Chapter 3: Conformal Prediction Under Exchangeability

from Theoretical Foundations of Conformal Prediction

by Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates

Presentation by:

Anna Kosovskaia Matteo Gätzner Sarah Medding

ETH Zürich

May 6, 2025

Outline

Uncertainty Quantification for Predictions

- Data points $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, e.g. regression $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$.
- Common task: Predict Y_{n+1} given X_{n+1} and dataset $\{(X_i, Y_i)\}_{i=1}^n$ with some predictive model $\hat{f}(X_{n+1})$ e.g. neural net, linear model, smoothing spline ....
- **Problem:** Even if performance seems good empirically, e.g. high cross validation R^2 score, we have no real guarantee that inference time predictions are accurate.
- **Solution:** Use $\hat{f}$ to construct **prediction set** $\mathcal{C}(X_{n+1}) \subseteq \mathcal{Y}$ and get guarantees for $\mathcal{C}(X_{n+1})$!
- **Important**: Conformal prediction does **not** predict any probability, density or distribution! It predicts a set of plausible labels, i.e. labels that **conform** to the patterns in the observed data.

Nice Properties of Conformal Prediction

- Rigerous uncertainty quantification for predictive models.
- No assumptions about predictive model.
- No asymptotics, limit theorems, Gaussian approximations etc.
- Only minimal assumptions on data-generating distribution (exchangeability).

Marginal Coverage

- We aim for **marginal coverage**:

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

  with user-specified error level $\alpha \in (0, 1)$.
- In words: With high probability (at least $1 - \alpha$), the true value is in our prediction set $\mathcal{C}(X_{n+1})$.
- **Intuition**: $\mathcal{C}(X_{n+1})$ large $\implies$ high uncertainty.
  $\mathcal{C}(X_{n+1})$ small $\implies$ low uncertainty.
- Conformal prediction provides marginal coverage for all predictive models $\hat{f}$, even very bad ones, but better models yield smaller (more informative) prediction sets $\mathcal{C}(X_{n+1})$.

Sufficient condition

- How to construct prediction sets with marginal coverage?
- Sufficient condition: **exchangeability**.

Exchangeability Definition

Definition (Exchangeability)

Let $Z_1, \dots Z_n \in \mathcal{Z}$ be random variables with a joint distribution. We say that the random vector $(Z_1, \dots, Z_n)$ is *exchangeable* if, for every permutation $\sigma \in \mathcal{S}_n$,

$$(Z_1, \dots, Z_n) \stackrel{d}{=} (Z_{\sigma(1)}, \dots, Z_{\sigma(n)}),$$

where $\stackrel{d}{=}$ denotes equality in distribution, and S_n is the set of all permutations on $[n] := \{1, \dots, n\}$.

Reminder: Equality in Distribution

Definition (Equality in distribution)

A random variable $Z_1 \in \mathcal{Z}$ is said to be *equal in distribution* to another random variable $Z_2 \in \mathcal{Z}$, symbolically $Z_1 \overset{d}{=} Z_2$, their distribution functions match point-wise, i.e. if for all $z \in \mathcal{Z}$

$$P(Z_1 \leq z) = P(Z_2 \leq z).$$

Definition (Exchangeability for infinite sequences)

Let $Z_1, Z_2, \cdots \in \mathcal{Z}$ be an *infinite* sequence of random variables with a joint distribution. We say that this infinite sequence is exchangeable if $(Z_1, \ldots, Z_n)$ is exchangeable for every $n \geq 1$.

Exchangeability Intuition

- Statement about a random vector.
- **Intuition**: Sequence is equally likely to appear in any order.
- E.g. P is some distribution over $\{1, 2, 3\}$ and we draw two observations Z_1, Z_2, then

$$P(Z_1 = 1, Z_2 = 2) = P(Z_1 = 2, Z_2 = 1).$$

Scenarios Where We Have Exchangeability

- Exchangeability can arise in many scenarios. Non-exhaustive example scenarios:
- **Scenario 1**: If $Z_1, \ldots, Z_n$ are sampled uniformly without replacement from a potentially larger but finite set $\{z_1, \ldots, z_N\}$ with $N \geq n$.
- **Scenario 2**: If $Z_1, \ldots, Z_n$ are drawn i.i.d. from a distribution P on $\mathcal{Z}$.

