New Advances in Conformal Prediction

April 13, 2025

1 Introduction

Quantifying the uncertainty in a prediction produced by a regression or classification model is important to enable these models to be confidently deployed in real-world settings. This typically means that a procedure outputs a set of predictions, as opposed to a single prediction, which contains the true value of the response with a high probability. For example, it is well-known that one can construct exact prediction intervals in the normal linear model and asymptotically valid prediction intervals for generalised linear models. In Bayesian statistics, one can calculate the posterior predictive distribution to quantify uncertainty. An important limitation of all of these methods is that they make strong distributional assumptions on the data-generating process and in some cases, only provide asymptotic guarantees.

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 appeal of conformal prediction is that it can be combined with any existing predictive model to generate a valid prediction interval. In particular, we need not make any assumptions on this predictive model. Due to this, and the continuing success of modern machine learning methods at predictive tasks, conformal prediction is 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 we would like to predict the true response Y_{n+1} corresponding to a new predictor X_{n+1}. We aim to use $(X_1, Y_1), \ldots, (X_n, Y_n)$ to construct a prediction set $C(X_{n+1})$ such that

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

where $\alpha \in (0, 1)$ is a chosen confidence 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 above coverage guarantee. 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 above coverage guarantee. To conclude Section 2, the aim of Section 2.4 is to examine specific instances of split conformal prediction that improve the empirical performance of the procedure and demonstrate this with numerical experiments.

In Section 3, we explore extensions to the standard conformal prediction framework presented in Section 2. Whilst exchangeability was 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 designed to account for this is presented in Section 3.2.

1.1 Notation

Throughout the essay, we use the following notation.

For any positive integer n, we define $[n] = \{1, \ldots, n\}$ and S_n denotes the symmetric group on n elements - i.e. the permutations on n elements. If $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and $\sigma \in S_n$, then $\sigma(v) = (v_{\sigma(1)}, \ldots, v_{\sigma(n)})$.

We write $\mathcal{X}$ and $\mathcal{Y}$ to denote the (measurable) spaces of predictors and responses, respectively. We write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Typically, we think of $\mathcal{X} = \mathbb{R}^p$ for some positive integer p and $\mathcal{Y} = \mathbb{R}$ in the regression setting and $\mathcal{Y} = [K]$, for some positive integer K, in the classification setting.

2 Foundations of Conformal Prediction

In this chapter, we introduce the key mathematical tools for this essay. The main theoretical result of this section is Theorem 2.1 which establishes the coverage guarantee for full conformal prediction.

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], 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

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

Equivalently, $Z_1, ..., Z_n$ are exchangeable if for any measurable set A and for any $\sigma \in S_n$, we have that

$$\mathbb{P}\left((X_1, \ldots, X_n) \in A\right) = \mathbb{P}\left((X_{\sigma(1)}, \ldots, X_{\sigma(n)})\right) \in A.$$

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, we see that exchangeability is a weaker condition that being i.i.d.

Remark 2.2. Another way to view exchangeability is as follows. Suppose the random variables $Z_1, \ldots, Z_n \in \mathbb{R}$ 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 ordering.

Definition 2.2. Let $w = (w_1, \ldots, w_n) \in [0, 1]^n$ be such that $\sum_{i=1}^n w_i = 1$ and let $z \in \mathbb{R}^n$. We define the following quantities.

(i) The *weighted empirical cumulative distribution function* of z with respect to the *weights* w is the function $F_z^w : (-\infty, \infty] \to [0,1]$ given for $x \in (-\infty, \infty]$ by

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

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

$$\hat{F}_z(x) = \hat{F}^{(1/n,\ldots,1/n)}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{z_i \leq x\}.$$

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

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

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

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

The weighted versions of the quantities above will play an important role in Section 3. Therefore, some of the results of this section are proven for the weighted quantities even if only the special case $w = (1/n, \ldots, 1/n)$ is required. The following lemma formalises a sense in which the empirical cumulative distribution and quantile functions are inverses of each other.

Lemma 2.1. *Let $z \in \mathbb{R}^n$, $w = (w_1, \ldots, w_n) \in [0,1]^n$ be such that $\sum_{i=1}^n w_i = 1$ 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) The first claim follows from upon noting that $\hat{F}_z^w$ is right-continuous. Indeed, by the definition of the infimum, there exists a sequence (x_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, $\hat{F}_z$ jumps by $\frac{1}{n}$ at each z_i, for $i \in [n]$. Therefore

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

□

We now prove a lemma which is fundamental in the proof of the coverage guarantee in section Section 2.2 and links all three of the concepts introduced 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 $S_\beta = \left\{y \in \mathbb{R}^n : \; y_j \leq \hat{Q}_y(\beta)\right\}$. Define $\tau \in S_n$ to be the transposition exchanging 1 and j. We have that

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

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

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

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. This proves the lemma. □

Remark 2.3. In the case where $Z_1, \ldots, Z_n$ are almost surely distinct, we can obtain the above result using an even simpler argument. Fix $i \in [n]$. As discussed in Remark 2.2, each of the $n!$ orderings of $Z = (Z_1, \ldots, Z_n)$ 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 $\lceil \beta n \rceil$ gives that

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

2.2 Full conformal prediction

In this subsection, we present the full conformal prediction algorithm and prove its coverage guarantee. The presentation of the material in this subsection and the proof of Theorem 2.1 is inspired by [ABB24].

