb{P} \circ U^{-1}$ is $\lambda-$weighted-exchangeable according to Definition 3.2, where $U = (U_1, \ldots, U_n)$. Indeed, defining $g(u_1, \ldots, u_n) = \frac{f(u_1, \ldots, u_n)}{\lambda_1(u_1) \cdots \lambda_n(u_n)}$, we note that for any Borel-measurable $B_1, \ldots, B_n \subseteq \mathbb{R}$ and $\sigma \in S_n$, we have

$$\int_{B_1 \times \cdots \times B_n} \frac{\mathrm{d}(\mathbb{P} \circ U^{-1})(u_1, \ldots, u_n)}{\lambda_1(u_1) \cdots \lambda_n(u_n)} = \int_{B_1 \times \cdots \times B_n} g(u_1, \ldots, u_n)\, \mathrm{d}u_1 \cdots \mathrm{d}u_n,$$

and

$$\begin{aligned}
&\int_{B_{\sigma(1)}\times\cdots\times B_{\sigma(n)}} \frac{\mathrm{d}(\mathbb{P}\circ U^{-1})(u_1,\ldots,u_n)}{\lambda_1(u_1)\cdots\lambda_n(u_n)} \\
&= \int_{\mathbb{R}^n} g(u_1,\ldots,u_n)\,\mathbb{1}\left\{(u_{\sigma^{-1}(1)},\ldots,u_{\sigma^{-1}(n)})\in B_1\times\cdots\times B_n\right\}\mathrm{d}u_{\sigma^{-1}(1)}\cdots\mathrm{d}u_{\sigma^{-1}(n)} \\
&= \int_{B_1\times\cdots\times B_n} g(u_{\sigma(1)},\ldots,u_{\sigma(n)})\,\mathrm{d}u_1\cdots\mathrm{d}u_n,
\end{aligned}$$

from which the equivalence follows.

The lemma below demonstrates how the notion of weighted exchangeability applies to the distribution shift setting.

Lemma 3.1. *Let $Z_1,\ldots,Z_n \overset{\text{i.i.d}}{\sim} P$ be data points in $\mathcal{Z}$ and suppose $Z_{n+1}\sim Q$ is a data point in $\mathcal{Z}$ independent of $(Z_i)_{i=1}^n$, for some distributions P,Q on $\mathcal{Z}$ such that Q is absolutely continuous with respect to P. Then the distribution of $((X_i,Y_i))_{i=1}^{n+1}$ is $\lambda-$weighted exchangeable, where $\lambda=\left(1,\ldots,1,\frac{\mathrm{d}Q}{\mathrm{d}P}\right)$ and $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is the Radon-Nikodym derivative.*

Proof. Note that $(Z_i)_{i=1}^{n+1}\sim P^n\times Q$. For any measurable sets $B_1,\ldots,B_{n+1}$ and $\sigma\in S_{n+1}$, we have that

$$\begin{aligned}
(\overline{P^n\times Q})(B_1\times\cdots\times B_{n+1}) &= \int_{B_1\times\cdots\times B_{n+1}} \frac{\mathrm{d}(P^n\times Q)(u_1,\ldots,u_{n+1})}{\frac{\mathrm{d}Q}{\mathrm{d}P}(u_{n+1})} \\
&= P(B_1)\cdots P(B_n)\int_{B_{n+1}} \frac{1}{\frac{\mathrm{d}Q}{\mathrm{d}P}(u_{n+1})}\,\mathrm{d}Q(u_{n+1}) \\
&= P(B_1)\cdots P(B_n)\int_{B_{n+1}} \frac{1}{\frac{\mathrm{d}Q}{\mathrm{d}P}(u_{n+1})}\frac{\mathrm{d}Q}{\mathrm{d}P}(u_{n+1})\,\mathrm{d}P(u_{n+1}) \\
&= P^{n+1}(B_1\times\cdots\times B_{n+1}).
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
(\overline{P^n\times Q})(B_1\times\cdots\times B_{n+1}) &= P^{n+1}(B_1\times\cdots\times B_{n+1}) \\
&= P^{n+1}(B_{\sigma(1)}\times\cdots\times B_{\sigma(n+1)}) \\
&= (\overline{P^n\times Q})(B_{\sigma(1)}\times\cdots\times B_{\sigma(n+1)}),
\end{aligned}$$

so $\overline{P^n\times Q}$ is exchangeable. □

We now state an important lemma on weighted exchangeability that will be used to establish the coverage guarantee. The interpretation of this result, as mentioned in [Bar+24], is that conditional on the unordered multiset set of values $\{U_1,\ldots,U_k\}$, the distribution of U_i is a discrete distribution over $\{U_1,\ldots,U_k\}$ where the probabilities may be explicity expressed in terms of λ. The conditional probability of taking on the value U_j is denoted $w_j^{i,k}(U_1,\ldots,U_k)$ below.

