Split Conformal Prediction for Regression
Tutorial Notes

Mikołaj Mazurczyk & Christian Igel
{mima,igel}@di.ku.dk
University of Copenhagen

August 30, 2025

Abstract

This tutorial introduces Conformal Prediction (CP) for quantifying the uncertainty of machine learning (ML) models. Conformal Prediction provides prediction sets guaranteed to contain (in expectation) the true outcome with a user-specified probability without strong assumptions about the underlying data distribution. We focus on regression tasks and on Split CP, which leverages a hold-out calibration data set and can be applied to any ML model. We establish the foundational theory of Split CP, proving its marginal coverage guarantee under the assumption of i.i.d. data. Then we present Split Localized Conformal Prediction, an efficient method that approximates conditional coverage by adapting to the local data structure while preserving the rigorous marginal guarantee. The goal of this tutorial is to provide attendees with the basic foundations needed to apply CP in their research and to explore more advanced topics in distribution-free uncertainty quantification.

Please inform us about errors and send us suggestions how to improve the text.

Contents

Notation

- X, Y, x, y — random variables and their realizations are denoted by uppercase and lowercase letters, respectively.

- $\mathcal{X}$ — the feature/input space.

- $\mathcal{Y}$ — the label/output space.

- $2^{\mathcal{Y}}$ — power set of $\mathcal{Y}$, meaning a set of all possible subsets of $\mathcal{Y}$.

- $\{z_i\}_{i=1}^n$ — multiset of n elements $z_i, \ldots, z_n$ (collection of elements where duplicates can occur).

- $F_X(x)$ — cumulative distribution function (CDF) of a real-valued variable X, $F_X(x) = \mathbb{P}(X \leq x)$.

- $\widehat{F}_X(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i \leq z\}$ — empirical CDF computed with realizations $\{x_i\}_{i=1}^n$ of a random variable X.

- $\mathcal{Q}_\tau(F_X)$ — τ-quantile of a CDF F_X defined as

$$\mathcal{Q}_\tau(F_X) = \inf\{x : \tau \leq F_X(x)\}.$$

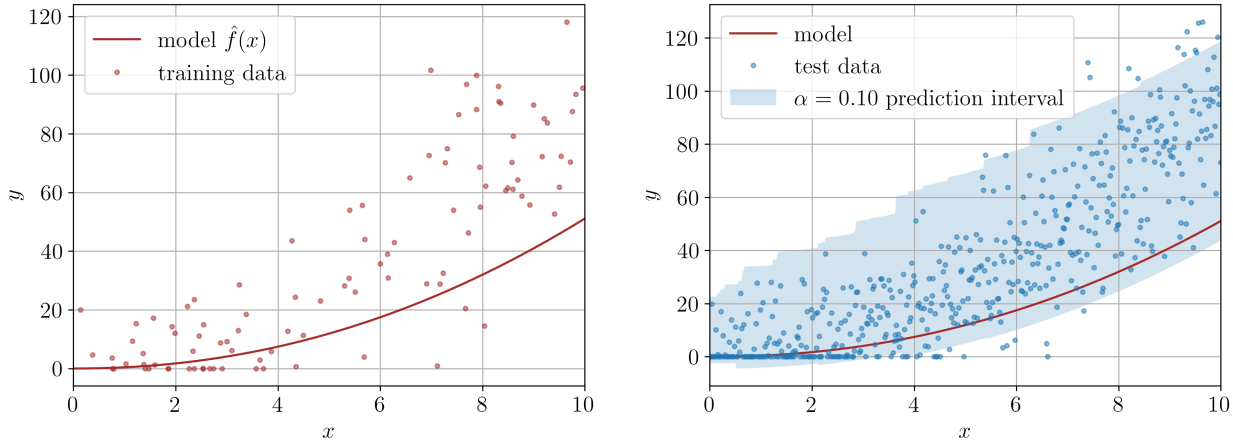

Figure 1.1: Left: Training data and a model $\hat{f}(x) = ax^b$ fitted to theses data after logarithmic transformation of the data. The data were generated from an allometric (power) equation plus Gaussian noise with variance increasing with x; the observations are clipped to remain non-negative. The logarithmic transformation is not uncommon when fitting allometric equations, but not recommended [Zar68] and just used here to have example with an asymmetric error distribution. Right: Prediction interval computes using conformal prediction.

- $\widehat{\mathcal{Q}}_\tau\left(\{z_i\}_{i=1}^n\right)$ — empirical τ-quantile of a collection $\{z_i\}_{i=1}^n$ of size n using the nearest-rank definition is given by

$$\widehat{\mathcal{Q}}_\tau\left(\{z_i\}_{i=1}^n\right) = \inf\left\{z' : \tau \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{z_i \leq z'\right\}\right\},$$

  which is the $\lceil \tau \cdot n \rceil$-th value in the list of $\{z_i\}_{i=1}^n$ values sorted in increasing order.

1. Introduction

Given a statistical model $\hat{f} : x \mapsto y$ learned from data, we want to quantify the probability that the true value y for an input x is within a certain set $\mathcal{C}(x)$ with a high probability. Consider regression with $\mathcal{Y} = \mathbb{R}$. Then one is typically interested in prediction intervals $\mathcal{C}(x) = [\hat{f}(x) - c_{\mathrm{lo}}, \hat{f}(x) + c_{\mathrm{hi}}] \subset \mathbb{R}$ with nonnegative scalars c_{lo} and c_{hi}. Figure 1.1 depicts an example, where a non-linear model fitted to data is shown in the left plot, and corresponding prediction intervals are visualized in the right plot.

There are many approaches that estimate prediction sets. If we had employed standard linear regression using optimal least squares, textbooks (and software packages) would suggest ways to analytically compute prediction intervals. If we had fitted a Gaussian process (GPs) from Bayesian machine learning, it would naturally come with uncertainty estimates for the predictions. However, do these approaches give us rigorous guarantees in practice, that is, can we guarantee that the predictions will indeed be in the prediction set $\mathcal{C}(x)$ with a probability of $(1 - \alpha) \in (0, 1]$?

Let us assume that our training data are representative for the problem at hand and i.i.d., that is, the observations can be viewed as being drawn independently of each other from the same underlying (unknown) distribution. If the relationship between the independent variables x and the dependent variable y is indeed affine linear, the residuals (i.e., the errors the model makes) are statistically independent, and the residuals are normally distributed with a variance that is constant across all values of x, then least-squares linear regression leads to proper prediction intervals. For the GP prediction intervals to be correct, the GP prior (i.e., the mean function and the kernel function) must reflect the true underlying function's properties and it is typically assumed that the noise on the observations is additive Gaussian and i.i.d. [Pap24]. However, if these strong assumptions are violated, then the uncertainty estimates these approaches provide are not calibrated in the sense that we do not have (probabilistic) guarantees on the prediction sets. While much research has been conducted on loosening the assumptions and dealing with their violations both for least-squares linear regression and GPs, we would argue that fundamental assumptions are generally violated in practice. The statistical models are almost always misspecified. Is the relationship between input and output really linear? Does the GP kernel function really completely and exactly describe the covariance between points? Or, more generally, is the prior really correct or partly chosen because of mathematical convenience? In most applications, the answer is no. The Conformal Prediction (CP, [VGS99]) methods discussed in this tutorial provide a remedy. They allow us to compute prediction sets with formal guarantees without assumptions on the model. However, you need not set aside your favorite model and uncertainty estimation heuristic: Conformal prediction can be used to calibrate uncertainty estimates provided by your model, for example, to adjust prediction intervals from GPs.

In machine learning, there is a standard way to estimate the accuracy of a predictive model. To estimate the expected accuracy of a model, we estimate its accuracy on i.i.d. validation not used during training. In expectation over all draws of the validation data set, the mean error on the validation data set equals the expected error; and to account for finite sample effects, we apply finite sample concentration bounds. One can view the CP methods presented in this tutorial as adopting this approach to estimating the uncertainty of a model:

- To estimate the uncertainty of a model, we estimate its uncertainty on an i.i.d. calibration data set not used during training.

- In expectation over all draws of the calibration data set, the α-quantile of uncertainties on the calibration data should be the expected α-quantile of the uncertainty for an independent data point drawn from the same distribution.

- To account for finite-sample effects, we apply finite sample concentration bounds.

Split conformal prediction (SCP) is a variant of CP that requires an additional calibration data set. This is a drawback because in many applications labeled data is scarce. However, *uncertainty quantification does not come for free. You have to pay for it, either in the form* *of strong assumptions on the model or by using data.* We focus on SCP using a hold-out calibration data set $\mathcal{D}_{\text{cal}}$ [BCRT21, ABB25]. You can apply CP in a cross-validation setting to use the available data more efficiently. However, this requires training of several models on different subsets of the data, which can be computationally expensive. Here, we focus on a setting common in deep learning, where data is plentiful but training models is very costly.

Conformal prediction goes back to Vovk and Gammerman (e.g., [VGS99]), there is an excellent review by Angelopoulos and Bates [AB23], and a new textbook draft by Angelopoulos, Barber, and Bates [ABB25]. This tutorial is based on [HTGL23, AB23, ABB25], most concepts and proofs are taken from [ABB25].

2. Split Conformal Prediction

This tutorial focuses on SCP applied to regression tasks with $\mathcal{Y} = \mathbb{R}$, however, most of the results are derived for the general case.

2.1. Basic idea

Let us first formulate the problem. Given a

- distribution P over $\mathcal{X} \times \mathcal{Y}$,
- fitted model $\hat{f} : \mathcal{X} \to \mathcal{Y}$,
- miscoverage level $\alpha \in \mathbb{R}$,
- calibration data set $\mathcal{D}_{\text{cal}} = \{X_i, Y_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P$,

we want to construct a function $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ that generates a prediction interval $\mathcal{C}\left(X_{\text{test}}\right)$ such that

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) \geq 1 - \alpha \tag{2.1}$$

for any test sample $(X_{\text{test}}, Y_{\text{test}}) \sim P$ independent of $\mathcal{D}_{\text{cal}}$.