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

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

Remark 2.4. Whilst the conformity score may theoretically be an arbitary function - as in the above display - we now explain how the conformity score should be understood in practice. The first argument of the conformity score represents an arbitary 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, where a high 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 using the data in the second argument. Following this intuition, we will also refer to a value of s given by $s(z; D)$ as the conformity score of the test point 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 $i \in [n]$. 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 $x \mapsto \mathbb{E}(Y|X = x)$ based 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 in classification setting? Words before definition below.

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 $z \in \mathcal{Z}$, $D \in \cup_{j \geq 1} \mathcal{Z}^j$ and $j \in \mathbb{N}$, we have that

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

We now give a brief informal description of how the coverage guarantee is obtained. 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. As discussed in Remark 2.2 and Remark 2.3, this means that any ordering of the conformity scores is equally likely, so 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 primary idea used to construct the prediction set. However, since Y_{n+1} is unknown, will instead consider the test point (X_{n+1}, y) and select those $y \in \mathcal{Y}$ to be included the prediction set, whose conformity score falls in the bottom $1 - \alpha$ fraction.

Before formally proving the coverage guarantee, we introduce the notation for the quantities mentioned in the outline above. 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

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

and

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

Additionally, let

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

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 $S \coloneqq (S_1, \ldots, S_{n+1})$ is 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 $z = (z_1, \ldots, z_n) \in \mathcal{Z}^n$. 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 $(Z_1, \ldots, Z_n)$, 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

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

so S is exchangeable.

Finally, note 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 derive an equivalent characterisation of the prediction set that we will use in the remainder of this essay and that will be particularly useful in the next subsection. In order to do this, we first present a lemma.

Lemma 2.3. *Let $z \in \mathbb{R}^n$, $y \in \mathbb{R}$ and $w \in [0,1]^n$ such that $\sum_{i=1}^n w_i = 1$. 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

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

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. □

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 $\frac{1}{n+1}\sum_{i=1}^n \mathbb{1}\{x \geq z_i\} \leq \frac{n}{n+1}$. Now observe that for any $x < \infty$, we have that

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

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

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

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

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

□

Remark 2.5. 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^y_{n+1}, which is stated in Theorem 2.2 below. This will be particularly important in Section 2.3.

The following reformulation of Theorem 2.1 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*

$$C(X_{n+1}) = \left\{y \in \mathcal{Y} : \quad S^y_{n+1} \leq \hat{Q}_{(S^y_1,\ldots,S^y_n,\infty)}(1-\alpha)\right\}. \tag{2.3}$$

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

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

Remark 2.6. 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 $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)} = \infty$, which trivially covers all test points. 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

$$\hat{Q}_z\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}\right) = \inf\left\{y \in \mathbb{R} : \sum_{i=1}^{n} \mathbb{1}\left\{z_i \leq y\right\} \geq \lceil (1-\alpha)(n+1) \rceil\right\} = z_{(\lceil (1-\alpha)(n+1) \rceil)},$$

where $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$ are the *order statistics* of z, defined by

$$z_{(k)} = \inf\left\{y \in \mathbb{R} : \sum_{i=1}^{n} \mathbb{1}\left\{z_i \leq y\right\} \geq k\right\},$$