We also define the *empirical distribution* of $U_1,\ldots,U_k$ by

$$\hat{P}_k \coloneqq \frac{1}{k}\sum_{i=1}^k \delta_{U_i},$$

where δ_u denotes the Dirac measure at u, for all $u\in\mathcal{U}$.

Lemma 3.2 ([Bar+24] Proposition 7)**.** *For any $\lambda \in \Lambda^k$, any $\lambda-$weighted exchangeable Q on $\mathcal{U}^k$, and $U \sim Q$, we have that*

$$U_i \mid \hat{P}_k \sim \sum_{j=1}^{k} w_j^{i,k}(U_1, \ldots, U_k)\, \delta_{U_j},$$

where

$$w_j^{i,k}(u_1, \ldots, u_k) = \frac{\sum_{\sigma \in S_k:\, \sigma(i)=j} \lambda_1(u_{\sigma(1)}) \cdots \lambda_k(u_{\sigma(k)})}{\sum_{\sigma \in S_k} \lambda_1(u_{\sigma(1)}) \cdots \lambda_k(u_{\sigma(k)})}$$

The authors of [Bar+24] formalise conditioning on the unordered values $\{U_1, \ldots, U_k\}$ by conditioning on the sub-$\sigma-$algebra $\mathcal{E}_k$ of $\mathcal{B}(\mathcal{U}^k)$ defined by $(u_1, \ldots, u_k) \in \mathcal{E}_k \iff (u_\sigma(k), \ldots, u_\sigma(k)) \in \mathcal{E}_k$ for any $\sigma \in S_k$. As mentioned in [Bar+24], this can be shown to be equivalent to conditioning on the empirical distribution, which is how we stated Lemma 3.2. We omit the proof of Lemma 3.2 since it consists primarily of measure-theoretic calculations regarding conditional distributions; the proof may be found in [Bar+24] (Proposition 7). However, it is helpful to note a special case of this result.

Remark 3.1. If $U_1, \ldots, U_k$ are exchangeable, then $w_j^{i,k}(u_1, \ldots, u_k) = 1/k$, so the result states that

$$U_i \mid \hat{P}_k \sim \frac{1}{k} \sum_{j=1}^{k} \delta_{U_j} = \hat{P}_k.$$

The interpretation of this is that if $U_1, \ldots, U_k$ are exchangeable, then conditional on the unorderd multiset of values $\{U_1, \ldots, U_k\}$, U_i is equally likely to be any of the values in the multiset $\{U_1, \ldots, U_k\}$.

We now apply Lemma 3.2 to our setting by combining it with Lemma 3.1.

Lemma 3.3 ([ABB24] Proposition 7.6)**.** *Let $Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} P$ be data points in $\mathcal{Z}$ and suppose $Z_{n+1} \sim Q$ is a data point in $\mathcal{Z}$ independent of $(Z_i)_{i=1}^n$, for some distributions P, Q on $\mathcal{Z}$ such that Q is absolutely continuous with respect to P. Then*

$$Z_{n+1} \mid \hat{P}_{n+1} \sim \sum_{i=1}^{n+1} w_i\, \delta_{Z_i},$$

where

$$w_i = \frac{\frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)}{\sum_{j=1}^{n+1} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_j)}, \quad i \in [n+1]. \tag{3.4}$$

Proof. By Lemma 3.2 we have that

$$Z_{n+1} | \hat{P}_{n+1} \sim \sum_{i=1}^{n+1} w_i^{n+1,\,n+1}(Z_1, \ldots, Z_{n+1})\, \delta_{Z_i},$$

where

$$\begin{aligned} w_i^{n+1,\,n+1}(Z_1, \ldots, Z_{n+1}) &= \frac{\sum_{\sigma \in S_{n+1}:\ \sigma(n+1)=i} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_{\sigma(n+1)})}{\sum_{\sigma \in S_{n+1}} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_{\sigma(n+1)})} \\ &= \frac{n!\, \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)}{\sum_{j=1}^{n+1} \sum_{\sigma \in S_{n+1}:\ \sigma(n+1)=j} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_j)} \\ &= \frac{\frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)}{\sum_{j=1}^{n} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_j)}. \end{aligned}$$

□

We now state the main theorem of this subsection which derives the prediction set and coverage guarantee for weighted conformal prediction. The statement and proof follow that of Theorem 7.5 in [ABB24].