Conformal Prediction Under Exchangeability

- Exchangeability. ✓
- Next: How conformal prediction works and why.

Setting

- Exchangeable sequence of data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$.
- Task: Predict unobserved Y_{n+1} given everything else.
- Notation: Upper case letters are random variables, lower case letters are fixed values.
- Conformal prediction constructs $\mathcal{C}(X_{n+1}) \subset \mathcal{Y}$ with marginal predictive coverage $\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha$.

Scores

- Construction uses score function s that maps a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^k$ (of any size k) to a real value $s((x, y); \mathcal{D}) \in \mathbb{R}$.
- Interpretation: Score function measures error of model on a single test point.
- Similar to a loss function in machine learning. High score means bad prediction, low score means good prediction.
- Example: Residual score $s((x, y); \mathcal{D}) = |y - \hat{f}(x; \mathcal{D})|$ where $\hat{f}(x; \mathcal{D})$ is the prediction of a model trained on $\mathcal{D}$.
- We'll assume *symmetric* score functions, i.e. functions that are invariant to permutations of $\mathcal{D}$.

Symmetric Score Functions

Definition

A score function s is symmetric if for any data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, any dataset $\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^k$, and any permutation σ on $[k]$, we have deterministic equality

$$s((x, y); \mathcal{D}) = s((x, y); \mathcal{D}_\sigma)$$

where $\mathcal{D}_\sigma$ is the dataset, permuted by σ.

The Full Conformal Prediction Procedure

- Next: the *full* **conform***al prediction* procedure.
- High level idea: invert score function to identify possible values $y \in \mathcal{Y}$ for the response Y_{n+1} that agree (or **conform**) with the trends observed in the available data.

Notation

- Training set with n data points:

$$\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

- Training set plus one test point:

$$\mathcal{D}_{n+1} = ((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})).$$

- *Augmented dataset*:

$$\mathcal{D}^y_{n+1} = ((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)).$$

We call (X_{n+1}, y) the *hypothesized test point* and y the *hypothesized reponse value*.

- Score of ith data point within augmented dataset:

$$S^y_i = \begin{cases} s((X_i, y); \mathcal{D}^y_{n+1}), & \text{if } i = n+1 \\ s((X_i, Y_i); \mathcal{D}^y_{n+1}), & \text{if } i \in \{1, \ldots, n\}. \end{cases}$$