There is an important difference between *(marginal) coverage* defined in (2.1) and *conditional coverage*:

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid X_{\text{test}}\right) \geq 1 - \alpha \tag{2.2}$$

Marginal coverage is a much weaker property. For example, assume that the model is related to a medical problem in the context of a rare disease and that the prediction task is simple for healthy persons. If the prevalence of the disease is α, a model that works perfectly for healthy people and not at all for diseased people meets the marginal coverage criterion at miscoverage level α. From a patient's perspective, conditional coverage is desired; the patient wants to know the uncertainty of the method given their condition, not the uncertainty overall. Unfortunately, conditional coverage is difficult to achieve, in the continuous case in general impossible as we will see later. We will discuss CP and conditional coverage section 5.

To get an intuition of how SCP works, let us consider the following simple algorithm:

1. For each of n points $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$ compute a score function, say, the absolute error $s_i = s(x_i, y_i) = \left| y_i - \hat{f}\left(x_i\right) \right|$.
2. Sort $s_1, \ldots, s_n$ in increasing order and create an empirical cumulative distribution function (CDF) of errors.
3. Pick an empirical quantile $\hat{q}$ corresponding to the $(1-\alpha)$ quantile of the distribution.
4. Construct a prediction interval:

$$\mathcal{C}\left(x_{\text{test}}\right) = \{y \in \mathcal{Y} \mid s(x_{\text{test}}, y) \leq \hat{q}\} = \left[\hat{f}\left(x_{\text{test}}\right) - \hat{q}, \hat{f}\left(x_{\text{test}}\right) + \hat{q}\right]$$

Intuitively, since calibration samples and the test points are i.i.d., there is equal chance that the model will produce a test error similar to any other calibration sample, meaning it will approximately follow the empirical distribution, see Figure 2.1. Now, by construction, this means that there is a $(1-\alpha)$ chance that the test error would fall below $\hat{q}$ in our empirical CDF, so we have $1-\alpha$ probability that it will be within the error margin of $\pm\hat{q}$.

Continuing the example from Figure 1.1, the left plot in Figure 2.1 visualizes the scores computed in step 1 and the right plot depicts the corresponding empirical CDF considered in step 2. To make step 3 more concrete, we need the concept of the empirical τ-quantile of a collection $\{z_i\}_{i=1}^n$. Using the nearest-rank definition, it is given by

$$\widehat{\mathcal{Q}}_\tau\left(\{z_i\}_{i=1}^n\right) = \inf\left\{z' : \tau \leq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\left\{z_i \leq z'\right\}\right\},$$

which is the $\lceil \tau \cdot n \rceil$-th value in the list of $\{z_i\}_{i=1}^n$ values sorted in increasing order (e.g., $\widehat{\mathcal{Q}}_{0.5}\left(\{1,2,3,4,5\}\right) = 3$; $\widehat{\mathcal{Q}}_{0.5}\left(\{1,2,3,4\}\right) = 2$; $\widehat{\mathcal{Q}}_{0.25}\left(\{1,2,3,\ldots,100\}\right) = 25$). The empirical quantile that we choose in the above algorithm is $\hat{q} := \widehat{\mathcal{Q}}_{(1-\alpha)((n+1)/n)}\left(\{s_i\}_{i=1}^n\right)$, which is the k-th value when sorting $\{s_i\}_{i=1}^n$ in increasing order with $k := \lceil (1-\alpha)(n+1) \rceil$. The value of k can be higher than $\lceil (1-\alpha) n \rceil$, that is, $\widehat{\mathcal{Q}}_{(1-\alpha)}\left(\{s_i\}_{i=1}^n\right) \leq \widehat{\mathcal{Q}}_{(1-\alpha)((n+1)/n)}\left(\{s_i\}_{i=1}^n\right)$. This choice allows us to prove the main marginal coverage guarantee as stated in Theorem 1 below using the proof in section 3. The plot on the left in Figure 2.2 shows a prediction interval generated by the procedure for the example introduced in Figure 1.1.

The procedure is valid no matter the choice of $\hat{f}$, but of course the results depend of how well $\hat{f}$ fits the data. The data distribution does not affect the validity of the confidence set, as long as our calibration and test samples are i.i.d. (actually, they only need to be exchangeable, see definition 5 in Appendix B). However, the quality of the results, of course, depends on the distribution of the data – and on the choice of the score function. The results in the left plot of Figure 2.2 do not look very good. The reason is that the noise is heteroscedastic in our example from Figure 1.1, that is, it depends on the input x. Furthermore, the errors are not symmetric in the sense that the model is more under- than overestimating. The simple procedure given above produces prediction intervals that are centered on the model's prediction and have the same width independent of the input x, see left plot of Figure 2.2.

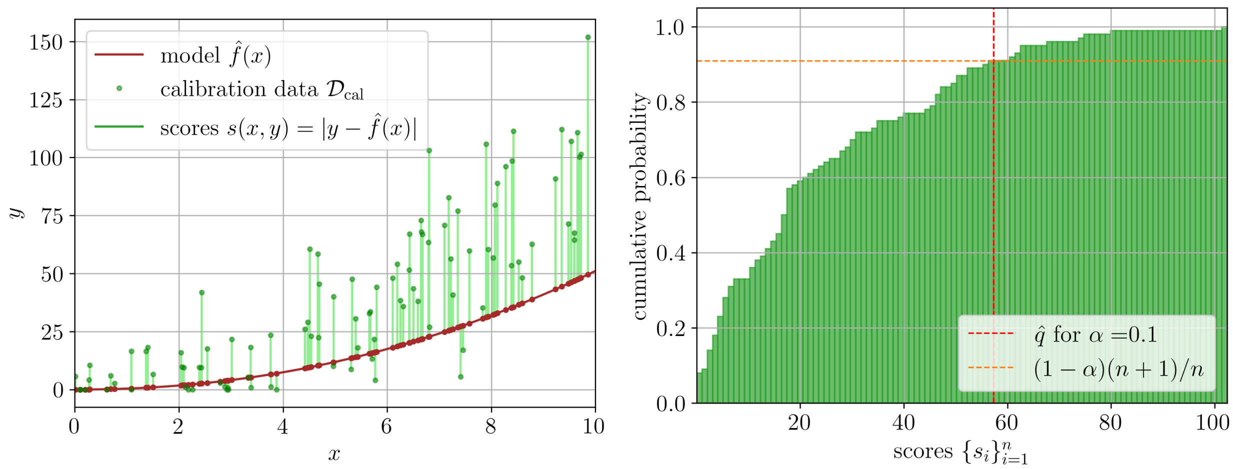

Figure 2.1: Score function and its empirical cumulative distribution. Left: Visualization of the scores $s_i = \left| y_i - \hat{f}\left(x_i\right) \right|$ for $(x_i, y_i) \in \mathcal{D}_{\mathrm{cal}}$. Right: Empirical cumulative distribution function of the non-conformity scores.

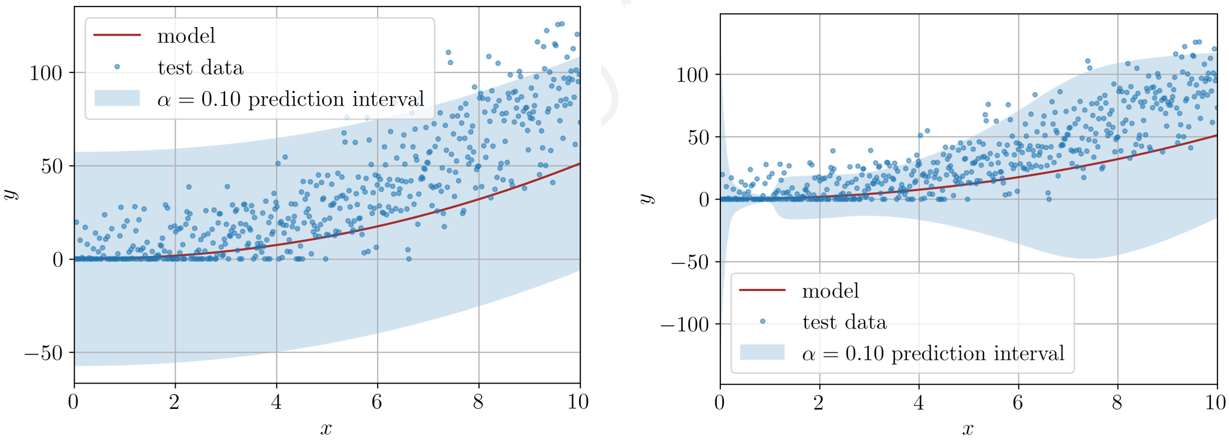

Figure 2.2: SCP using different score functions applied to the task introduced in Figure 1.1. Left: Absolute error. Right: Scaled absolute error (using a simple neural network to predict the logarithmic variance optimizing the likelihood as suggested by [NW94]).

However, these properties can be changed by changing the score function, as discussed in detail in section 4.