Theorem 3.2. *[[ABB24] Theorem 7.5] Let $Z_1, \ldots, Z_{n+1} \overset{\text{i.i.d}}{\sim} P$ be data points in $\mathcal{Z}$ and suppose $Z_{n+1} \sim Q$ is a data point in $\mathcal{Z}$ independent of $(Z_i)_{i=1}^n$, for some distributions P, Q on $\mathcal{Z}$ such that Q is absolutely continuous with respect to P. For $y \in \mathcal{Y}$ and $i \in [n]$, define*

$$w_i^y = \frac{\frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)}{\sum_{j=1}^n \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_j) + \frac{\mathrm{d}Q}{\mathrm{d}P}(X_{n+1}, y)} \quad \text{and} \quad w_{n+1}^y = \frac{\frac{\mathrm{d}Q}{\mathrm{d}P}(X_{n+1}, y)}{\sum_{j=1}^n \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_j) + \frac{\mathrm{d}Q}{\mathrm{d}P}(X_{n+1}, y)}. \tag{3.5}$$

Define the prediction set

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \hat{Q}_{S^y}^{w^y}(1-\alpha) \right\},$$

where $w^y = (w_i^y)_{i=1}^{n+1}$. Then we have that

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

Proof. We have that

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) = \mathbb{P}\left(S_{n+1} \leq Q_S^{w^{Y_{n+1}}}(1-\alpha)\right) = \mathbb{E}\left[\mathbb{P}\left(S_{n+1} \leq Q_S^w(1-\alpha) \;\middle|\; \hat{P}_{n+1}\right)\right],$$

where w_i is as in Equation (3.4) and $w = (w_i)_{i=1}^{n+1}$. We first note that

$$\hat{F}_S^w(x) = \frac{\sum_{i=1}^{n+1} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i) \mathbb{1}\left\{x \geq s(Z_i; D)\right\}}{\sum_{i=1}^{n+1} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)},$$

for any $x \in \mathbb{R}$. Thus, we have that $\hat{F}_S^w$ is invariant under permutations of $(Z_1, \ldots, Z_{n+1})$ since s is symmetric. Therefore, $\hat{Q}_S^w(1-\alpha)$ is $\hat{P}_{n+1}-$measurable. Moreover, Lemma 3.3 and the symmetry of s imply that

$$S_{n+1} \mid \hat{P}_{n+1} \sim \sum_{i=1}^{n+1} w_i \delta_{S_i}.$$

Therefore, we have that

$$\mathbb{P}\left(S_{n+1} \leq Q_S^w(1-\alpha) \;\middle|\; \hat{P}_{n+1}\right) = \hat{F}_S^w(\hat{Q}_S^w(1-\alpha)) \geq 1 - \alpha$$

by Lemma 2.1. The result follows upon taking expectations. □

Theorem 3.2 gives rise to the following algorithm.

Algorithm 4 Weighted conformal prediction algorithm

Input: Data $((X_i, Y_i))_{i=1}^n$; test predictor X_{n+1}; miscoverage level $\alpha \in (0,1)$, conformity score s.
  Initialise $C \leftarrow \emptyset$
  for $y \in \mathcal{Y}$ **do**
    Compute w_i^y for $i \in [n+1]$ according to (3.5).
    Compute $S_i^y = s((X_i, Y_i); D^y)$ for $i \in [n+1]$.
    Compute $\hat{Q}_{S^y}^w(1-\alpha)$
    if $S_{n+1}^y \leq \hat{Q}_{S^y}^w(1-\alpha)$ **then**
      $C \leftarrow C \cup \{y\}$.
    end if
  end for
Output: C

An important limitation of Theorem 3.2 and Algorithm 4 is that it assumes $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is known (up to a constant factor). In practice, this is unlikely to hold and so $\frac{\mathrm{d}Q}{\mathrm{d}P}$ must be estimated. In the case where Q and P have densities q and p, respectively, it is equivalent to estimate the density ratio q/p, for which numerous methods exist [SSK12]. As discussed in [Tib+19], estimating the density ratio can be reframed as a binary classification problem by attaching a label $C = 1$ for data points from Q and $C = 0$ for data points from P. Then

$$\frac{\mathbb{P}(C = 1 \mid Z = z)}{\mathbb{P}(C = 0 \mid Z = z)} = \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)} \frac{q(z)}{p(z)}.$$

