New Advances in Conformal Prediction

April 4, 2025

1 Introduction

1.1 Notation

2 Foundations of Conformal Prediction

The primary appeal of conformal prediction, as originally developed by Vovk et. al. [VGS], lies in the fact that the coverage guarantee holds in a distribution-free sense and in finite samples. Compare this to the construction of prediction sets in classical statistics, which often relies on strong distributional assumptions (such as normality) and may only hold asymptotically.

In this chapter, we develop the mathematical foundations of conformal prediction. We begin by defining the key concepts of quantiles, empirical cumulative distribution functions and exchangeability in Section 2.1. In sections Section 2.2 and Section 2.3, we present the two main variants of conformal prediction: full conformal prediction and split conformal prediction, respectively, and prove their coverage guarantees. In Section 2.4, we consider specific instances of conformal prediction and compare these through numerical experiments.

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 is inspired by the proof of Theorem 1 in [Bar+23].

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 P be a probability distribution on $\mathbb{R}$ with cumulative distribution function F. The *quantile function* of P is defined for $\beta \in (0, 1)$ by

$$Q(P; \beta) := \inf \{z \in \mathbb{R} : F(z) \geq \beta\}.$$

We may also use the notation $Q(F; \beta)$ in place of $Q(P; \beta)$.

Definition 2.3. For $z \in \mathbb{R}^n$, we define the following quantities.

(i) The *empirical cumulative distribution function* of z is the function $F_z : \mathbb{R} \to [0, 1]$ given by

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

Equivalently, $\hat{F}_z$ is the cumulative distribution function of the probability distribution

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{z_i},$$

where δ_a denotes the Dirac measure at a, for any $a \in \mathbb{R}$.

(ii) The *quantile function* of z is defined for $\beta \in (0, 1)$ by

$$\hat{Q}_z(\beta) := \inf \left\{x \in \mathbb{R} : \hat{F}_z(x) \geq \beta\right\}.$$

Equivalently,

$$\hat{Q}_z(\beta) := Q\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}; \beta\right).$$

The following lemma formalises a sense in which the empirical cumulative distribution and quantile functions are inverses of each other.

Lemma 2.1. *Let $z \in \mathbb{R}^n$ and $\beta \in (0, 1)$. Then we have that $\hat{F}_z\left(\hat{Q}_z(\beta)\right) \geq \beta$. If, moreover, the components of z are distinct, then $\hat{F}_z\left(\hat{Q}_z(\beta)\right) = \frac{\lceil n\beta \rceil}{n}$.*

Proof. The first claim follows from the definitions upon noting that $\hat{F}_z$ is right-continuous. For the second claim, we observe that for all $x \in \mathbb{R}$, we have that $\hat{F}_z(x) \in \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}$. 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

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

$\square$

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 $Z := (Z_1, \ldots, Z_n)$. 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. $\qquad \square$

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 kth smallest element of Z, we deduce that the probability that Z_i is the kth 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 2.4. 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

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

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*

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

Then we have that

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

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

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

4

Proof. We begin by showing that $S := (S_1, \ldots, S_{n+1})$ 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

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

$\square$

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. *For any $z \in \mathbb{R}^n$ and $\beta \in (0, 1)$, we have that*

$$y \leq \hat{Q}_{(z,y)}(\beta) \iff y \leq \hat{Q}_z\left(\frac{\lceil (n+1)\beta \rceil}{n}\right)$$

Proof. First note that by Definition 2.3, we have that

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

Therefore, we have that

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

by Lemma 2.1. We deduce that

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

which shows the forward direction of the result. Now suppose $y > \hat{Q}_{(z,y)}(\beta)$, then

$$\begin{aligned}
\hat{F}_z(\hat{Q}_z(\beta)) &= \frac{n+1}{n}\hat{F}_{(z,y)}(\hat{Q}_z(\beta)) - \frac{1}{n}\mathbb{1}\left\{y \leq \hat{Q}_{(z,y)}(\beta)\right\} \\
&= \frac{n+1}{n}\hat{F}_{(z,y)}(\hat{Q}_{(z,y)}(\beta)) \\
&= \frac{n+1}{n}\hat{F}_{(z,y)}\left(\hat{Q}_{(z,y)}\left(\frac{\lceil\beta(n+1)\rceil}{n+1}\right)\right) \\
&\geq \frac{\lceil\beta(n+1)\rceil}{n},
\end{aligned}$$

where the final inequality follows from Lemma 2.1. $\square$

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*

$$C(X_{n+1}) = \left\{y \in \mathcal{Y}: \quad S_{n+1}^y \leq \hat{Q}_{(S_1^y, \ldots, S_n^y)}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}\right)\right\}. \tag{2.2}$$

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

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

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 $C \leftarrow \emptyset$

    for $y \in \mathcal{Y}$ **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 $S^y = (S_1^y, \dots, S_n^y)$.

        Compute $\hat{Q}_{S^y}(1 - \alpha)$ as the $\lceil (1 - \alpha)(n + 1) \rceil$ smallest of $S_1^y, \dots, S_n^y$.

        if $S_{n+1}^y \leq \hat{Q}_{S^y}(1 - \alpha)$ **then**

            $C \leftarrow C \cup \{y\}$.

        end if

    end for

Output: C
___

Remark 2.5. 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 the advantages and disadvantages of one over the other.

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 $D_{\text{cal}} \coloneqq ((X_i, Y_i))_{i=1}^n$. 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 $\hat{S} = (\hat{S}_1, \dots, \hat{S}_n)$.

Theorem 2.3. *Suppose* $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{Z}$ *are exchangeable. Define the prediction set*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \hat{s}(X_{n+1}, y) \leq \hat{Q}_{\hat{S}}\left(\frac{\lceil (1 - \alpha)(n + 1) \rceil}{n} \right) \right\}. \tag{2.3}$$

Then we have that

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

If, moreover, the scores $\hat{S}_1, \dots, \hat{S}_{n+1}$ *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}. $\qquad\square$

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 $C \leftarrow \emptyset$

Compute $\hat{S}_i = \hat{s}(X_i, Y_i)$ for each $i \in [n]$ and set $\hat{S} = (\hat{S}_1, \ldots, \hat{S}_n)$.

for $y \in \mathcal{Y}$ **do**

    Compute $\hat{s}(X_{n+1}, y)$.

    Compute $\hat{Q}_{\hat{S}}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}\right)$ as the $\lceil(1-\alpha)(n+1)\rceil$ smallest of $\hat{S}_1, \ldots, \hat{S}_n$.

    if $\hat{s}(X_{n+1}, y) \leq \hat{Q}_{\hat{S}}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}\right)$ **then**

        $C \leftarrow C \cup \{y\}$.

    end if

end for

Output: C

Remark 2.6. 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 Algorithm 2 to the calibration data to obtain the prediction set.

Remark 2.7. 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 $\hat{s}$ and compute the conformity scores on the calibration data. 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. 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 $\hat{\mu}(X_{n+1})$ since

$$C(X_{n+1}) = \left\{ y \in \mathbb{R} : |y - \hat{\mu}(X_{n+1})| \le \hat{Q}_{\hat{S}}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}\right) \right\}$$

$$= \left[\hat{\mu}(X_{n+1}) - \hat{Q}_{\hat{S}}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}\right), \ \hat{\mu}(X_{n+1}) + \hat{Q}_{\hat{S}}\left(\frac{\lceil(1-\alpha)(n+1)\rceil}{n}\right)\right].$$

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}.

2.4 Choice of conformity score

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`.

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

[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.

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