for $k \in [n]$. This means the calculation of the quantile $\hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}$ is simplified to finding the $\lceil (1-\alpha)(n+1) \rceil$ smallest element in the list $(\hat{S}_1, \ldots, \hat{S}_n)$. This leads to Algorithm 1 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 ← ∅
if ⌈(1 − α)(n + 1)⌉ > n then
    C = 𝒴
else
    for y ∈ 𝒴 do
        Compute S_i^y = s((X_i, Y_i); D^y).
        Compute S_{n+1}^y = s((X_{n+1}, y); D^y).
        Set S^y = (S_1^y, ..., S_n^y).
        Compute Q̂ as the ⌈(1 − α)(n + 1)⌉ element in the list S_1^y, ..., S_n^y.
        if S_{n+1}^y ≤ Q̂ then
            C ← C ∪ {y}.
        end if
    end for
end if
```

Output: C

Remark 2.7. Something about the prediction set above not being obviously computable. Can do in specific cases (need references) but may not be possible in general. In above algo, if $\mathcal{Y}$ is not discrete, problematic, need to discretise. Also, the form of the prediction set not at all obvious Split conformal resolves some of this... (e.g. below we will see we can get interval). Dependence of the form of the prediction set on the conformity score also not obvious.

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 the coverage guarantee from Section 2.2 also holds for split conformal prediction. We will also compare full and split conformal prediction, discussing their respective advantages and disadvantages.

For split conformal prediction, 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 disjoint from the calibration data $D_{\text{cal}} := ((X_i, Y_i))_{i=1}^n$. Split conformal prediction then uses the calibration data together with $\hat{s}$ and X_{n+1}, to form a prediction set for Y_{n+1}.

Before proving the coverage guarantee for split conformal prediction, we introduce the necessary notation. Let $\hat{S}_i = \hat{s}(X_i, Y_i)$ for $i \in [n]$ and let $\hat{S}_{n+1}^y = \hat{s}(X_{n+1}, y)$ 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*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ \hat{S}_{n+1}^y \leq \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1-\alpha) \right\}. \tag{2.4}$$

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 non-random 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

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

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

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

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

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

The result follows by marginalising 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**
    $C = \mathcal{Y}$
  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

Remark 2.8. Although D_{tr} and $\hat{s}$ can be arbitary for the coverage guarantee to hold, in practice, we usually split the full training set into a proper training set D_{tr} and a calibration set D_{cal}. We fit a model on the proper training set to obtain a conformity score $\hat{s}$, and then apply the split conformal prediction procedure given in Section 2.3 to the calibration data to obtain the prediction set.

Remark 2.9. Note that the split conformal coverage guarantee 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. Moreover, split conformal prediction has the advantage that it is more computationally efficient than full conformal prediction. Indeed - as discussed in the above remark - in split conformal prediction, we must only fit the model once to obtain $\hat{s}$. However, in full conformal prediction, we must refit the model for each $y \in \mathcal{Y}$ since we compute the conformity score with respect to a dataset that includes the point (X_{n+1}, y). On the other hand, full conformal prediction has the advantage that it uses all of the training data to fit the model, whereas in split conformal prediction, we may typically only use half of the full training set as the proper training set and the other half as the calibration set.

Example 2. Consider the regression setting as in Example 1 and assume $p = 1$ so that $\mathcal{X} = \mathbb{R}$. Suppose $\hat{\mu}$ is an estimate of the regression function obtained from the 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*. Note that in this case the prediction set is an interval centered at $\hat{\mu}(x)$ since

$$\begin{aligned} C(x) &= \left\{ y \in \mathbb{R} : \, |y - \hat{\mu}(x)| \leq \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha) \right\} \\ &= \left[\hat{\mu}(X_{n+1}) - \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha), \; \hat{\mu}(x) + \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha) \right]. \end{aligned} \tag{2.5}$$