For now, let us have a look at how SCP is typically used – as a method to recalibrate given uncertainty estimates, rather than to create intervals from scratch. We start with the same problem formulation as before, but now we assume that the model $\hat{f}$ also provides an uncertainty estimate $u(x) > 0$ for its prediction given inut x. Uncertainty estimates could come from variances predicted by the model (e.g., computed following [NW94]), ensemble variances (as suggested as a measure for uncertainty already by [HS90]), variances from Bayesian modeling (for GPs this is discussed in [Pap24]), etc. Then we can define $s_i = \left| y_i - \hat{f}(x_i) \right| / u(x_i)$, similar as before, but now also rescaling the error by u. Repeating the procedure of obtaining $\hat{q}$ leads to valid prediction sets of the form

$$\mathcal{C}(x) = \left[\hat{f}(x) - \hat{q} \cdot u(x), \hat{f}(x) + \hat{q} \cdot u(x) \right].$$

Essentially here we are rescaling all u by a constant factor, so that we obtain the desired coverage, allowing to model non-homoscedastic noise thanks to dependence of $u(x)$ on x. The plot on the right in Figure 2.2 shows the prediction interval when using (based on some a priori expert knowledge) $u(x) \coloneqq x$. As desired, the widths of the prediction intervals now depend on x, and in fact the average width of the prediction intervals is considerably lower compared to simply using the absolute error. However, the intervals are still centered around the prediction. This is suboptimal in our example, and we will see how this can be addressed in section 4.

Perhaps it is less obvious that this way of computing s_i leads to prediction sets with $1-\alpha$ coverage guarantee, but we will prove an even more general result in the following. So far, we have seen that we can adapt to the data distribution with the appropriate choice of the model and the way of computing s_i. In the following, we provide a proper definition of SCP, generalizing the process of computing s_i and $\mathcal{C}$.

2.2. The SCP algorithm

When defining SCP we generalize the absolute errors to the non-conformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which can be any function that fulfills two crucial requirements. First, higher scores $s(x, y)$ should encode worse agreement between prediction $\hat{y}$ based on x and the ground-truth y, as they are used to generate the quantile $\hat{q}$ in a similar way as in our intuitive example, and thus the order matters. Second, the score function must be independent of $\mathcal{D}_{\text{cal}}$ so that the test and calibration scores are i.i.d. Again, in our intuitive example we required the data to be i.i.d. so that the errors on $\mathcal{D}_{\text{cal}}$ and test samples are i.i.d., which motivates this requirement.

When applying SCP to Machine Learning (ML) tasks, typically the training dataset is used to construct s, in particular, s utilizes a pre-trained model $\hat{f}$. Thus, hereafter, we combine these assumptions on s by calling it the pre-trained non-conformity score function $s(x, y)$. It is called pre-trained as it uses the pre-trained model $\hat{f}$, and the term non-conformity score indicates that the higher the score, the less y conforms to [agrees with] the estimate $\hat{y} \coloneqq \hat{f}(x)$). Now, we are ready to introduce the SCP method in Algorithm 1.

Algorithm 1: Split Conformal Prediction

```
Input: non-conformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; calibration set
       $\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i=1}^{n}$ of size $n \in \mathbb{Z}^+$; test point $X_{\text{test}}$ such that
       $\forall_{i \in \{1,\dots,n\}} (x_i, y_i), (x_{\text{test}}, y_{\text{test}}) \in \mathcal{X} \times \mathcal{Y}$; target miscoverage level $\alpha \in (0,1)$
Output: prediction set $\mathcal{C}(x_{\text{test}})$ with coverage $\geq 1 - \alpha$, where $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$
for $i = 1$ to $n$ do  $s_i \leftarrow s(x_i, y_i)$          // Compute scores on calibration set
order scores $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(n)}$          // Calculate conformal quantile
set rank $k \leftarrow \lceil (1-\alpha)(n+1) \rceil$
if $k > n$ then  set threshold $\hat{q} \leftarrow \infty$                    // Handle edge case
else  set threshold $\hat{q} \leftarrow s_{(k)}$
return $\mathcal{C}(x_{\text{test}}) \leftarrow \{y \in \mathcal{Y} \mid s(X_{\text{test}}, y) \leq \hat{q}\}$          // Construct prediction set
```

In the pseudo-code, we use the notion $s_{(1)}, \dots, s_{(n)}$ for elements $\{s_i\}_{i=1}^{n}$ sorted in increasing order. Algorithm 1 provides a prediction set with the following fundamental marginal coverage property.

Theorem 1 (Marginal coverage guarantee; [VGS99]). *Assume that $\mathcal{C}$ is defined as in Algorithm (1) for $\alpha \in (0,1)$, $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^{n}$, and $\forall_{i \in \{1,\dots,n\}} (X_i, Y_i), (X_{test}, Y_{test}) \overset{i.i.d.}{\sim} P$, where P is some data distribution. Then, the following holds*

$$1 - \alpha \leq \mathbb{E}_{\mathcal{D}_{cal}} \left[\mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right) \mid \mathcal{D}_{cal} \right) \right] \leq 1 - \alpha + \frac{1}{n+1} + \epsilon_{tie},$$

where ϵ_{tie} captures the likelihood that the non-conformity score of the test data point S_{test} is the same as another calibration score S_i,

$$\mathbb{P}\left(\exists_{i \in \{1,\dots,n\}} \; S_i = S_{test} \right).$$

Before presenting the proof in section 3, we make some remarks and observations. First, the conformal coverage guarantee only holds in expectation over the calibration and test samples. Hence the name *marginal* coverage guarantee, as we marginalize over $\mathcal{D}_{\text{cal}}$ and $(X_{\text{test}}, Y_{\text{test}})$. Later in the text, we will consider how the limited data setting (i.e., conditioning on $\mathcal{D}_{\text{cal}}$) and considering specific test samples (i.e., conditioning on X_{test} and/or Y_{test}) influence this guarantee.

The second takeaway is that when the calibration dataset is large and the probability is low, the guarantee is tight, meaning $\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) \approx 1 - \alpha$. How realistic are these requirements? In the next section, we will consider how the size of $\mathcal{D}_{\text{cal}}$ influences the coverage guarantee in limited data setting, but the rough guideline is to have $n \geq 1000$ [AB23], making the $1/(n+1)$ term negligible. Then the value of ϵ_{tie} depends heavily on the data and thus also score function's distribution. For example, if data and scores are continuous $\epsilon_{\text{tie}} = 0$ (although in reality computations are done with limited precision, so almost always $\epsilon_{\text{tie}} \geq 0$), while if we have categorical variables or scores the tie probability grows. However, even when the target variable is categorical we can still have a continuous score function.

However, assuming that the gap is small, marginal coverage guarantee provides a crucial insight: prediction sets produced by SCP are as tight as possible, given a score function. For example, in the case we studied previously, a poor fit of $\hat{f}$ led to wide prediction intervals. However, the result in Theorem 1 tells us that the prediction intervals are not wider than is needed to compensate for the errors in $\hat{f}$.

We note that the result of Theorem 1 also holds under exchangeability of the data, a weaker assumption than i.i.d. (see definition 5 in Appendix B). However, several other results in these notes require i.i.d., thus we stick to that assumption. For the proof using exchangeability we refer to Chapter 3 of [ABB25].

3. Proof of the Split Conformal Prediction Marginal Coverage Guarantee

This section presents (which can be safely skipped) the proof of Theorem 1 using steps similar to [ABB25]. We start by deriving two lemmata used in the main proof. The notation for the ordered scores in Algorithm 1 is drawn from the standard conventions of *order statistics*. Let us formally define it:

Definition 1 (Order statistics of a finite list). *For a collection of n random variables $Z_1, \dots, Z_n$ the k-th order statistic, denoted by $Z_{(k)}$, is the random variable corresponding the k-th value when these random variables are arranged in non-decreasing (ascending) order. This implies*

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(k)} \leq \cdots \leq Z_{(n)}.$$

We have the following basic property:

Lemma 1. *For a list of n i.i.d. random variables $\mathcal{Z} = (Z_1, \dots, Z_n)$ and any $k \in \{1, \dots, n\}$ we have*

$$\mathbb{P}\left(Z_i \leq Z_{(k)}\right) \geq \frac{k}{n} \quad \textit{and} \quad \mathbb{P}\left(Z_i < Z_{(k)}\right) \leq \frac{k-1}{n}.$$

Proof: Starting from the definition of order statistics we have

$$k \leq \sum_{i=1}^{n} \mathbb{1}\left\{Z_i \leq Z_{(k)}\right\}$$

and because expectation preserves inequalities we can rewrite

$$\begin{aligned} k &\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\left\{Z_i \leq Z_{(k)}\right\}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{Z_i \leq Z_{(k)}\right\}\right] \\ &= \sum_{i=1}^{n} \mathbb{P}\left(Z_i \leq Z_{(k)}\right) \\ &= n \cdot \mathbb{P}\left(Z_n \leq Z_{(k)}\right) \text{ for any } i, \end{aligned}$$

where first equality is due to linearity of expectation and third follows by the fact that each Z_i is i.i.d. Dividing both sides by n leads to the first inequality in Lemma 1, and the proof of the second follows equivalent steps. ■

Now we introduce the replacement lemma, which shows that, from the point of view of order statistics, comparing the test sample with the samples in the calibration dataset is equivalent to comparing it with the union of the samples:

Lemma 2 (Replacement lemma). *Let $S_{(n;k)}$ denote the k-th order statistic computed over $\{S_i\}_{i=1}^n$, and $S_{(n+1;k)}$ be the k-th order statistic computed over $\{S_i\}_{i=1}^n \cup \{S_{test}\}$. Then we have*

$$S_{test} \leq S_{(n;k)} \iff S_{test} \leq S_{(n+1;k)}.$$

Proof:

1. $S_{\text{test}} \leq S_{(n;k)} \implies S_{\text{test}} \leq S_{(n+1;k)}$: We show it this proving the contrapositive $((A \implies B) \iff (\neg B \implies \neg A))$. Suppose that $S_{\text{test}} > S_{(n+1;k)}$. In this case, we must have $S_{(n+1;k)} = S_{(n;k)}$, because if we add S_{test} and it is strictly greater than $S_{(n+1;k)}$ its addition did not change the order of the first k lowest values. Therefore $S_{\text{test}} > S_{(n+1;k)} \implies S_{\text{test}} > S_{(n;k)}$

2. $S_{\text{test}} \leq S_{(n+1;k)} \implies S_{\text{test}} \leq S_{(n;k)}$: We must have $S_{(n+1;k)} \leq S_{(n;k)}$ because the k-th smallest entry in the list cannot increase if we add a new value to the list. ■

Now we are in the position to prove the main result.

