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 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 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

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

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 $F_z : (-\infty,\infty] \to [0,1]$ given for $x \in (0,\infty)$ by

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

We define $\hat{F}_z(\infty) = 1$.

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

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

Remark 2.3. Note that $\hat{F}_z$ 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 $a \in \mathbb{R}$. We also have that

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

so $\hat{Q}_z$ is the quantile function of the probability distribution $\frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}$.

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 $\hat{F}_z(x) \in \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\}$. 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 2.4. 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 2.5. 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 2.4, 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

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

□

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

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

Proof. First suppose that $y \leq \hat{Q}_{(z,y)}(\beta)$. Note that by Definition 2.3, we have that

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

If $\frac{\lceil \beta(n+1) \rceil}{n} \leq 1$, then it follows 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

$$\hat{Q}_{(z,y)}(\beta) \leq \hat{Q}_z\left(\frac{\lceil \beta(n+1) \rceil}{n}\right). \tag{2.2}$$

Clearly $y \leq \infty$, so this proves the forward direction of the lemma. To prove the converse direction, first note that if $y > \hat{Q}_{(z,y)}(\beta)$, then we have that

$$\begin{aligned} \hat{F}_z(\hat{Q}_{(z,y)}(\beta)) &= \frac{n+1}{n}\hat{F}_{(z,y)}(\hat{Q}_{(z,y)}(\beta)) - \frac{1}{n}\mathbb{1}\left\{y \leq \hat{Q}_{(z,y)}(\beta)\right\} \\ &= \frac{n+1}{n}\hat{F}_{(z,y)}(\hat{Q}_{(z,y)}(\beta)) \\ &= \frac{n+1}{n}\hat{F}_{(z,y)}\left(\hat{Q}_{(z,y)}\left(\frac{\lceil \beta(n+1) \rceil}{n+1}\right)\right) \\ &\geq \frac{\lceil \beta(n+1) \rceil}{n}, \end{aligned} \tag{2.3}$$

where the final inequality follows from Lemma 2.1. If $\frac{\lceil \beta(n+1) \rceil}{n} \leq 1$, then this implies that

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

a contradiction, so $y \leq \hat{Q}_{(z,y)}(\beta)$. If $\frac{\lceil \beta(n+1) \rceil}{n} > 1$, then it implies $\hat{F}_z(\hat{Q}_{(z,y)}(\beta)) > 1$, a contradiction, so $y \leq \hat{Q}_{(z,y)}(\beta)$. □

Remark 2.6. Note that for any $z \in \mathbb{R}^n$ and $\beta \in (0,1]$,

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

This follows from (2.2) and (2.3) upon noting that $\hat{Q}_{(z,\infty)}(\beta) < \infty$ if $\beta \leq \frac{n}{n+1}$, which is equivalent to $\frac{\lceil \beta(n+1) \rceil}{n} \leq 1$.

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*

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \quad S^y_{n+1} \leq \hat{Q}_{(S^y_1, \ldots, S^y_n, \infty)}(1-\alpha) \right\}. \tag{2.4}$$

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 ← ∅
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 $\hat{S}^y_{n+1} = \hat{s}(X_{n+1}, y)$ for any $y \in \mathcal{Y}$.

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

$$C(X_{n+1}) = \left\{ y \in \mathcal{Y} : \ \hat{S}^y_{n+1} \leq \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1-\alpha) \right\}. \tag{2.5}$$

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

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

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

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

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

$$\begin{aligned} C(x) &= \left\{ y \in \mathbb{R} : \, |y - \hat{\mu}(x)| \leq \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha) \right\} \\ &= \left[\hat{\mu}(X_{n+1}) - \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha), \; \hat{\mu}(x) + \hat{Q}_{(\hat{S}_1, \ldots, \hat{S}_n, \infty)}(1 - \alpha) \right]. \end{aligned} \tag{2.6}$$

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 (2.6). 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.