It is important to note that this simplified form of the prediction set is a consequence of Lemma 2.3 and 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 the proper training set D_{tr}. The latter is a consequence of both the split conformal algorithm, which ensures $S_i^y = \hat{s}(X_i, Y_i)$ is independent of y (as in the proof of Theorem 2.3), and Lemma 2.3, which ensures the quantile used in the prediction set depends only on D_{cal}.

We now present a numerical experiment to provide a concrete example of split conformal prediction to illustrate the ideas from this subsection. As mentioned in the above remarks, both the form of the conformity score $\hat{s}$ and the fitting procedure used to obtain it may, in theory, be arbitary. However, in practice, both of these can significantly affect the final prediction set. The choice of the form of the conformity score will be discussed in more detail in Section 2.4. In the numerical experiment below, we demonstrate the effect of the choice of the fitting procedure on the prediction set.

We conduct a numerical experiment with 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

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

We use this data-generating process to generate independent proper training, calibration and test datasets with sizes 1000, 3000 and 3000, respectively. We train regression models on the proper training set using both linear regression and random forests, where hyperparameters of the random forest are chosen via $5-$fold cross-validation. We then use split conformal prediction with the absolute residual score on the calibration dataset to prediction sets for each data point in the test dataset. The code for this numerical experiment is provided in [insert filename]. We now compare the prediction intervals obtained using both of these models.

Given a test dataset D_{test} with i.i.d data points, as above, we may compute the *empirical coverage*

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

as an estimate of the true coverage. By Theorem 2.3, we expect the empirical coverage to be close to $1 - \alpha$ for any choice of fitting procedure and conformity score, provided $|D_{\text{test}}|$ is large. Indeed, for this

numerical example, we find that the empirical coverage is 0.8993 and 0.9053 for the linear regression and random forest model, respectively. Although both methods provide the desired coverage, it is clear from Figure 1 that the random forests model is a better fit compared to the linear regression model. This is reflected by the fact that the average length of the conformal prediction intervals in the random forest model (1.6978) is smaller than that in the linear regression model (2.0773). Mathematically, we can explain this by considering (2.5). Since the distribution of the absolute residual scores of the calibration data points is more skewed towards zero (due to the random forests model having a better fit), the quantile $\hat{Q}_{\hat{S}}\left(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}\right)$ will be lower for the random forests model and so the prediction intervals will be shorter on average. Overall, this numerical experiment highlights that whilst any model fitting procedure can theoretically be used, a better fitting procedure is more desirable as the resulting conformal prediction intervals are on average narrower.

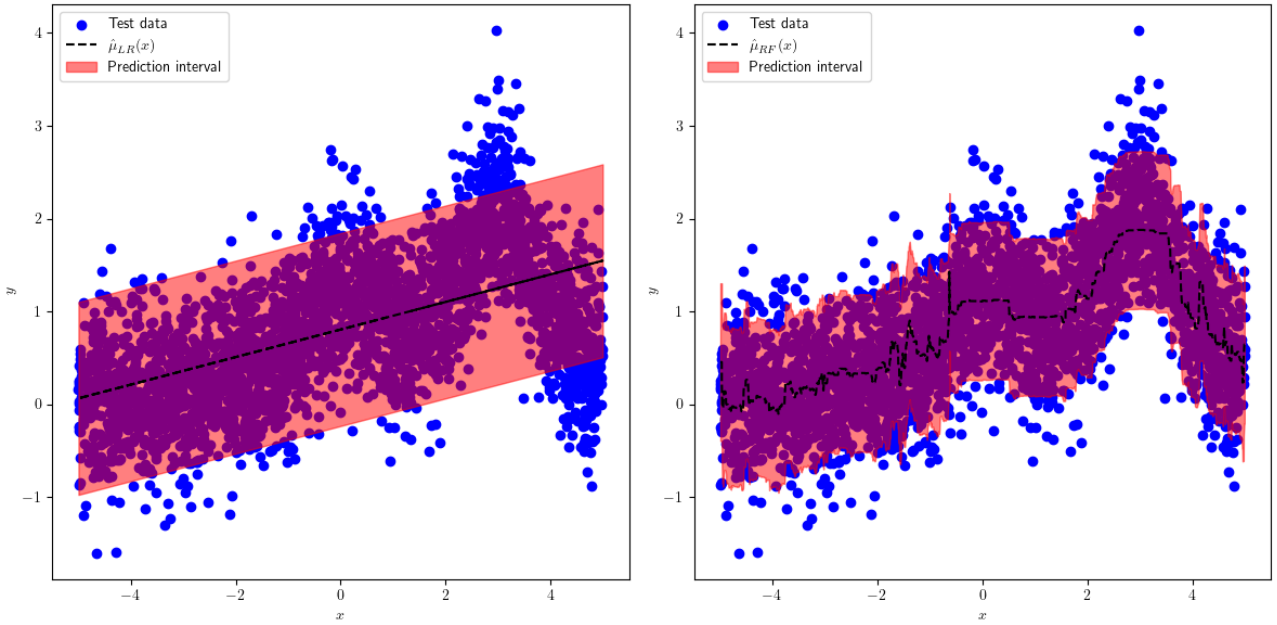


Figure 1: Plot of the test data, the fitted regression function and the split conformal prediction interval using linear regression (left) and random forests (right). The fitted regression function is denoted $\hat{\mu}_{LR}$ (left) and $\hat{\mu}_{RF}$ (right).

2.3.1 Training conditional coverage

Note that the coverage guarantee in Theorem 2.3 is marginal over the calibration data D_{cal}, i.e. the coverage guarantee is equivalent to

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

This means that for a randomly drawn calibration set and test point, the prediction set contains the test point with probability $1-\alpha$. However, in practice, it may often be the case that only one calibration set is available, and we wish to understand the probability of a random test point being covered conditional on this calibration set, i.e. the quantity

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