Proof of Theorem 1: The empirical quantile $\hat{Q} := \widehat{\mathcal{Q}}_{(1-\alpha)((n+1)/n)}\left(\{S_i\}_{i=1}^n\right)$ is k-th smallest score S_i of the n samples in $\mathcal{D}_{\text{cal}}$, where $k := \lceil (1-\alpha)(n+1) \rceil$ ($\hat{Q}$ is now capitalized since we treat it as a random variable). For $\alpha < \frac{1}{n+1}$ we get an undefined $\widehat{\mathcal{Q}}$ as $(1-\alpha)\left((n+1)/n\right) > 1 \iff k > n$, so in that edge case we set $\hat{q} = \infty$, resulting in $\mathcal{C}\left(X_{\text{test}}\right) = \mathcal{Y}$, which trivially satisfies marginal coverage.

Otherwise, by definition, we have that $\hat{Q} = S_{(k)}$, allowing us to reformulate the event in the conformal coverage guarantee:

$$\left\{Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right\} = \left\{s\left(X_{\text{test}}, Y_{\text{test}}\right) \leq \hat{Q}\right\} = \left\{S_{\text{test}} \leq S_{(k)}\right\}, \tag{3.1}$$

where $S_{\text{test}} := s\left(X_{\text{test}}, Y_{\text{test}}\right)$. Lemma 2 allows us to expand Eq. (3.1) into

$$\left\{Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right\} = \left\{S_{\text{test}} \le \hat{Q}\right\} = \left\{S_{\text{test}} \le S_{(n;k)}\right\} = \left\{S_{\text{test}} \le S_{(n+1;k)}\right\}. \tag{3.2}$$

We start with the lower bound in Theorem 1. Knowing that the non-conformity score function s is a fixed function from the perspective of $\mathcal{D}_{\text{cal}}$ and the test samples, the fact that the samples from $\mathcal{D}_{\text{cal}}$ and $\left(X_{\text{test}}, Y_{\text{test}}\right)$ are i.i.d. implies that the scores S_i and S_{test} are i.i.d.. Thus, combining Eq. (3.2) with Lemma 1 and using $k = \lceil (1-\alpha)(n+1) \rceil$ results in

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) = \mathbb{P}\left(S_{\text{test}} \le S_{(n+1;k)}\right) \ge \frac{k}{n+1} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \ge 1-\alpha,$$

where we used simplified notation $\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) := \mathbb{E}_{\mathcal{D}_{\text{cal}}}\left[\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right)\right]$.

To prove the upper bound, we first decouple the upper bound on S_{test} and the tie probability, noting that

$$S_{\text{test}} \le S_{(n+1;k)} \iff \text{either } S_{\text{test}} < S_{(n+1;k+1)} \text{ or } S_{\text{test}} = S_{(n+1;k)} = S_{(n+1;k+1)}$$

as shonw formally in Appendix D. From that we have

$$\begin{aligned}
\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right)\right) &= \mathbb{P}\left(S_{\text{test}} \le S_{(n+1;k)}\right) \\
&= \mathbb{P}\left(S_{\text{test}} < S_{(n+1;k+1)}\right) + \mathbb{P}\left(S_{\text{test}} = S_{(n+1;k)} = S_{(n+1;k+1)}\right) \\
&\le \mathbb{P}\left(S_{\text{test}} < S_{(n+1;k+1)}\right) + \epsilon_{\text{tie}} \\
&\le \frac{(k+1)-1}{n+1} + \epsilon_{\text{tie}} \\
&= \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} + \epsilon_{\text{tie}} \\
&\le 1 - \alpha + \frac{1}{n+1} + \epsilon_{\text{tie}},
\end{aligned}$$

where the second equality holds since the two events are mutually exclusive, and the second inequality is by Lemma 1. ■

4. Examples of score functions

With SCP derived using a general score function $s(x, y)$. Now we present common and useful examples of score functions that demonstrate the versatility of the framework and highlight the importance of choosing the appropriate score function for each task. We consider the regression setting, where $\mathcal{Y} \subseteq \mathbb{R}$. The appropriate choice of the score function depends on the distribution of the error residuals. The distribution of the residuals (i.e., the errors of the model) can be categorized as

- homoscedastic, not dependent on input x, or

Table 4.1: Common score functions for regression. The column *heteroscedastic* indicates whether the width of prediction interval depends on the input x. The column *asymmetric* indicates whether the score function can result in prediction intervals where the prediction $\hat{f}\left(x\right)$ is not in the center. Furthermore, some scores require additional uncertainty estimates to be computed.

Score function	heteroscedastic	asymmetric	only requires $\hat{f}\left(x\right)$
Absolute error score	no	no	yes
Scaled absolute error score	yes	no	no
Conformalized Quantile Regression	yes	yes	no
Signed-error Split Conformal Regression	no	yes	yes

- heteroscedastic, dependent on input x.

If $\mathcal{Y} = \mathbb{R}$ and the prediction set has the form $[\hat{f}(x) - c_{\mathrm{lo}}, \hat{f}(x) + c_{\mathrm{hi}}]$ then we distinguish (extends canonically to multi-dimensional outputs)

- symmetric uncertainties, where the upper and lower confidence bounds are the same, and

- aymmetric uncertainties, where upper and lower confidence bounds can differ.

Conformal prediction can be tailored to these settings by changing the score function. In the following, we present some common score functions. Basic properties of these functions are summarized in Table 4.1

For completeness, we start with the score functions we have seen before, the absolute error score and the scaled absolute error score. Both provide symmetric uncertainties, but the latter can model heteroscedastic residuals. Applied to our example problem, they produce prediction intervals as shown in Figure 2.2.

The the other SCP procedures introduced in this section are Conformalized Quantile Regression, which corresponds to just using a different score function, and Signed-error Split Conformal Regression, which additionally requires a minor modification of the SCP algorithm. Their performance on our test problem are visualized in Figure 4.1.

4.1. Absolute error score

Our initial setting $s\left(x, y\right) \coloneqq |y - \hat{f}\left(x\right)|$ results in

$$\mathcal{C}\left(x\right) = \left\{ y \in \mathbb{R} \;\middle|\; |y - \hat{f}\left(x\right)| \leq \hat{q} \right\} \iff y \in \left[\hat{f}\left(x\right) - \hat{q}, \hat{f}\left(x\right) + \hat{q} \right].$$

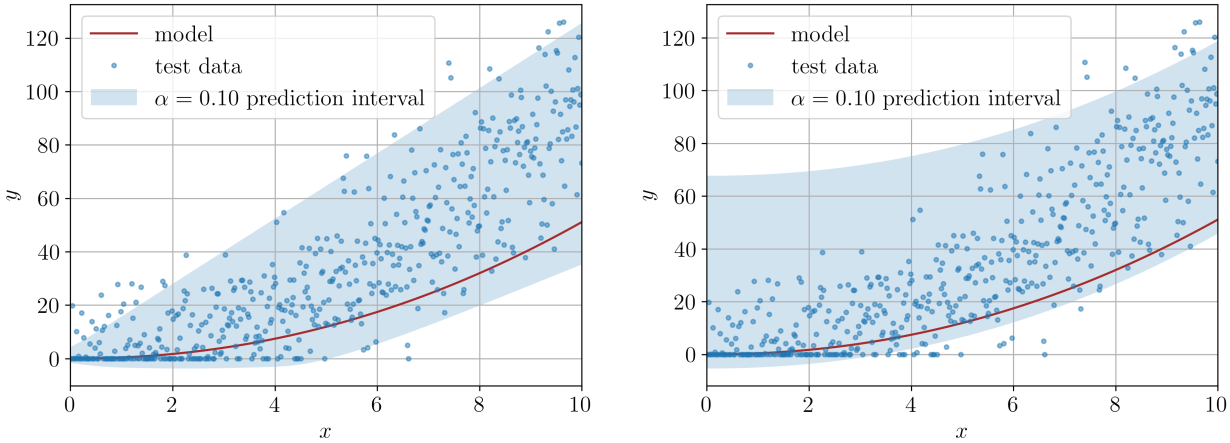

Figure 4.1: SCP using different score functions applied to the task introduced in Figure 1.1. Left: Conformalized Quantile Regression using simple neural networks to predict quantiles trained using the pinball loss. Right: Signed-error Split Conformal Regression.

4.2. Scaled absolute error score

When we assume that the model also predicts a positive symmetric uncertainty estimate $u(x)$, we can set $s(x, y) \coloneqq |y - \hat{f}(x)|/u(x)$, resulting in

$$\mathcal{C}(x) = \left\{ y \in \mathbb{R} \,\middle|\, \frac{|y - \hat{f}(x)|}{u(x)} \leq \hat{q} \right\} \iff y \in \left[\hat{f}(x) - \hat{q} \cdot u(x), \hat{f}(x) + \hat{q} \cdot u(x)\right].$$

4.3. Conformalized Quantile Regression (CQR)

Conformalized Quantile Regression (CQR) was introduced by [RPC19] and allows the modeling of asymmetric uncertainty. It requires a model that in addition to the mean estimate $\hat{f}(x)$ provides estimates of the lower and upper quantiles $\hat{q}_{\tau_{\text{lo}}}(x)$ and $\hat{q}_{\tau_{\text{hi}}}(x)$ at, typically predefined, levels τ_{lo} and τ_{hi}. To correct both at the same time, the score function

$$s(x, y) \coloneqq \max\left\{\hat{q}_{\tau_{\text{lo}}}(x) - y, y - \hat{q}_{\tau_{\text{hi}}}(x)\right\}$$

can be used, which results in

$$\mathcal{C}(x) = \left\{ y \in \mathbb{R} \,\middle|\, \max\left\{\hat{q}_{\tau_{\text{lo}}}(x) - y, y - \hat{q}_{\tau_{\text{hi}}}(x)\right\} \leq \hat{q} \right\} \iff y \in \left[\hat{q}_{\tau_{\text{lo}}}(x) - \hat{q}, \hat{q}_{\tau_{\text{hi}}}(x) + \hat{q}\right].$$

Let us convince ourselves that this definition of $s(x, y)$ results in the desired prediction interval: $\hat{q}_{\tau_{\text{lo}}}(x) - y \leq \hat{q}$ implies $y \geq \hat{q}_{\tau_{\text{lo}}}(x) - \hat{q}$; $y - \hat{q}_{\tau_{\text{hi}}}(x) \leq \hat{q}$ implies $y \leq \hat{q}_{\tau_{\text{hi}}}(x) + \hat{q}$; and $\max\left\{\hat{q}_{\tau_{\text{lo}}}(x) - y, y - \hat{q}_{\tau_{\text{hi}}}(x)\right\} \leq \hat{q}$ implies both $\hat{q}_{\tau_{\text{lo}}}(x) - y \leq \hat{q}$ and $y - \hat{q}_{\tau_{\text{hi}}}(x) \leq \hat{q}$.