Therefore, if $\hat{\rho}(z)$ is an estimate of $\mathbb{P}(C = 1 \mid Z = z)$ obtained from the data (e.g. using logistic regression), then

$$\frac{\hat{\rho}(z)}{1 - \hat{\rho}(z)}$$

can be used as an estimate of $\frac{\mathrm{d}Q}{\mathrm{d}P}$ in weighted conformal prediction. Another limitation of weighted conformal prediction is that we cannot, in general, derive a split conformal prediction variant of it. In the notation of Section 2.3, if we take a conformity score $\hat{s}$, then the quantile $\hat{Q}^{w^y}_{S^y} = \hat{Q}^{w^y}_{(\hat{S}_1,\ldots,\hat{S}_{n+1})}$ still depends on y, in general. In the special case where $\frac{\mathrm{d}Q}{\mathrm{d}P} \equiv \frac{\mathrm{d}Q}{\mathrm{d}P}(x, y)$ is only a function of x, referred to as *covariate shift*, the quantile no longer depends on y, and so we can derive a split conformal prediction variant of weighted conformal prediction in the same way as in Theorem 2.3.

Although both the NexCP method and weighted conformal prediction are weighted variants of conformal prediction, it is important note there are several differences in the methods, which are also discussed in [Bar+23]. The weights in weighted conformal prediction are data-dependent and are specialised to the distribution shift setting. Therefore, the weighted conformal predicition method has no coverage guarantee outside the distribution shift setting in Theorem 3.2. On the other hand, the NexCP method uses fixed weights and provides a coverage guarantee for a general violation of exchangeability and also allows for s to not be symmetric. The NexCP method also provides valid coverage if there exchangeability does in fact hold, whereas weighted conformal prediction has no guarantee in this situation.

3.3 Conformal Risk Control

In this subsection, we present the *conformal risk control* method developed by [Ang+24]. The presentation of this subsection is inspired by [Ang+24; ABB24] and the proof of Theorem 3.3 follows [Ang+24].

Whilst in Section 3.1 and Section 3.2, we considered extensions of conformal prediction under violation of exchangeability, we now initially return to the setting of exchangeable data. Conformal risk control extends the scope of conformal prediction beyond coverage guarantees. Specifically, recall that for split conformal prediction, we may rewrite the coverage guarantee as follows.

$$\mathbb{P}(Y_{n+1} \notin C(X_{n+1})) = \mathbb{E}[\mathbb{1}\{Y_{n+1} \notin C(X_{n+1})\}] \leq \alpha, \tag{3.6}$$

where

$$C = \left\{ y \in \mathcal{Y} : \hat{S}^y_{n+1} \leq \hat{Q}_{\hat{S}} \left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n} \right) \right\},$$

and $\hat{S} = (\hat{S}_1, \ldots, \hat{S}_n)$. Conformal risk control provides a *risk control* guarantee of the form

$$\mathbb{E}\left[\ell(Y_{n+1}, C_{\hat{\lambda}}(X_{n+1}))\right] \leq \alpha, \tag{3.7}$$

where ℓ is a *loss function* and $\hat{\lambda}$ is a parameter dependent on $((X_i, Y_i))_{i=1}^n$. Indeed, we will show later on that the coverage guarantee of split conformal prediction is a special case of (3.7) obtained by choosing

$$C_\lambda(x) = \{y \in \mathcal{Y} : \hat{s}(x,y) \leq \lambda\}, \ \ell(y, C_\lambda(x)) = \mathbb{1}\{y \notin C_\lambda(x)\} \text{ and } \hat{\lambda} = \hat{Q}_{\hat{S}}\left(\frac{\lceil (1-\alpha)(n+1)\rceil}{n}\right), \quad (3.8)$$

as suggested by comparing the form of (3.6) and (3.7).

As mentioned in [Ang+24], an example of a problem setting where this is particularly useful is multiclass classification, or, more generally, when predicting a set-valued output. If there are K classes in total, standard conformal prediction would output a collection of subsets of $[K]$, containing the true set of classes with probability at least $1-\alpha$. However, it may be more desirable to output a single set that contains at least a fraction $1-\alpha$ of the true classes. This can be achieved by applying conformal risk control with the *false negative rate* loss function given by

$$\ell(y, C_\lambda(x)) = 1 - \frac{|y \cap C_\lambda(x)|}{|y|}.$$

We now discuss the assumptions behind this method and prove the risk control guarantee in Theorem 3.3. Throughout, we will draw analogies to split conformal prediction to aid comprehension.