Note that this is a random quantity since it is a function of the calibration data. 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 [Tib24a].

Lemma 2.5 ([Tib24b]). *Suppose* $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ *are i.i.d. and suppose that* $\hat{S}_i$ *has a continuous distribution. Then*

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1}) \,|\, D_{\text{cal}}\right) \sim \text{Beta}(k_\alpha, n + 1 - k_\alpha), \tag{2.6}$$

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

Proof. 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 at least k of the random variables $U_1, \ldots, U_n$ must be less than or equal to x and $\mathbb{P}(U_i \leq x) = x$ for any $i \in [n]$. Therefore, if $g(z) = z^r(1-z)^{n-r}$we have that

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

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 (maybe include quick proof) 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

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

by remark to be added, and so the result follows.

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 (2.5). However, it is not immediately clear from (2.5) how the choice of conformity score affects the prediction set. We gain some insight into this from Example 2, 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. 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 them through numerical experiments. Throughout this subsection, we use split conformal prediction.

2.4.1 Regression

We work in the regression setting as in Example 2 with $\mathcal{X} = \mathbb{R}$. Consider the absolute residual score and its corresponding predicition interval (2.5). 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. $\mathrm{Var}(Y|X = x)$ is not constant in x, then the prediction interval (2.5) does not accurately capture the uncertainty in Y given $X = x$. This is closely related to the fact that Theorem 2.3 only guarantees *marginal coverage*

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

as opposed to *test-conditional coverage*

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

This means that conformal prediction does not guarantee $1 - \alpha$ level coverage at every $x \in \mathcal{X}$, but rather only on average over x. The two conformity scores we present aim to make the prediction interval adaptive to heteroscedasticity.

2.4.2 Locally Weighted Residual Score

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

$$\hat{s}(x, y) = \frac{|y - \hat{\mu}(x)|}{\hat{\sigma}(x)}. \tag{2.7}$$

The corresponding prediction set is given by

$$C(x) = \left[\hat{\mu}(x) - \hat{\sigma}(x)\, \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha),\ \hat{\mu}(x) + \hat{\sigma}(x)\, \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)\right], \tag{2.8}$$

using the notation of Section 2.3.

This conformity score was originally introduced in [Lei+18] and aims to account for heteroscedasticity but scaling the width of the interval in (2.5) 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 $(X_i, Y_i) \in D_{\mathrm{tr}}$ to obtain $\hat{\mu}$ and then regress $|Y_i - \hat{\mu}(X_i)|$ onto X_i for $(X_i, Y_i) \in D_{\mathrm{tr}}$.

2.4.3 Conformalised Quantile Regression

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

$$q_\tau(x) = \inf\left\{z \in \mathbb{R} : \mathbb{P}\left(Y \leq z | X = x\right) \geq \tau\right\}, \quad \tau \in (0,1) \tag{2.9}$$

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 test-conditional coverage. The approach of estimating $q_\tau(x)$ is referred to as *quantile regression*. In this essay, we do not discuss details of the numerous methods for constructing quantile regression estimators. However, we note that the following important fact.

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

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

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 $[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$ may not be guaranteed to have $(1-\alpha)-$level coverage in finite samples. *Conformalised quantile regression* [RPC19] calibrates this interval using conformal prediction to provide it with a finite-sample 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

$$\hat{s}(x, y) = \max\left\{\hat{q}_{\alpha/2}(x) - y,\, y - \hat{q}_{1-\alpha/2}(x)\right\}. \tag{2.10}$$

This results in the conformal prediction interval