The estimates of the lower and upper quantiles can be learned from data. We either use different models to represent $\hat{q}_{\tau_{\text{lo}}}(x)$ and $\hat{q}_{\tau_{\text{hi}}}(x)$ or add additional outputs to our model representing $\hat{f}$ (e.g., if $\hat{f}$ is a deep neural network, one can add an additional output head). These models can be trained by optimizing the pinball loss (see [SC11] and references therein).

4.4. Signed-error Split Conformal Regression

In all examples above, both upper and lower end of the prediction intervals shifted by the same value $\hat{q}$. This can serve as a potential limitation if the model errors are asymmetric. In that case, large errors on one side of $\hat{f}(x)$ could inflate $\hat{q}$ by dominating the CDF of non-conformity scores, resulting in large overcoverage on the the other side of $\hat{f}(x)$. Importantly, this argument holds even for the asymmetric quantiles of CQR – there is nothing guaranteeing that the corrections to the upper and lower quantiles should be identical. To address this issue, we need to construct two upper and lower conformal quantiles – an idea introduced by [LJL14]. As in this original work, we will also present the method based on the absolute-error score function.

Instead of defining $s(x, y) = |y - \hat{f}(x)|$, two signed-error scores are used

$$s_{i,\mathrm{lo}} \coloneqq s_{\mathrm{lo}}(x_i, y_i) \coloneqq \hat{f}(x_i) - y_i, \quad s_{i,\mathrm{hi}} \coloneqq s_{\mathrm{hi}}(x_i, y_i) \coloneqq y_i - \hat{f}(x_i),$$

for which we run the conformal prediction at two miscoverage levels $\alpha_{\mathrm{lo}}, \alpha_{\mathrm{hi}}$, resulting in

$$\hat{q}_{\mathrm{lo}} \coloneqq \widehat{\mathcal{Q}}_{(1-\alpha_{\mathrm{lo}})((n+1)/n)}\left(\{s_{i,\mathrm{lo}}\}_{i=1}^{n}\right) \quad \text{and} \quad \hat{q}_{\mathrm{hi}} \coloneqq \widehat{\mathcal{Q}}_{(1-\alpha_{\mathrm{lo}})((n+1)/n)}\left(\{s_{i,\mathrm{hi}}\}_{i=1}^{n}\right),$$

and since both are valid score function by Theorem 1 we have that

$$\begin{aligned}\mathbb{P}\left(Y_{\mathrm{test}} \in \mathcal{C}_{\mathrm{lo}}(X_{\mathrm{test}})\right) &= \mathbb{P}\left(Y_{\mathrm{test}} \geq \hat{f}(X_{\mathrm{test}}) - \hat{q}_{\alpha_{\mathrm{lo}}}\right) \geq 1 - \alpha_{\mathrm{lo}}, \\ \mathbb{P}\left(Y_{\mathrm{test}} \in \mathcal{C}_{\mathrm{hi}}(X_{\mathrm{test}})\right) &= \mathbb{P}\left(Y_{\mathrm{test}} \leq \hat{f}(X_{\mathrm{test}}) + \hat{q}_{\alpha_{\mathrm{hi}}}\right) \geq 1 - \alpha_{\mathrm{hi}}.\end{aligned}$$

Now, because miscoverage events are disjoint, we have

$$\begin{aligned}\mathbb{P}\left(Y_{\mathrm{test}} \notin \mathcal{C}_{\mathrm{lo}}(X_{\mathrm{test}}) \cup Y_{\mathrm{test}} \notin \mathcal{C}_{\mathrm{hi}}(X_{\mathrm{test}})\right) &= \mathbb{P}\left(Y_{\mathrm{test}} \notin \mathcal{C}_{\mathrm{lo}}(X_{\mathrm{test}})\right) + \mathbb{P}\left(Y_{\mathrm{test}} \notin \mathcal{C}_{\mathrm{hi}}(X_{\mathrm{test}})\right) \\ &\leq \alpha_{\mathrm{lo}} + \alpha_{\mathrm{hi}},\end{aligned} \tag{4.1}$$

which implies

$$\mathbb{P}\left(Y_{\mathrm{test}} \in \mathcal{C}_{\mathrm{lo}}(X_{\mathrm{test}}) \cup Y_{\mathrm{test}} \in \mathcal{C}_{\mathrm{hi}}(X_{\mathrm{test}})\right) \geq 1 - \alpha_{\mathrm{lo}} - \alpha_{\mathrm{hi}}.$$

Combining $\mathcal{C}_{\mathrm{lo}}$ and $\mathcal{C}_{\mathrm{hi}}$ results in

$$\mathcal{C}(X_{\mathrm{test}}) = \left[\hat{f}(X_{\mathrm{test}}) - \hat{q}_{\mathrm{lo}}, \hat{f}(X_{\mathrm{test}}) + \hat{q}_{\mathrm{hi}}\right]$$

and by setting $\alpha = \alpha_{\mathrm{lo}} + \alpha_{\mathrm{hi}}$ we get the standard marginal coverage guarantee

$$\mathbb{P}\left(Y_{\mathrm{test}} \in \mathcal{C}(X_{\mathrm{test}})\right) \geq 1 - \alpha.$$

At first glance the method might seem superior when compared to standard SCP for regression – combining $\mathcal{C}_{\mathrm{lo}}$ and $\mathcal{C}_{\mathrm{hi}}$ does not cost us anything (i.e., there are no bounds that make the final coverage guarantee looser since we have an equality in Eq. (4.1)), while we obtain asymmetric correction of confidence intervals. However, in practice, signed-error conformal regression often produces wider prediction intervals than ones obtained with standard SCP. The reason is that in standard SCP, the probability of the true value falling outside the entire interval is at most α, where the errors can be distributed arbitrarily above or below the interval. On the other hand, signed-error conformal regression provides a stronger, two-part guarantee. It ensures that the probability of the true value falling above the upper bound is at most α_{hi}, and the probability of it falling below the lower bound is also at most α_{lo}. Thus, standard SCP may produce tighter confidence intervals, as it can balance the errors on both sides which signed-error conformal regression can not achieve as it is forced to provide this stronger guarantee.

4.5. Calibration-conditional coverage

The result in Theorem 1 does not depend on the size of the calibration data set $\mathcal{D}_{\text{cal}}$, so it also holds for calibration data sets that contain only a single element. This should make us suspicious. Theorem 1 shows that we have a strong coverage guarantee in expectation over all possible test and calibration samples. It does not make a statement about a specific fixed calibration data set $\mathcal{D}_{\text{cal}}$. However, since in practice we are always confined to the limited data settings, we need to consider how this guarantee behaves when conditioning on the specific realizations of the random variables. First, we take a look at the calibration-conditional coverage

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right).$$

We can exploit the concentration properties to derive the following bound.

Theorem 2 (Calibration-conditional coverage). *Under the assumptions that the data points $\{(X_i, Y_i)_i\}_{i=1}^n = \mathcal{D}_{cal}$ and (X_{test}, Y_{test}) are i.i.d. and that $\mathcal{C}$ was constructed via split conformal prediction (Algorithm 1) using any pre-trained non-conformity score function s at a miscoverage level β, then the calibration-conditional coverage stochastically dominates the beta distribution*

$$\mathbb{P}\left(\mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right) \mid \mathcal{D}_{cal}\right) \leq 1-\alpha\right) \leq F_{Beta((1-\beta)(n+1), \beta(n+1))}\left(1-\alpha\right),$$

where $F_{Beta(a,b)}$ denotes the CDF of the Beta$\,(a, b)$ distribution.