The parameter λ controls the conservativeness of the prediction sets with a larger value of λ indicating a larger, and hence more conservative, prediction set. Theorem 3.3 explains how λ should be chosen to provide the risk control guarantee. The corresponding parameter in split conformal prediction is the quantile threshold $\hat{Q}_{\hat{S}}\left(\frac{\lceil (1-\alpha)(n+1)\rceil}{n}\right)$. If this threshold equals ∞, then the resulting prediction set is $\mathcal{Y}$ - the most conservative, but also uninformative, prediction set possible. This motivates the following assumptions.

We assume that λ takes values in $\Lambda \subseteq \mathbb{R}$ such that $\lambda_{\max} = \sup \Lambda \in \Lambda$. We assume that if $\lambda' \geq \lambda$, then $C_\lambda(x) \subseteq C_{\lambda'}(x)$ for all $x \in \mathcal{X}$, i.e. as we increase λ, the prediction sets become larger. Moreover, we assume that the function $\lambda \mapsto \ell(y, C_\lambda(x))$ is decreasing and right-continuous for all $(x,y) \in \mathcal{Z}$. This means that loss of larger prediction sets is smaller. We also define the *empirical risk*

$$\hat{R}_k(\lambda) = \frac{1}{k}\sum_{i=1}^{k} \ell(Y_i, C_\lambda(X_i)),$$

for $k \in [n+1]$. Note that R_k is decreasing and right-continous. We now prove the risk control guarantee.

Theorem 3.3. *Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ are exchangeable and that the prediction sets $C_\lambda(x)$ and the map $\lambda \mapsto \ell(y, C_\lambda(x))$ satisfy the assumptions above. Additionally, assume that*

$$\ell(y, C_{\lambda_{\max}}(x)) \leq \alpha \quad \text{and} \quad \sup_{\lambda \in \Lambda} \ell(y, C_\lambda(x)) \leq B < \infty,$$

for some $B > \alpha$ and for all $(x,y) \in \mathcal{Z}$. Define

$$\hat{\lambda} = \inf\left\{\lambda \in \Lambda : \hat{R}_n(\lambda) \leq \alpha - \frac{B-\alpha}{n}\right\}, \quad (3.9)$$

if the set (3.9) is non-empty, else define $\hat{\lambda} = \lambda_{\max}$. Then we have that

$$\mathbb{E}\left[\ell(Y_{n+1}, C_{\hat{\lambda}}(X_{n+1}))\right] \leq \alpha.$$

Proof. There are two key steps to this proof. In the first step, we consider

$$\tilde{\lambda} := \inf\left\{\lambda \in \Lambda : \hat{R}_{n+1}(\lambda) \leq \alpha\right\},$$

which depends on $(X_i, Y_i)_{i=1}^n$ and the test point (X_{n+1}, Y_{n+1}), and we use exchangeability to show that $\mathbb{E}\left[\ell(Y_{n+1}, C_{\tilde{\lambda}}(X_{n+1}))\right] \leq \alpha$. In the second step, we bound $\hat{R}_{n+1}(\lambda)$ by a quantity depending on $\hat{R}_n(\lambda)$ and $((X_i, Y_i))_{i=1}^n$. Combining this with the monotonicity of $\lambda \mapsto \ell(y, C_\lambda(x))$ will give the result.

First note that $\tilde{\lambda}$ is well-defined since $\hat{R}_{n+1}(\lambda_{\max}) \leq \alpha$. Note that right-continuity implies that $\hat{R}_{n+1}(\tilde{\lambda}) \leq \alpha$. We now use exchangeability. Specifically, recall from Remark 3.1 that if $Z_1, \ldots, Z_{n+1}$ are exchangeable, then $Z_{n+1} \mid \hat{P}_{n+1} \sim \frac{1}{n+1}\sum_{i=1}^{n+1} \delta_{Z_i}$, where $\hat{P}_{n+1}$ is the empirical distribution. Since $\hat{R}_{n+1}$ is invariant under permutations of $Z_1, \ldots, Z_{n+1}$, we have that $\tilde{\lambda}$ is $\hat{P}_{n+1}-$measurable. This implies that

$$\begin{aligned}
\mathbb{E}\left[\ell(Y_{n+1}, C_{\tilde{\lambda}}(X_{n+1}))\right] &= \mathbb{E}\left[\mathbb{E}\left[\ell(Y_{n+1}, C_{\tilde{\lambda}}(X_{n+1})) \mid \hat{P}_{n+1}\right]\right] \\
&= \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} \ell(Y_i, C_{\tilde{\lambda}}(X_i))\right] \\
&= \mathbb{E}\left[\hat{R}_{n+1}(\tilde{\lambda})\right] \leq \alpha.
\end{aligned}$$