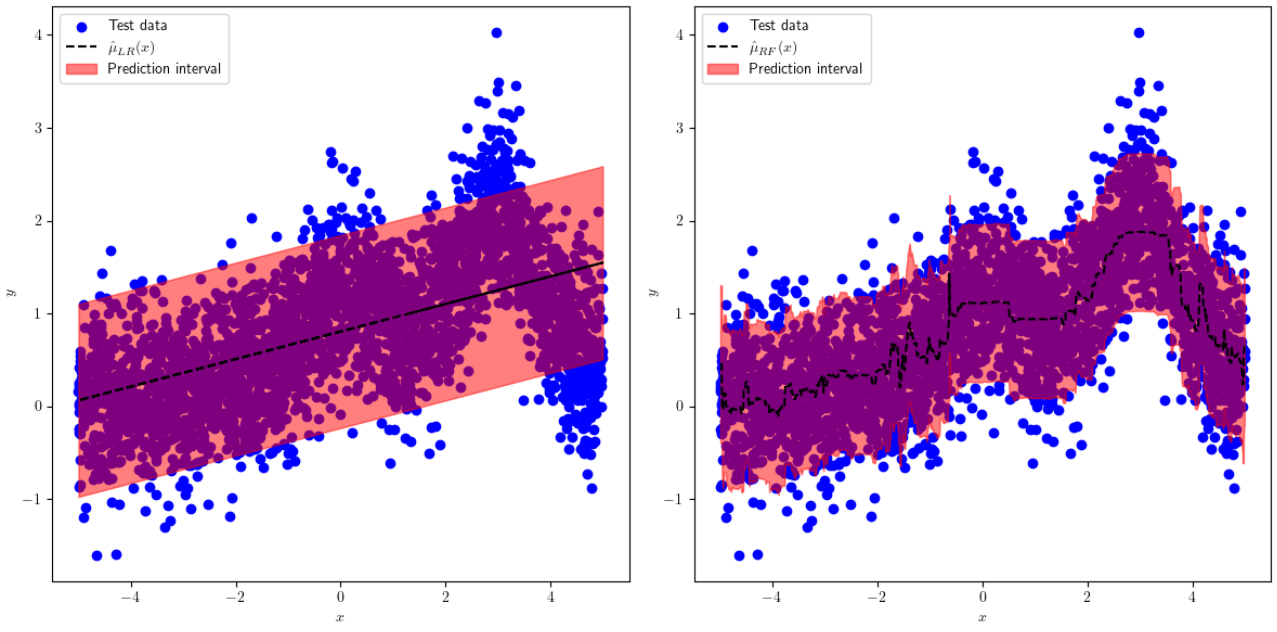


Figure 1: Plot of the test data, the fitted regression function and the split conformal prediction interval using linear regression (left) and random forests (right). The fitted regression function is denoted $\hat{\mu}_{LR}$ (left) and $\hat{\mu}_{RF}$ (right).

2.4 Choice of conformity score

In Section 2.3, we showed that any conformity score can be used to construct the split conformal prediction set (2.6). However, it is not immediately clear from (2.6) 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 (2.6). 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 (2.6) 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 (2.6) 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 quantiles directly. This is referred to as *quantile regression*. In this essay, we do not discuss the numerous methods for performing quantile regression. However, we note that they rely on the following fact.

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

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

2.4.4 Numerical Experiments

In this section, we 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 $i \in [n]$.

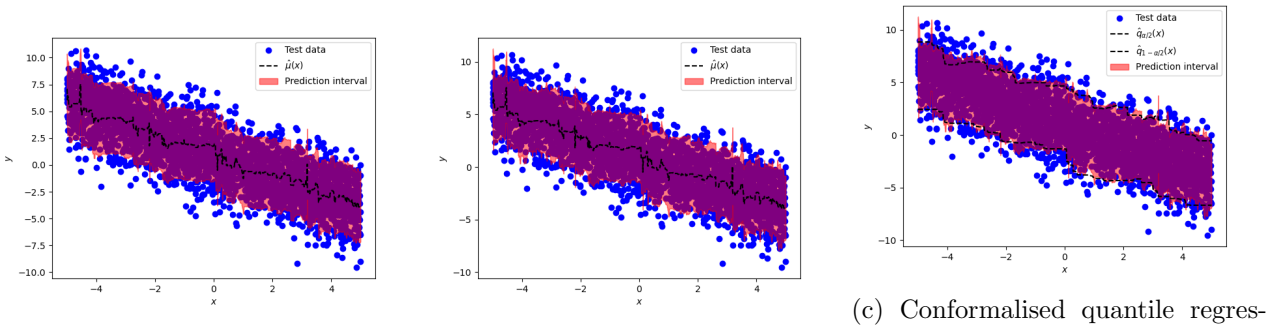
In each setting, we generate independent proper training, calibration and test datasets with sizes 1000, 3000 and 3000, respectively, set $\alpha = 0.1$ and generate conformal prediction intervals for the test dataset using the absolute residual score, the locally weighted residual score and conformalised 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 $\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})|X_{n+1}\right)$ 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.