Proof: Let F_S be the CDF of the distribution of scores $s(X, Y)$. Define $S_i = s(X_i, Y_i)$, which are i.i.d. draws from the distribution with CDF F_S, and let $S_{(1)} \leq \cdots \leq S_{(n)}$ be the order statistics of $S_1, \ldots, S_n$. By the definition of SPC, where $k = \lceil (1-\beta)(n+1) \rceil$, we have

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) = \mathbb{P}\left(S_{\text{test}} \leq S_{(k)} \mid \mathcal{D}_{\text{cal}}\right) = F_S\left(S_{(k)}\right),$$

where the last step holds since S_{test} is independent of $\mathcal{D}_{\text{cal}}$ and has CDF F_S. Therefore

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) \leq 1-\alpha\right) = \mathbb{P}\left(F_S\left(S_{(k)}\right) \leq 1-\alpha\right).$$

Now, let $Z_i = F_S(S_i)$. Since F_S is non-decreasing, the order statistics $Z_{(1)} \leq \cdots \leq Z_{(n)}$ of $Z_1, \ldots, Z_n$ satisfy $Z_{(k)} = F_S\left(S_{(k)}\right)$. Moreover, by the probability integral transform (see Theorem 5 in Appendix B), if F_S is continuous, each Z_i is an i.i.d. sample from the uniform distribution $\mathcal{U}_{[0,1]}$. Otherwise, if F_S has discontinuities, the resulting Z_i stochastically dominate $U_1, \ldots, U_n \sim \mathcal{U}_{[0,1]}$, meaning $\forall_{i \in [n], x \in [0,1]}$ we have $F_{Z_i}(x) \leq F_{U_i}(x)$. Combining both observations, we get

$$\mathbb{P}\left(F_S\left(S_{(k)}\right) \leq 1-\alpha\right) = \mathbb{P}\left(Z_{(k)} \leq 1-\alpha\right) \leq \mathbb{P}\left(U_{(k)} \leq 1-\alpha\right).$$

Finally, by the definition of Beta distribution, the k-th order statistic of a sample of size n from $\mathcal{U}_{[0,1]}$ is a beta random variable

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

and since $k \geq (1-\beta)(n+1)$, the distribution above stochastically dominates
$\text{Beta}\left((1-\beta)(n+1), \beta(n+1)\right)$. ■

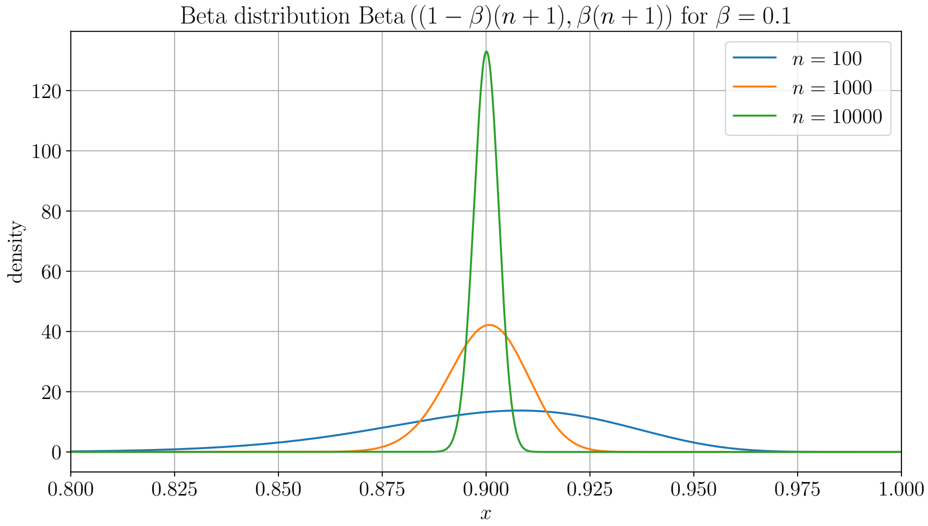

Figure 4.2: Beta distribution.

Theorem 2 shows that the internal coverage level $1-\beta$ and the calibration-conditional coverage level $1-\alpha$ that we want to achieve are connected through a beta distribution, which is visualized in Figure 4.2. Perhaps the tradeoff between the two is not immediately apparent since there is no closed form for CDF of a beta distribution. Thus, we derive the following upper bound to get an intuition of the rate at which calibration-conditional coverage approaches the marginal coverage guarantee with growing size n.

Lemma 3 (Calibration-conditional coverage upper bound). *Under the conditions of Theorem 2 we have that*

$$\mathbb{P}\left(\mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right) \mid \mathcal{D}_{cal}\right) \leq 1-\alpha\right) \leq e^{-2n\Delta^2},$$

where the miscoverage rate $\beta := \alpha - \Delta$ is used when computing $\mathcal{C}\left(X_{test}\right)$.

Proof: In the proof of Theorem 2 we established that

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) \leq 1-\alpha\right) \leq \mathbb{P}\left(U_{(k)} \leq 1-\alpha\right),$$

where $U_{(k)}$ is the k-th order statistic of n uniformly distributed random variables $U_1, \ldots, U_n \sim \mathcal{U}_{[0,1]}$, where $k = \lceil (1-\beta)(n+1) \rceil = \lceil (1-\alpha+\Delta)(n+1) \rceil$. This is equivalent to saying that at least k realizations of U_i are $\leq 1-\alpha$. Thus, by using $V_i = \mathbb{1}\left\{U_i \leq 1-\alpha\right\}$ the probability can be rewritten as

$$\mathbb{P}\left(U_{(k)} \leq 1-\alpha\right) = \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq k\right) \leq \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq (1-\alpha+\Delta)(n+1)\right).$$

Since $V_1, \ldots, V_n$ are i.i.d. with expected value

$$\mathbb{E}\left[V_i\right] = \mathbb{E}\left[\mathbb{1}\left\{U_i \leq 1-\alpha\right\}\right] = \mathbb{P}\left(U_i \leq 1-\alpha\right) = 1-\alpha,$$

for all $i \in [n]$, we can apply Hoeffding's inequality (see Theorem 6 in Appendix B) to get

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} V_i \geq (1-\alpha+\Delta)(n+1)\right) &\leq \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq (1-\alpha+\Delta)\, n\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} V_i - n\mathbb{E}\left[V\right] \geq (1-\alpha+\Delta)\, n - n(1-\alpha)\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} V_i - n\mathbb{E}\left[V\right] \geq n\Delta\right) \\
&\leq e^{-2n\Delta^2}
\end{aligned}$$

ending the proof. ■

This result shows that there is a price to pay: to achieve calibration-conditional coverage of $1-\alpha$ we need to increase the internal coverage of Split CP to $1-\alpha+\Delta$, making the predicted confidence sets less conservative. The proof of Lemma 3 follows identical steps if instead of increasing the internal coverage level by Δ we decrease the calibration-conditional coverage level by the same value (it has to be decreased for the Hoeffding's bound to apply). Thus, we pay the price of comparable order of magnitude in either the prediction set size or calibration-conditional coverage.

Before proceeding, we wish to clarify a crucial distinction. While the marginal coverage guarantee (Theorem 1) applies to both i.i.d. and exchangeable data, calibration-conditional coverage strictly requires the i.i.d. assumption. This is because exchangeable data generally lacks the concentration properties necessary for calibration-conditional coverage. For an extended discussion, please refer to Chapter 4 of [ABB25].

PAC-style formulation. Using the result of Theorem 2, we can derive a probably approximately correct (PAC) bound on the calibration-conditional coverage, allowing us to be fully rigorous even in the limited data setting. Writing in the standard form, for some $\delta \in (0, 1)$ we want the prediction sets $\mathcal{C}\left(X_{\text{test}}\right)$ to satisfy

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\text{test}} \notin \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) \leq \alpha\right) \geq 1 - \delta \tag{4.2}$$

which is equivalent to

$$\mathbb{P}\left(\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(X_{\text{test}}\right) \mid \mathcal{D}_{\text{cal}}\right) \leq 1 - \alpha\right) \leq \delta. \tag{4.3}$$

By applying Theorem 2 to Eq. (4.3) we can connect the probability bound δ and the internal $\beta := \alpha - \Delta$ to achieve the desired coverage level $1 - \alpha$ even in the calibration-conditional setting.

Corollary 1. *For any $\alpha, \delta \in (0, 1)$ and β being a solution to $F_{Beta((1-\beta)(n+1),\beta(n+1))}\left(1 - \alpha\right) = \delta$, we have that Split Conformal Prediction ran at miscoverage level β achieves $1 - \alpha$ level with probability at least $1 - \delta$*

$$\mathbb{P}\left(1 - \alpha \leq \mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right) \mid \mathcal{D}_{cal}\right)\right) \geq 1 - \delta.$$

Numerically solving for β leads to a unique solution, while solving through Lemma 3 provides a more loose but closed-form expression $\beta = \alpha - \sqrt{\ln\left(1/\delta\right)/2n}$. The second bound is looser, as it can be shown that

$$F_{\text{Beta}((1-\alpha+\Delta)(n+1),(\alpha+\Delta)(n+1))}\left(1 - \alpha\right) \leq e^{-2n\Delta^2}$$

using the fact that the beta distribution is sub-gaussian [MA17]. We omit the proof here for brevity. Nevertheless, to reiterate the key point: internally we need to use $1-\beta$ increased by $O\left(n^{-1/2}\right)$ compared to $1 - \alpha$ so that we obtain calibration-conditional coverage of at least $1 - \alpha$ with probability at least $1 - \delta$.

5. Test-conditional coverage

The standard evaluation of ML models is often limited to estimating their performance in expectation over test samples. However, being able to provide guarantees about the error of specific samples is arguably important and in many cases practically more relevant than a kind of average accuracy or confidence. In medical diagnosis, for example, a patient will be interested in the reliability of a test result in her particular case, not in the reliability of the test on average.

5.1. Theoretical limitations of conditional coverage

Can conformal prediction be tailored in such way? As often, it depends critically on the nature of the feature space $\mathcal{X}$.

Definition 2 (Atoms, nonatomic distributions). *For a distribution P on $\mathcal{Z}$, the atoms of P are all points $z \in \mathcal{Z}$ with positive probability,*

$$atom\,(P) = \{z \in \mathcal{Z} \;:\; \mathbb{P}_P\,(Z = z) > 0\}\,.$$

The distribution P is called nonatomic if it has no atoms, i.e., $atom\,(P) = \varnothing$ (the empty set).