$$\begin{aligned}&\left\{y \in \mathbb{R} : \hat{q}_{\alpha/2}(X_{n+1}) - y \leq \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha) \quad \text{and} \quad y - \hat{q}_{1-\alpha/2}(X_{n+1}) \leq \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)\right\} \\ &= \left[\hat{q}_{\alpha/2}(X_{n+1}) - \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha), \hat{q}_{1-\alpha/2}(X_{n+1}) + \hat{Q}_{(\hat{S}_1,\ldots,\hat{S}_n,\infty)}(1-\alpha)\right]. \end{aligned} \tag{2.11}$$

To intuitively understand this score, we note the following. 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, $\hat{s}(x, y) \geq 0$, 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. Note that this contrasts to the conformity scores we have seen previously, which could only take on non-negative values. We see from the above that if the estimated interval $[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$ overcovers, then conformalised quantile regression narrows the interval, and if the estimated interval undercovers, then conformalised quantile regression widens the interval.

2.4.4 Numerical Experiments

We now present numerical experiments 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:**

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

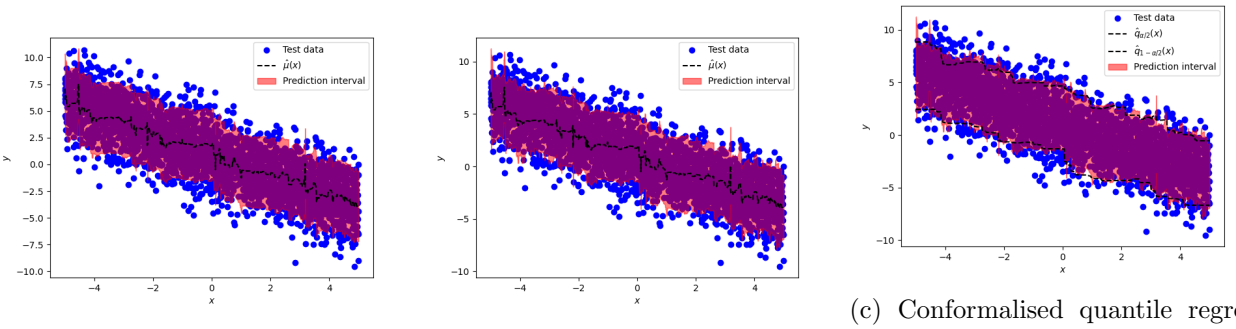
for all $i \in [n]$.

In each setting, we generate independent proper training, calibration and test datasets, set $\alpha = 0.1$ and generate conformal prediction intervals for the test dataset using the absolute residual score, the locally weighted residual score and conformalised quantile regression. Further details on the exact implementation and the code are provided in the appendix. In Table 1, we record the empirical coverage and the average length of the prediction intervals calculated on the test dataset. In Figure 2 and Figure 3, we plot the prediction intervals obtained using these three methods.

We observe that all three methods provide the target coverage in both settings. This is guaranteed by the theoretical coverage guarantee for split conformal prediction in Theorem 2.3 since all the data points are i.i.d., and thus exchangeable. However, we clearly see in Figure 3 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. Since the absolute residual score generates prediction intervals whose width is constant (in x), we see in Figure 3a that these overcover in some regions and undercover in others. Therefore, the absolute residual score yields wider prediction intervals as compared to the other two methods, which can be seen in Table 1. This is further emphasised in Figure 3d, where we estimate the conditional coverage $\mathbb{P}(Y_{n+1} \in C(X_{n+1})|X_{n+1})$ by dividing the covariate space $(-5, 5)$ into smaller subintervals and calculating the empirical coverage for the test dataset on each subinterval. This plot demonstrates that the locally weighted score and conformalised quantile regression provide improved conditional coverage compared to the absolute residual score.

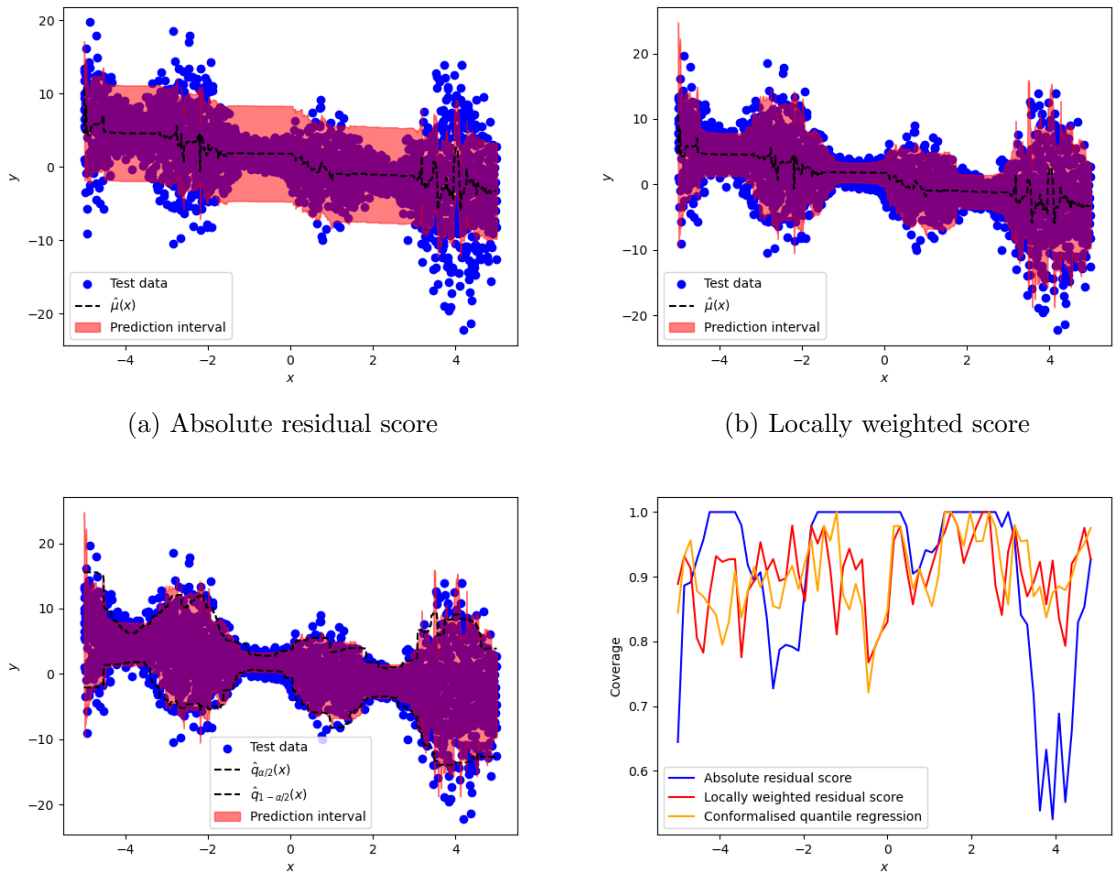
	Absolute residual		Locally weighted		Conformalised quantile regression	
	Coverage	Average length	Coverage	Average length	Coverage	Average length
Setting 1	0.9010	6.616	0.9027	6.632	0.9017	6.712
Setting 2	0.9003	13.03	0.9063	11.58	0.9027	11.55

Table 1: Empirical 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 2: Conformal prediction intervals in setting 1.



(a) Absolute residual score

(b) Locally weighted score

(c) Conformalised quantile regression

(d) Local coverage

Figure 3: Conformal prediction intervals in setting 2.

3 Extensions of Conformal Prediction

In Section 2, we saw that the coverage guarantee of full conformal prediction relied on two key assumptions: the exchangeability of the data and the symmetry of the conformity score. We saw that if the exchangeability assumption is violated, the coverage may deviate significantly from the target $1 - \alpha$ level. Something about why it is important to generalise to nonexchangeable data and examples (e.g. distribution shift) and non-symmetric conformity scores (e.g. can use algorithm that does not treat data symmetrically such as WLS.)

3.1 Nonexchangeable Conformal Prediction

In this subsection, we present the *nonexchangeable conformal prediction* (NexCP) method developed in [Bar+23]. The key theoretical contribution of [Bar+23], presented in Theorem 3.1, is the development of a weighted conformal prediction procedure for which a coverage guarantee can be derived when the data is not assumed to be exchangeable and the conformity score is not assumed to be symmetric.