We now relate $\hat{R}_{n+1}$ to $\hat{R}_n$. If the set in (3.9) is empty, then $\hat{\lambda} = \lambda_{\max}$, and so certainly $\hat{\lambda} \geq \tilde{\lambda}$. Otherwise, note that

$$\hat{R}_{n+1}(\hat{\lambda}) = \frac{n}{n+1}\hat{R}_n(\hat{\lambda}) + \frac{1}{n+1}\ell(Y_{n+1}, C_{\hat{\lambda}}(X_{n+1})) \leq \frac{n}{n+1}\hat{R}_n(\hat{\lambda}) + \frac{B}{n+1} \leq \alpha$$

by the definition of $\hat{\lambda}$ and right-continuity. Therefore, the definition of $\tilde{\lambda}$ implies that $\hat{\lambda} \geq \tilde{\lambda}$. Since $\lambda \mapsto \ell(C_\lambda(X_{n+1}), Y_{n+1})$ is decreasing, this implies that

$$\ell(C_{\hat{\lambda}}(X_{n+1}), Y_{n+1}) \leq \ell(C_{\tilde{\lambda}}(X_{n+1}), Y_{n+1}).$$

Thus, we have that

$$\mathbb{E}\left[\ell(C_{\hat{\lambda}}(X_{n+1}), Y_{n+1})\right] \leq \mathbb{E}\left[\ell(C_{\tilde{\lambda}}(X_{n+1}), Y_{n+1})\right] \leq \alpha.$$

□

We now show that split conformal prediction is a special case of conformal risk control. Take $\Lambda = (-\infty, \infty]$ and C_λ, ℓ as in (3.8). Then we may take $B = 1$ in the proof of Theorem 3.3, which gives that

$$\begin{aligned}
\hat{\lambda} &= \inf\left\{\lambda \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{Y_i \notin C_\lambda(X_i)\right\} \leq \alpha - \frac{1-\alpha}{n}\right\} \\
&= \inf\left\{\lambda \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\hat{s}(X_i, Y_i) > \lambda\right\} \leq \alpha - \frac{1-\alpha}{n}\right\} \\
&= \inf\left\{\lambda \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\hat{s}(X_i, Y_i) \leq \lambda\right\} \geq \frac{(1-\alpha)(n+1)}{n}\right\} \\
&= \inf\left\{\lambda \in \mathbb{R} : \hat{F}_{(\hat{S}_1, \ldots, \hat{S}_n)}(\lambda) \geq \frac{(1-\alpha)(n+1)}{n}\right\} = \hat{Q}_{\hat{S}}\left(\frac{\lceil (1-\alpha)(n+1)\rceil}{n}\right),
\end{aligned}$$

as claimed.

Overall, we see that the coverage guarantee provided by split conformal prediction is a special case of the much wider class of guarantees provided by conformal risk control, which control the expectation of any bounded, decreasing loss function.

We remark that it is possible to combine conformal risk control with the methods from Section 3.1 and Section 3.2 to extend its guarantees beyond the exchangeable setting, although we do not discuss the details here. The extension to the distribution shift setting is discussed in [Ang+24] (Proposition 3). To conclude this section, we state a result below relating conformal risk control with the NexCP method; note that we use the same notation as in Section 3.1. We omit the proof of this result, but we note that it combines the main ideas in the proofs of Theorem 3.3 and Theorem 3.1 and can be found in [Far+24].

Theorem 3.4 ([Far+24] Theorem 1)**.** *Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$ and that the prediction sets $C_\lambda(x)$ and the map $\lambda \mapsto \ell(y, C_\lambda(x))$ satisfy the same assumptions as in Theorem 3.3. Suppose that $A \leq \ell(y, C_\lambda(x)) \leq B$ for some constants A, B and any λ and (x, y). Given fixed weights $w_1, \ldots, w_n \geq 0$, define $\tilde{w}_i$ for $i \in [n+1]$ as in (3.1). Define the weighted empirical risk*

$$\hat{R}_n(\lambda) = \frac{\sum_{i=1}^n w_i \ell(Y_i, C_\lambda(X_i))}{\sum_{i=1}^n w_i}$$

and define

$$\hat{\lambda} = \inf\left\{\lambda \in \Lambda : \hat{R}_n(\lambda) \leq \alpha - \frac{B-\alpha}{\sum_{i=1}^n w_i}.\right\}$$

Then for any $\alpha \in [A, B]$,

$$\mathbb{E}\left[\ell(Y_{n+1}, C_{\hat{\lambda}}(X_{n+1}))\right] \leq \alpha + (B-A)\sum_{k=1}^n \tilde{w}_k \, \mathrm{d_{TV}}(D, D^{(k)}).$$

4 Conclusion

To conclude this essay, we give an overview the topics discussed in this essay, the advantages and disadvantages of conformal prediction and possible directions for future work.