As a concrete example, a distribution on $\mathbb{R}$ is nonatomic if and only if it is continuous (i.e., has a density). In that case, we are faced with an impossibility result [ABB25]:

Theorem 3 (Impossibility of test-conditional coverage (nonatomic)). *Suppose $\mathcal{C}$ is any procedure that satisfies distribution-free conditional coverage, i.e., for any distribution P on $\mathcal{X} \times \mathcal{Y}$,*

$$\mathbb{P}\,(Y_{test} \in \mathcal{C}\,(X_{test}) \;|\; X_{test}) \geq 1 - \alpha$$

holds almost surely, where the probability is taken with respect to $(X_{test}, Y_{test})\,, (X_1, Y_1)\,, \cdots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$. Then, for any distribution P on $\mathcal{X} \times \mathcal{Y}$ for which the marginal P_X is nonatomic and $\forall\,(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\mathbb{P}\,(y \in \mathcal{C}\,(x)) \geq 1 - \alpha.$$

In other words, Theorem 3 shows that in the case of nonatomic P_X we get deterministic coverage from the perspective of test and calibration samples. This is only achieved when the prediction sets are uninformative. For example, if we construct them such that $\forall_{x \in \mathcal{X}}\, \mathcal{C}\,(x) = \mathcal{Y}$, or in a randomized fashion:

$$\mathcal{C}\,(x) = \begin{cases} \mathcal{Y}, & \text{with probability } 1 - \alpha \\ \varnothing, & \text{with probability } \alpha \end{cases}$$

For the proof and an extended discussion on Theorem 3, please refer to [ABB25]. The intuition behind this impossibility result is that a truly distribution-free method must remain valid for any data distribution, including pathological “worst-case” scenarios. To see why this is a problem, consider a critic who constructs a malicious distribution, P', by mixing a smooth distribution P with an infinitesimal point mass (a “spike”) at an arbitrary location (x_0, y_0). Because the feature space is assumed to be continuous (non-atomic), the calibration set will almost surely not contain the exact point x_0, rendering it “blind” to the spike’s existence. However, to maintain its guarantee, the method must cover y_0 when tested at $X_{n+1} = x_0$. Since this must hold for any arbitrarily chosen spike location, the method is forced to produce uninformatively large prediction sets (e.g., intervals of infinite expected length) for all distributions—even smooth ones—to defend against an adversarial example it cannot rule out.

Thus, to derive any meaningful test-conditional coverage guarantees, we need some relaxations. One simple approach is to perform a binning on $\mathcal{X}$ as described below.

5.2. Bin-wise conditional coverage

One simple way to introduce some weak form of dependence on X_{test} is to partition $\mathcal{X}$ and to run SCP for each partition independently.

Theorem 4 (Bin-wise conditional coverage). *Assuming setup of Theorem 1, partition $\mathcal{X}$ into B bins $\mathcal{X}_1, \ldots, \mathcal{X}_B$. Further, let $\mathcal{I}_b = \{i \in \{1, \ldots, n\} \mid X_i \in \mathcal{X}_b\}$ be the index set of features X_i belonging to bin $\mathcal{X}_b$, $n_b = |\mathcal{I}_b|$ and $\hat{q}_b := \widehat{\mathcal{Q}}_{(1-\alpha)((n_b+1)n_b)}\left(\{S_i\}_{i\in\mathcal{I}_b}\right)$ be the empirical quantile computed according to Algorithm 1 for each bin $b \in \{1, \ldots, B\}$ separately by partitioning $\mathcal{D}_{cal}$ into $\{(X_i, Y_i)\}_{i\in\mathcal{I}_b}$. By defining*

$$\mathcal{C}\left(X_{test}\right) = \{y \ : \ X_{test} \in \mathcal{X}_b \wedge S_{test} \leq \hat{q}_b\}$$

we get the bin-wise conditional coverage guarantee

$$\mathbb{P}\left(Y_{test} \in \mathcal{C}\left(Y_{test}\right) \ \mid X_{test} \in \mathcal{X}_b\right) \geq 1 - \alpha$$

for all $b \in \{1, \ldots, B\}$ with $\mathbb{P}\left(X_{test} \in \mathcal{X}_b\right) > 0$.

We do not provide a rigorous proof, but it is based on the fact that if samples are i.i.d. on $\mathcal{X} \times \mathcal{Y}$, they are also i.i.d. in their respective bins. One could argue that Theorem 4 is trivial, it simply tells us to split the data into bins and apply confomal prediction independently per bin. Immediately we also see that prediction set constructed as in Theorem 4 satisfies marginal coverage guarantee, since

$$\mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(Y_{\text{test}}\right)\right) = \sum_{b=1}^{B} \mathbb{P}\left(X_{\text{test}} \in \mathcal{X}_b\right) \cdot \mathbb{P}\left(Y_{\text{test}} \in \mathcal{C}\left(Y_{\text{test}}\right) \ \mid X_{\text{test}} \in \mathcal{X}_b\right) \geq 1 - \alpha.$$

However, in some instances there is no natural way of binning the data, and finding an appropriate bin separation is highly task-dependent, and for each b the number of calibrations samples $|\mathcal{I}_b|$ in that bin need to be large enough. For example, how could we bin satellite images? Binning by biome could be meaningful. Even then, it is just an approximate version of test-conditional coverage, since coverage is limited to bin-average (i.e., expectation over X values in the bin), and we would need enough samples in each bin so that calibration-conditional coverage is reasonable.

Next, we will introduce the Split Localized Conformal Prediction (SLCP), that through leveraging the training data computes an estimate of local quantile and employs SCP to correct the local quantiles in limited data regime so that we still have the coverage guarantees. Importantly, with growing training data this method achieves asymptotic test-conditional coverage.

5.3. Split Localized Conformal Prediction

An ideal solution to the test-conditional coverage would be a ground-truth quantile of the score's conditional CDF, $q^*_{1-\alpha}(x) := \mathcal{Q}_{1-\alpha}\left(F_{S|X}\left(S|x\right)\right)$, as for that quantile we would have

$$\mathbb{P}\left(S \leq q^*_{1-\alpha}\left(x\right) \ \mid X = x\right) = 1 - \alpha. \tag{5.1}$$

Naturally we do not have access to $q^*_{1-\alpha}(x)$, but we can try to approximate it with some $\tilde{q}_{1-\alpha}(x) \approx q^*_{1-\alpha}(x)$. However, then we lose any coverage guarantees

$$\mathbb{P}\left(S \leq \tilde{q}_{1-\alpha}(x) \mid X = x\right) \quad \overset{?}{\geq} \quad 1 - \alpha.$$

Now, the idea on how to enforce these guarantees introduced by [HTGL23] is to transform Eq. (5.1) into

$$\mathbb{P}\left(S - \tilde{q}_{1-\alpha}(x) \leq q^*_{1-\alpha}(x) - \tilde{q}_{1-\alpha}(x) \mid X = x\right) = 1 - \alpha,$$

and then run the conformal prediction on the residual of ideal and approximated quantiles

$$\mathbb{P}\left(S' \leq \hat{q}\right) \geq 1 - \alpha,$$

with the new score $S' = S - \tilde{q}_{1-\alpha}(x)$. If our approximation is correct to a similar degree for all x, then the residual becomes a constant somewhat close to zero motivating the use of SCP that returns a single $\hat{q}$ responsible for providing coverage guarantee for all X, implicitly enforcing test-conditional coverage. For example, let us assume that the original score function is the absolute error, then we get:

$$\mathcal{C}(x) = \left\{ y \in \mathbb{R} \;\middle|\; \overbrace{\overbrace{|y - \hat{f}(x)|}^{s(x,y)} - \tilde{q}_{1-\alpha}(x)}^{s'(x,y)} \leq \hat{q} \right\}$$
$$\iff y \in \left\{ \hat{f}(x) - \underbrace{\hat{q}}_{\substack{\text{CP}\\\text{correction}}} - \underbrace{\tilde{q}_{1-\alpha}(x)}_{\substack{\text{estimated}\\\text{quantile}}}, \hat{f}(x) + \hat{q} + \tilde{q}_{1-\alpha}(x) \right\}$$

To learn $\tilde{q}_{1-\alpha}$ from data, [HTGL23] suggested the Nadaraya–Watson density estimator [Nad64, Wat64], see Appendix C, to model CDFs. Adapting Definition 6 in Appendix C to our case of modelling $\widehat{F}_{S|X}(S|x)$, we naturally replace Y with S, and then obtaining $\tilde{q}_{1-\alpha}(x)$ reduces to computing the quantile over the empirical CDF. Crucially, however, $\widehat{F}_{S|X}(S|x)$ cannot be computed using $\mathcal{D}_{\text{cal}}$ as it would violate the i.i.d. assumption of the calibration scores S' with S'_{test}. For that, we can leverage the training dataset, resulting in the following definition

$$\tilde{q}_{1-\alpha}(x) = \widetilde{\mathcal{Q}}_{1-\alpha}\left(\widehat{F}_{S|X}(S|x)\right) = \inf\left\{ s : 1 - \alpha \leq \sum_{(x',y') \in \mathcal{D}_{\text{train}}} w(x'|x)\, \mathbb{1}\left\{s \leq s(x'y')\right\} \right\} ,$$

where the weights $w(x'|x)$ are computed as shown in Appendix C. Under the assumption that the marginal distribution P_X belongs to the Hölder class (see Definition 4 in Appendix A), [HTGL23] show that as $n \to \infty$ the empirical CDF $\widehat{F}_{S|X}$ approximated by the Nadaraya–Watson estimator asymptotically approaches the ground truth $F_{S|X}$, resulting in asymptotic test-conditional coverage guarantee. This result holds true regardless of the choice of the non-conformity score function s. However, does that theoretical result translate to practical performance in limited data setting? Using the signed-error score function, [HTGL23] compare this approach with other state-of-the-art methods, naming the algorithm Split Localized Conformal Prediction (SLCP). Algorithm 2 describes SLPC, using the signed-error score (see subsection 4.4) and the Nadaraya–Watson estimator, in pseudo-code. The corresponding results on our example problem are shown in the right plot in Figure 1.1.