Before we present the main theorem of this subsection, we introduce the required notation. Given data

points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \in \mathcal{Z}$, $y \in \mathcal{Y}$ 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 $n+1$.

Let s be conformity score. As mentioned above, we will not assume that s is symmetric. This means that 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

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

The NexCP method assigns *weights* $w_1, \ldots, w_n \in [0, \infty)$ to the data points $(X_1, Y_1), \ldots, (X_n, Y_n)$, respectively. The weights are fixed, non-negative real numbers and 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 $i \in [n]$.

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.

Lastly, we recall the definition of the total variation distance.

Definition 3.1. Let U and V be random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *total variation distance* between U and V is defined as

$$\mathrm{d_{TV}}(U, V) = \sup_{A \in \mathcal{F}} |\mathbb{P}(U \in A) - \mathbb{P}(V \in A)|.$$

We now prove the main theorem of this subsection which provides the predicition set and coverage guarantee for the NexCP method.

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

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha - \sum_{k=1}^n \tilde{w}_k \mathrm{d_{TV}}(S, S^{(k)}). \tag{3.4}$$

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 $K = n+1$, this is shown in the proof of Lemma 2.3. If $K \leq n$, we have that for any $x \in \mathbb{R}$,

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

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 $i = 1, \ldots, n+1$
    Compute $\hat{Q}^{\tilde{w}}_{S^{y,(K)}}$
    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

Before discussing the implications of Theorem 3.1, we note that, in the same way that split conformal prediction is 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*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ \hat{S}^y_{n+1} \leq \hat{Q}^w_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1-\alpha) \right\}. \tag{3.5}$$

Then we have that

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

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 interpret the result in Theorem 3.1, following the points made in [Bar+23]. 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

$$\sum_{k=1}^{n} \tilde{w}_k \mathrm{d_{TV}}(S, S^{(k)}). \tag{3.6}$$

may be interpreted as the loss in coverage that occurs due to not assuming exchangeability or that the conformity score is symmetric. We highlight some important special cases of Theorem 3.1.

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*

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

Proof. Since s is symmetric, we have that $S^{y,(K)}_i = s((X_i, Y_i); D^{y,(K)}) = s((X_i, Y_i); D^y) = S^y_i$ for $i \in [n]$ and similarly, $S^{y,(K)}_{n+1} = S^y_{n+1}$. 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. □

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 a measure of the extent to which exchangeability is violated, then Corollary 3.2 shows that smaller the violation of exchangeability, the smaller the loss in coverage.

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*

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

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. □

We see that Corollary 3.3 is in fact a generalisation of Theorem 2.1 since it has the same assumptions and coverage guarantee as Theorem 2.1 but allows for conformity scores that are not necessarily symmetric. Therefore Corollary 3.3 shows that in the special case that the data is exchangeable but the conformity score is not symmetric, the NexCP method may be used to recover the same $(1-\alpha)-$level coverage guarantee as in Theorem 2.1.

Overall, this shows that in addition to quantifying the coverage loss in the case where the data

3.2 Distribution Shift

In this subsection, we focus on a specific kind of violation of exchangeability - 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 $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ from 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, so Theorem 2.1 does not apply. We discuss a conformal predicition procedure originally developed in [Tib+19] and [PR21] which enables the construction of a prediction set with at least $1-\alpha$ coverage in the above setting.

We first introduce a notion that generalises exchangeability, referred to as *weighted exchangeability* ([Tib+19; Bar+24; Tan23]). In what follows $\mathcal{U}$ 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.