The results of Section 2 illustrate some of the key advantages of conformal prediction. In particular, we demonstrated that conformal predicition can be easily and efficiently combined with any base predicition method and guarantees coverage assuming only exchangeability of the data. Therefore, conformal prediction has the scope to be combined with a range of machine learning and deep learning methods, where uncertainty quantification is otherwise highly challenging. For example, conformal prediction has been applied to image segmentation [Ang+24; MDA24] and tasks in natural language processing [Cam+24; CGC24], and future work might explore further ways to tailor conformal prediction to specific application areas.

The methods presented in Section 3 are important as they extend the fundamental theory of conformal prediction to settings where exchangeability does not hold, thus widening its applicability. Indeed, the weighted conformal prediction procedure of Section 3.2 has found application in a range of areas of statistics including causal inference [LC21; JRC23], survival analysis [CLR23] and selective inference [JC23]. As mentioned in Section 3.1, tackling the question of choosing the fixed weights in the NexCP method may be a direction for future work. A signficant theoretical development is given by the recent work of [BT25], which consolidates the methods of Section 3.1 and Section 3.2 into a single theoretical framework. The authors of [BT25] not only show that their framework includes the NexCP method

and weighted conformal prediction as special cases, but also give examples of how it can be used to derive new results in more challenging settings, e.g. a variant of weighted conformal prediction where distribution shift also occurs within the training data. A possible avenue for further work may involve using the framework of [BT25] to develop variants of conformal prediction that provide theoretical guarantees in a variety of settings.

Whilst not the focus of this essay, an important limitation of conformal prediction, as mentioned in Section 2.3.1, is that it does not provide a test-conditional coverage guarantee. The topic of conditional coverage is a major research topic in conformal prediction and works in the area include results showing conditional coverage is impossible in some settings [LW14] and theoretical guarantees for certain relaxations of conditional coverage [GCC25].

Overall, conformal prediction is a promising approach to uncertainty quantification, which has recently seen significant developments. In this essay, we provided an introduction to the field by discussing both the mathematical foundations of conformal prediction, as well as a selection of recent advances in the area.

References

[VGS] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Vol. 29. Springer.

[Joh24] Reid A. Johnson. "quantile-forest: A Python Package for Quantile Regression Forests". In: *Journal of Open Source Software* 9.93 (2024), p. 5976. DOI: 10.21105/joss.05976. URL: https://doi.org/10.21105/joss.05976.

[MR06] Nicolai Meinshausen and Greg Ridgeway. "Quantile regression forests." In: *Journal of machine learning research* 7.6 (2006).

[ABB24] Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. *Theoretical Foundations of Conformal Prediction*. 2024. arXiv: 2411.11824 [math.ST]. URL: https://arxiv.org/abs/2411.11824.

[RPC19] Yaniv Romano, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression". In: *Advances in neural information processing systems* 32 (2019).

[Lei+18] Jing Lei et al. "Distribution-free predictive inference for regression". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1094–1111.

[AB21] Anastasios N Angelopoulos and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". In: *arXiv preprint arXiv:2107.07511* (2021).

[Tib24a] Ryan Tibshirani. *Advanced Topics in Statistical Learning*. https://www.stat.berkeley.edu/~ryantibs/statlearn-s24/lectures/conformal.pdf. [Online lecture notes]. 2024.

[Tib24b] Ryan Tibshirani. *Advanced Topics in Statistical Learning Homework 4*. https://www.stat.berkeley.edu/~ryantibs/statlearn-s24/homeworks/homework4.pdf. [Online homework sheet]. 2024.

[Koe05] Roger Koenker. *Quantile regression*. Vol. 38. Cambridge university press, 2005.

[Bar+23] Rina Foygel Barber et al. "Conformal prediction beyond exchangeability". In: *The Annals of Statistics* 51.2 (2023), pp. 816–845.