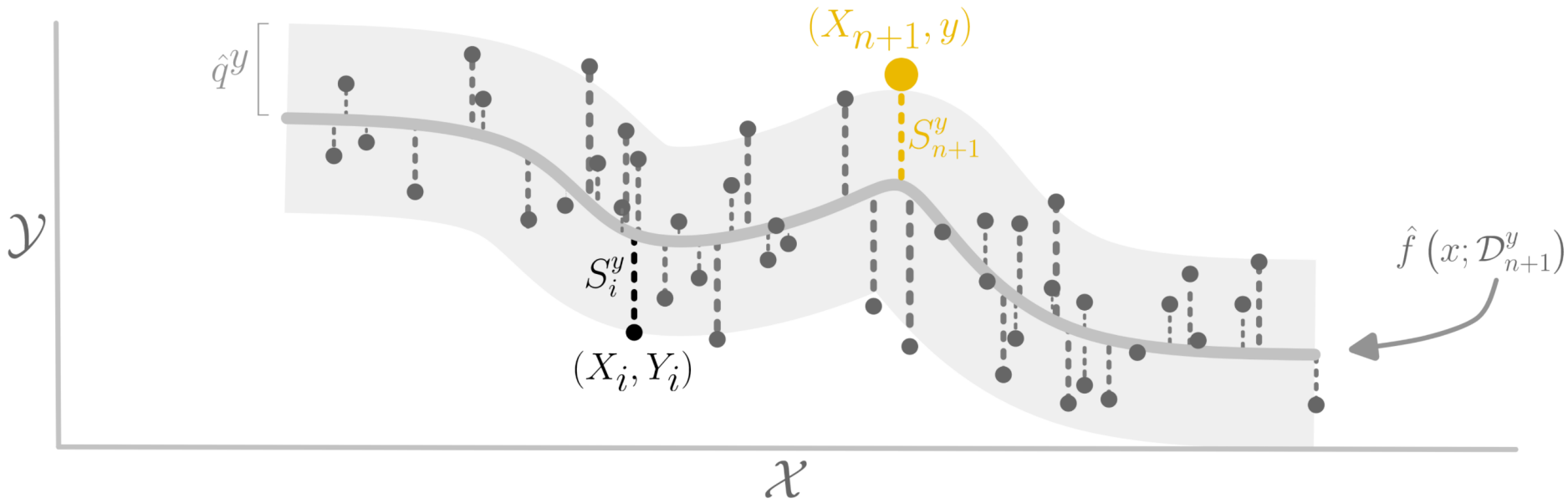


Figure: **Illustration of notation for a single hypothesized response** y. Grey curve is regression model $\hat{f}(x; \mathcal{D}^y_{n+1})$. Grey dots are known data points (X_i, Y_i) . Yellow dot is hypothesized data point (X_{n+1}, y). Dotted lines are residual scores S^y_i. $\hat{q}^y$ is the conformal quantile (defined next).

Full Conformal Prediction Set

- Consider some hypothesized response $y \in \mathcal{Y}$.
- If conformal score S_{n+1}^y is large compared to $S_1^y, \ldots, S_n^y$, then hypothesized response y is **inconsistent with the data** and should be excluded from $\mathcal{C}(X_{n+1})$.
- Construct $\mathcal{C}(X_{n+1})$ by simply taking all y wich **are consistent with the data**, i.e. have sufficiently small scores.
- Formalizing this idea, we get

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \hat{q}^y \right\}$$

with *conformal quantile*

$$\hat{q}^y = \text{Quantile}(S_1^y, \ldots, S_n^y; (1-\alpha)(1+1/n)).$$

- We'll later see that $\mathcal{C}(X_{n+1})$ actually provides correct coverage.

How to Compute the Conformal Prediction Set

- Easy to compute $\hat{q}^y$ for a specific $y \in \mathcal{Y}$.
- How about computing $\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \hat{q}^y \right\}$?
- Discrete $\mathcal{Y}$: Iterate over elements of $\mathcal{Y}$ and add to collection if the score is at most $\hat{q}^y$. ✓
- Continuous $\mathcal{Y}$: More difficult but possible. Ideas:
  1. Exploit model specific properties of e.g. linear regression or LASSO.
  2. Discretize label space $\mathcal{Y}$.
- For details, see section 9.2 of the book.

Marginal Coverage Guarantee of Conformal Prediction

Theorem (Marginal coverage guarantee of conformal prediction)

Suppose that $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ *are exchangeable and that* s *is a symmetric score function. Then the full conformal prediction set* $\mathcal{C}(X_{n+1})$ *satisfies the marginal coverage guarantee*

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

We'll see a proof of this theorem later.