Algorithm 2: Split Localized Conformal Prediction (SLCP)

```
Input: $\mathcal{D}_{\text{cal}}, \mathcal{D}_{\text{train}} \in 2^{\mathcal{X}\times\mathcal{Y}}$; model $\hat{f} : \mathcal{X} \to \mathcal{Y}$; test input $x_{\text{test}} \in \mathcal{X}$; $\alpha_{\text{lo}}, \alpha_{\text{hi}} \in (0,1)$;
       kernel $K : \mathbb{R} \to \mathbb{R}$
Output: Prediction set $\mathcal{C}(x_{\text{test}}) \in 2^{\mathcal{Y}}$ with coverage $\geq 1 - (\alpha_{\text{lo}} + \alpha_{\text{hi}})$ and
        $\forall_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \lim_{|\mathcal{D}_{\text{cal}}|\to\infty} \mathbb{P}\left(y \in \mathcal{C}\left(x\right) |X_{\text{test}} = x\right) \geq 1 - \alpha$.
$s_{\text{lo}}(x,y) := -s_{\text{hi}}(x,y) := \hat{f}(x) - y$
for $l \in \{lo, hi\}$ do
    $w(x_i|x) := \frac{K(\|x_i - x\|)}{\sum_{(x_j,y_j)\in\mathcal{D}_{\text{train}}} K(\|x_j - x\|)}$
    $\tilde{q}_{1-\alpha_l}(x) := \inf\left\{s : 1 - \alpha_l \leq \sum_{(x_i,y_i)\in\mathcal{D}_{\text{train}}} w(x_i|x)\mathbb{1}[s \leq s(x_i,y_i)]\right\}$
    $\hat{q}_l \leftarrow$ run SCP on $\mathcal{D}_{\text{cal}}$ with miscoverage $\alpha_l$ and score $s(x,y) = s_l(x,y) - \tilde{q}_{1-\alpha_l}(x)$
return
    $\mathcal{C}(x_{test}) := \left\{y \in \mathcal{Y} \,\middle|\, y \in \left[\hat{f}(x_{test}) - \tilde{q}_{1-\alpha_{lo}}(x_{test}) - \hat{q}_{lo}, \hat{f}(x_{test}) + \tilde{q}_{1-\alpha_{hi}}(x_{test}) + \hat{q}_{hi}\right]\right\}$
```

6. Summary

Conformal prediction

- gives prediction sets with rigorous coverage guarantees,
- requires no assumption on model, data generating distribution, and algorithm except calibration data and test data being i.i.d.,
- can deal with heteroscedastic, asymmetric noise/uncertainty,
- can turn your favourite uncertainty estimates into proper prediction sets,
- is easy to implement (there also exists a Python library MAPIE providing various conformal prediction algorithms [CBL+23]).

References

[AB23] A. N. Angelopoulos and S. Bates. Conformal prediction: a gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4): 494–591, 2023.

[ABB25] A. N. Angelopoulos, R. F. Barber, and S. Bates. *Theoretical Foundations of Conformal Prediction*. 2025. arXiv: `2411.11824 [math.ST]`.

[BCRT21] R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1): 486–507, 2021.

[CBL+23] T. Cordier, V. Blot, L. Lacombe, T. Morzadec, A. Capitaine, and N. Brunel. Flexible and systematic uncertainty estimation with conformal prediction via the MAPIE library. *Conformal and Probabilistic Prediction with Applications,* PMLR. 549–581, 2023.

[HS90] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 12(10): 993–1001, 1990.

[HTGL23] X. Han, Z. Tang, J. Ghosh, and Q. Liu. *Split Localized Conformal Prediction*. 2023. arXiv: `2206.13092 [stat.ML]`.

[LJL14] H. Linusson, U. Johansson, and T. Löfström. Signed-error conformal regression. *Advances in Knowledge Discovery and Data Mining (PKDD),* 224–236, 2014.

[MA17] O. Marchal and J. Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22: 1–14, 2017.

[Nad64] E. A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1): 141–142, 1964.

[NW94] D. A. Nix and A. S. Weigend. Estimating the mean and variance of the target probability distribution. *International Conference on Neural Networks (ICNN),* vol. 1. IEEE. 55–60, 1994.

[Pap24] H. Papadopoulos. Guaranteed coverage prediction intervals with Gaussian process regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9072–9083, 2024.

[RPC19] Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems (NeurIPS),* 3543–3553, 2019.

[SC11] I. Steinwart and A. Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1): 211–225, 2011.

[VGS99] V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. *International Conference on Machine Learning (ICML),* 444–453, 1999.

[Wat64] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4): 359–372, 1964.

[Zar68] J. H. Zar. Calculation and miscalculation of the allometric equation as a model in biological data. *BioScience*, 18(12): 1118–1120, 1968.

A. Concepts from analysis

Definition 3 (Lipschitz continuity). *A real-valued function $g : \mathbb{R}^d \to \mathbb{R}$ where $d \in \mathbb{Z}^+$ is Lipschitz continuous if its rate of change is bounded by a constant $L > 0$:*

$$|g(x) - g(y)| \leq L\|x - y\|$$

Definition 4 (Hölder class). *A function $g : \mathbb{R}^d \to \mathbb{R}$ where $d \in \mathbb{Z}^+$ belongs to the Hölder class $\Sigma(\beta, L)$ if all of its partial derivatives up to order $\beta - 1$ exist and are themselves Lipschitz continuous with a constant $L > 0$. The parameter $\beta \geq 1$ is the smoothness index.*

B. Statistics background

Theorem 5 (Probability integral transform). *Suppose that a random variable X has a continuous distribution for which the cumulative distribution function (CDF) is F_X. Then the random variable Y defined as*

$$Y := F_X(X)$$

has a standard uniform distribution $\mathcal{U}_{[0,1]}$. If F_X is not continuous then Y is a superuniform random variable, meaning that

$$\forall_{u \in [0,1]}\ \mathbb{P}(Y \leq u) \leq u,$$

or equivalently, Y dominates the uniform distribution:

$$\forall_{u \in [0,1]}\ F_Y(u) \leq F_U(u), \quad \textit{where} \quad U \sim \mathcal{U}_{[0,1]}.$$

Theorem 6 (Hoeffding's inequality). *Let $X_1, \ldots, X_n$ be independent real-valued random variables, such that for each $i \in \{1, \ldots, n\}$ there exist $a_i \leq b_i$, such that $X_i \in [a_i, b_i]$. Then for every $\epsilon > 0$:*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \epsilon\right) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)}$$

and

$$\mathbb{P}\left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] - \sum_{i=1}^{n} X_i \geq \epsilon\right) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n}(b_i - a_i)}.$$

Definition 5 (Exchangeability). *Let $Z_1, \ldots, Z_n \in \mathcal{Z}$ be random variables with a joint distribution. We say that the random list $(Z_1, \ldots, Z_n)$ is exchangeable if, for every permutation $\sigma \in Sym([n])$*

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

where $\overset{d}{=}$ denotes equality in distribution, and $Sym([n])$ is the set of all permutations on $[n] := \{1, \ldots, n\}$ (symmetric group on n elements).

C. Density estimation

Definition 6 (Nadaraya–Watson estimator of a conditional CDF). *Let* (x_i, y_i) *for* $i \in [m]$ *be the set of* m *samples, and* K *be a kernel with a bandwidth* h *such that* $K(\cdot)$ *has a mean zero (i.e., is of order at least 1). Then, the Nadaraya–Watson estimator of the CDF* $F_{Y|X}$ *is defined as*

$$\widehat{F}_{Y|X}(y|x) = \sum_{i=1}^{m} w(x_i|x)\, \mathbb{1}\{y \leq y_i\}$$

with

$$w(x_i|x) = \frac{K(\|x_i - x\|)}{\sum_{j=1}^{m} K(\|x_j - x\|)} = \frac{K(\|x_i - x\|)}{\sum_{j=1}^{m} K(\|x_j - x\|)}.$$

D. Dealing with ties

In the proof of Theorem 1, we used

$$S_{\text{test}} \leq S_{(n+1;k)} \iff \text{either } S_{\text{test}} < S_{(n+1;k+1)} \text{ or } S_{\text{test}} = S_{(n+1;k)} = S_{(n+1;k+1)}.$$

While this statement may be rather intuitive, let us prove it formally. First note that we cannot decouple $S_{\text{test}} \leq S_{(n+1;k)}$ into $S_{\text{test}} < S_{(n+1;k)}$ and $S_{\text{test}} = S_{(n+1;k)}$, because we are computing the order statistic with the $n+1$ samples including the test data, and thus we could have $S_{\text{test}} = S_{(n+1;k)}$ without any ties.

Let us first prove the $\Longrightarrow$ direction. By definition, $S_{(n+1;k)} \leq S_{(n+1;k+1)}$ and thus $S_{\text{test}} \leq S_{(n+1;k)}$ implies $S_{\text{test}} \leq S_{(n+1;k+1)}$, that is, we have either $S_{\text{test}} < S_{(n+1;k+1)}$ or $S_{\text{test}} = S_{(n+1;k+1)}$. In the latter case, $S_{(n+1;k)} \leq S_{(n+1;k+1)}$ implies $S_{(n+1;k)} \leq S_{\text{test}}$. As by assumption $S_{\text{test}} \leq S_{(n+1;k)}$, we must have $S_{\text{test}} = S_{(n+1;k)}$.

For the $\Longleftarrow$ direction, let us first assume $S_{\text{test}} = S_{(n+1;k)} = S_{(n+1;k+1)}$. In this case, $S_{\text{test}} = S_{(n+1;k)}$ trivially implies $S_{\text{test}} \leq S_{(n+1;k)}$. If we assume $S_{\text{test}} < S_{(n+1;k+1)}$, then the largest possible S_{test} fulfilling this strict inequality would be the next smaller element $S_{(n+1;k)}$, that is, $S_{\text{test}} \leq S_{(n+1;k)}$.