[Ang+24] Anastasios Nikolas Angelopoulos et al. "Conformal Risk Control". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=33XGfHLtZg.

[Tib+19] Ryan J Tibshirani et al. "Conformal prediction under covariate shift". In: *Advances in neural information processing systems* 32 (2019).

[PR21] Aleksandr Podkopaev and Aaditya Ramdas. "Distribution-free uncertainty quantification for classification under label shift". In: *Uncertainty in artificial intelligence*. PMLR. 2021, pp. 844–853.

[Bar+24] Rina Foygel Barber et al. "De Finetti's theorem and related results for infinite weighted exchangeable sequences". In: *Bernoulli* 30.4 (2024), pp. 3004–3028.

[Tan23] Wenpin Tang. "Finite and infinite weighted exchangeable sequences". In: *arXiv preprint arXiv:2306.11584* (2023).

[SSK12] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[Far+24] António Farinhas et al. "Non-Exchangeable Conformal Risk Control". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=j511LaqEeP.

[MDA24] Luca Mossina, Joseba Dalmau, and Léo Andéol. "Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 3574–3584.

[Cam+24] Margarida M. Campos et al. *Conformal Prediction for Natural Language Processing: A Survey*. 2024. arXiv: 2405.01976 [cs.CL]. URL: https://arxiv.org/abs/2405.01976.

[CGC24] John Cherian, Isaac Gibbs, and Emmanuel Candes. "Large language model validity via enhanced conformal prediction methods". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 114812–114842.

[LC21] Lihua Lei and Emmanuel J Candès. "Conformal inference of counterfactuals and individual treatment effects". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 83.5 (2021), pp. 911–938.

[JRC23] Ying Jin, Zhimei Ren, and Emmanuel J Candès. "Sensitivity analysis of individual treatment effects: A robust conformal inference approach". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2214889120.

[CLR23] Emmanuel Candès, Lihua Lei, and Zhimei Ren. "Conformalized survival analysis". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85.1 (2023), pp. 24–45.

[JC23] Ying Jin and Emmanuel J Candès. "Model-free selective inference under covariate shift via weighted conformal p-values". In: *arXiv preprint arXiv:2307.09291* (2023).

[BT25] Rina Foygel Barber and Ryan J. Tibshirani. *Unifying Different Theories of Conformal Prediction*. 2025. arXiv: 2504.02292 [math.ST]. URL: https://arxiv.org/abs/2504.02292.

[LW14] Jing Lei and Larry Wasserman. "Distribution-free prediction bands for non-parametric regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76.1 (2014), pp. 71–96.

[GCC25] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. "Conformal prediction with conditional guarantees". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2025), qkaf008.

[Har12] Matthew T Harrison. "Conservative hypothesis tests and confidence intervals using importance sampling". In: *Biometrika* 99.1 (2012), pp. 57–69.

A Additional Results

Lemma A.1. *If V is a real-valued random variable with continuous cumulative distribution function F, then $F(V) \sim \text{Uniform}(0,1)$.*

Proof. Define $F^- : (0,1) \to \mathbb{R}$ by $F^-(y) \coloneqq \inf\{x \in \mathbb{R} : F(x) \geq y\}$. If $F(x) \geq y$ for some $x \in \mathbb{R}$ and $y \in (0,1)$, then $x \geq F^-(y)$ by the definition of the infimum. If $x \geq F^-(y)$, then $F(x) \geq F(F^-(y)) \geq y$ since F is increasing and right-continuous. Therefore, $\{x \in \mathbb{R} : F(x) \geq y\} = [F^-(y), \infty)$. Take a sequence $x_n \uparrow F^-(y)$, then $F(x_n) < y$, but the continuity of F implies $F(x_n) \to F(F^-(y))$. Therefore, $F(F^-(y)) \leq y$. We deduce that $F(F^-(y)) = y$ for any $y \in (0,1)$.

Now let $U \sim \text{Uniform}(0,1)$, then

$$\mathbb{P}\left(F^-(U) \leq x\right) = \mathbb{P}\left(U \leq F(x)\right) = F(x)$$

for any $x \in \mathbb{R}$, so $F^-(U) \overset{d}{=} V$. Thus for any $\beta \in (0,1)$, we have that

$$\mathbb{P}\left(F(V) \leq \beta\right) = \mathbb{P}\left(F(F^-(U)) \leq \beta\right) = \mathbb{P}\left(U \leq \beta\right) = \beta.$$

Thus $F(V) \sim \text{Uniform}(0,1)$. □