Note 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 $\lambda_1, \ldots, \lambda_n \equiv 1$. The lemma below demonstrates how the notion of weighted exchangeability applies to our 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

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

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

We now state an important lemma on weighted exchangeability that will be used to establish the desired coverage guarantee. The interpretation of this result, as mentioned in [Bar+24], is that conditional on the unordered values $\{U_1, \ldots, U_k\}$, the distribution U_i is a discrete distribution over $\{U_1, \ldots, U_k\}$ where the probabilities may be explicity expressed in terms of λ. The probability of taking on the value U_j is denoted by $(w_{k,i}(U_1, \ldots, U_k))_j$ below. The way the authors of [Bar+24] formally express conditioning on the unordered values $\{U_1, \ldots, U_k\}$ is by conditioning on the sub-$\sigma-$algebra $\mathcal{E}_k$ of $\mathcal{B}(\mathcal{U}^k)$ satisfying $(u_1, \ldots, u_k) \in \mathcal{E}_k \iff (u_\sigma(k), \ldots, u_\sigma(k)) \in \mathcal{E}_k$ for any $\sigma \in S_k$. We now state the lemma but omit the proof since it consists primarily of measure-theoretic calculations regarding conditional distributions. The proof may be found in [Bar+24]. Add reference to say that this is same as conditioning on empirical dist + defn of empirical dist.

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*

$$U_i | \hat{P}_k \sim \tilde{P}_{k,i},$$

where $\tilde{P}_{k,i} = \sum_{j=1}^k \left(w_{k,i}(U_1, \ldots, U_k)\right)_j \, \delta_{U_j}$ and

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

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*

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

where

$$w_i = \frac{\frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)}{\sum_{j=1}^{n+1} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_j)}, \quad i \in [n+1]. \tag{3.7}$$

Proof. By Lemma 3.2 we have that

$$Z_{n+1}|\hat{P}_{n+1} \sim \tilde{P}_{n+1,n+1},$$

where $\tilde{P}_{n+1,n+1} = \sum_{i=1}^{k} \left(w_{n+1,n+1}(Z_1, \ldots, Z_{n+1})\right)_i \delta_{Z_i}$ and

$$\begin{aligned} w_{n+1,n+1}(z_1, \ldots, z_{n+1})_i &= \frac{\sum_{\sigma \in S_{n+1}:\ \sigma(n+1)=i} \frac{\mathrm{d}Q}{\mathrm{d}P}(z_{\sigma(n+1)})}{\sum_{\sigma \in S_{n+1}} \frac{\mathrm{d}Q}{\mathrm{d}P}(z_{\sigma(n+1)})} \\ &= \frac{n! \frac{\mathrm{d}Q}{\mathrm{d}P}(z_i)}{\sum_{j=1}^{n+1} \sum_{\sigma \in S_{n+1}:\ \sigma(n+1)=j} \frac{\mathrm{d}Q}{\mathrm{d}P}(z_j)} \\ &= \frac{\frac{\mathrm{d}Q}{\mathrm{d}P}(z_i)}{\sum_{j=1}^{n} \frac{\mathrm{d}Q}{\mathrm{d}P}(z_j)}. \end{aligned}$$

□

We now state the main theorem of this subsection which is the foundation of the weighted conformal prediction procedure for distribution shift.

Some notation needs to be defined below.

Theorem 3.2. *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*

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

Define the prediction set

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} :\ S_{n+1}^y \leq \hat{Q}_{S^y}^{w^y}(1-\alpha) \right\}. \tag{3.8}$$

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 (3.7) and we note that for all $i \in [n+1]$, we have $w_i^{Y_{n+1}} = w_i$. We first note that

$$\sum_{i=1}^{n+1} w_i \delta_{S_i} = \frac{\sum_{i=1}^{n+1} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i) \delta_{s(Z_i;D)}}{\sum_{i=1}^{n+1} \frac{\mathrm{d}Q}{\mathrm{d}P}(Z_i)}$$

is invariant under permutations of $(Z_1, \ldots, Z_{n+1})$ since s is symmetric. Therefore, $\hat{Q}_S^w(1-\alpha)$ is a function of $\hat{P}_{n+1}$. 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. □

References

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

[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).

[Tib24a] Ryan Tibshirani. *Advanced Topics in Statistical Learning Homework 4*. https://www.stat.berkeley.edu/~ryantibs/statlearn-s24/homeworks/homework4.pdf. [Online homework sheet]. 2024.

[Tib24b] Ryan Tibshirani. *Advanced Topics in Statistical Learning*. https://www.stat.berkeley.edu/~ryantibs/statlearn-s24/lectures/conformal.pdf. [Online lecture notes]. 2024.

[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.

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

[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).

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