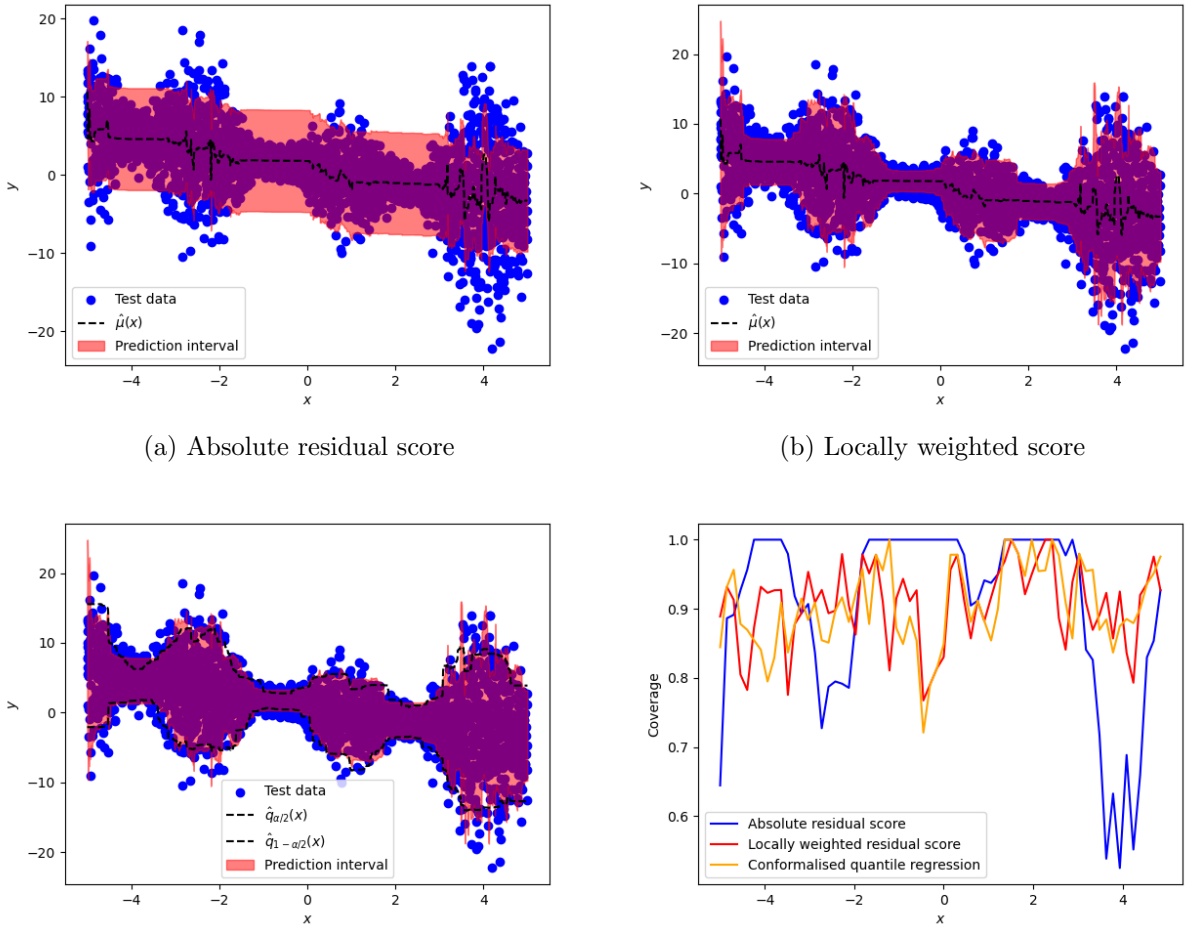
	Absolute residual		Locally weighted		Conformalised quantile regression	
	Coverage	Average length	Coverage	Average length	Coverage	Average length
Setting 1	0.9010	6.616	0.9027	6.632	0.9017	6.712
Setting 2	0.9003	13.03	0.9063	11.58	0.9027	11.55

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

3.1 Covariate Shift

3.2 Nonexchangeable Conformal Prediction

In this subsection, we present the *nonexchangeable conformal prediction* (NexCP) method developed in [Bar+23]. 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.

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).

[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.

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