Full Conformal Prediction Algorithm

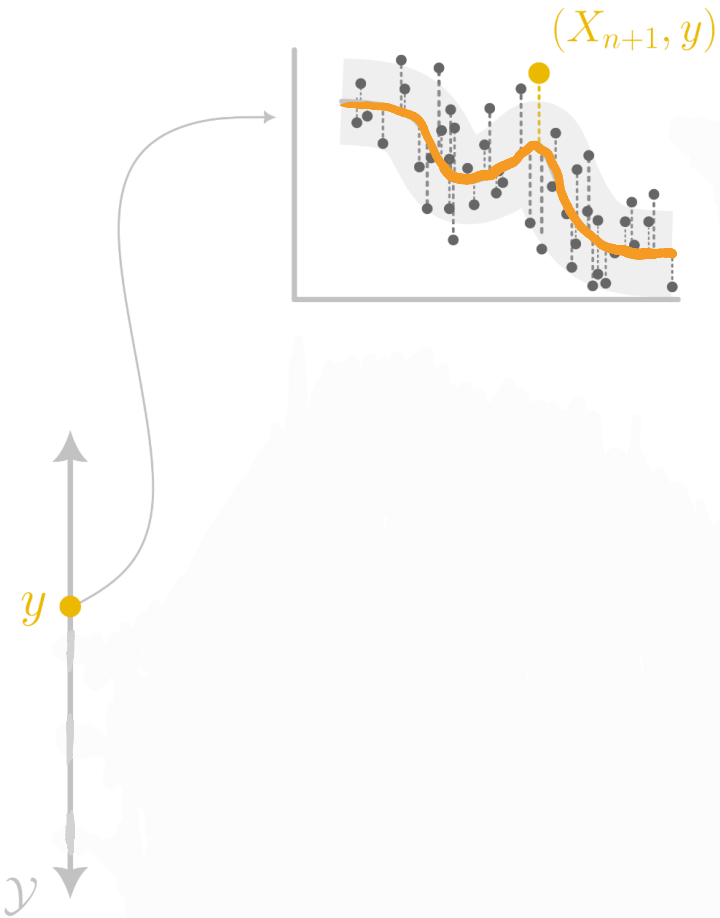
Algorithm 1: Full conformal prediction

Input : training data $(X_1, Y_1), \ldots, (X_n, Y_n)$,
test point X_{n+1},
target coverage level $1 - \alpha$,
conformal score function s

Output: prediction set $\mathcal{C}(X_{n+1})$

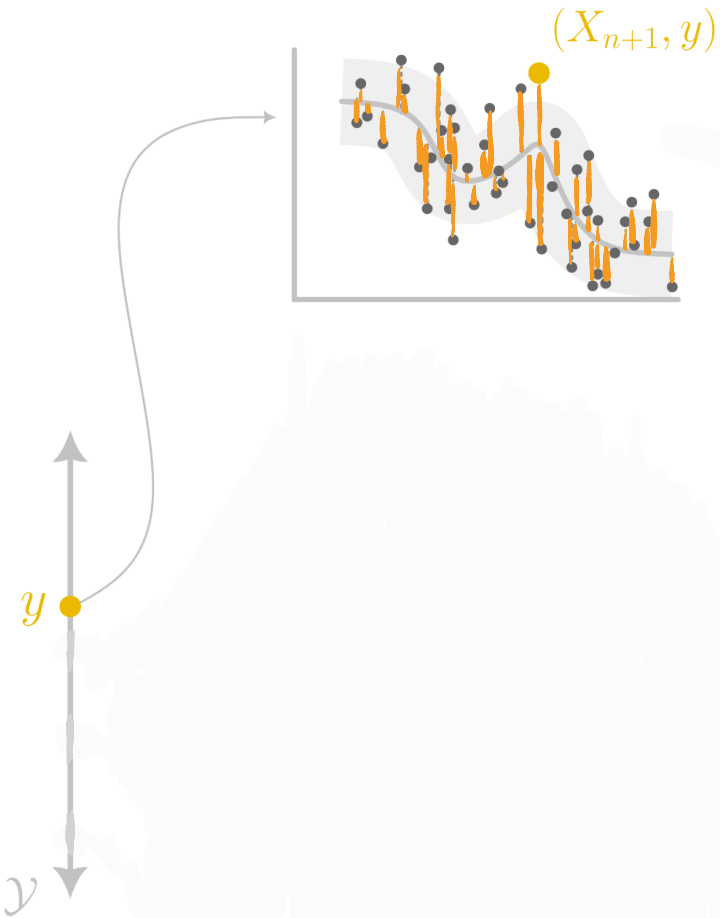
for $y \in \mathcal{Y}$ **do**
  for $i \in \{1, \ldots, n\}$ **do**
    $S_i^y \leftarrow s((X_i, Y_i); \mathcal{D}_{n+1}^y)$;
  $S_{n+1}^y \leftarrow s((X_{n+1}, y); \mathcal{D}_{n+1}^y)$;
  $\hat{q}^y \leftarrow \text{Quantile}(S_1^y, \ldots, S_n^y; (1 - \alpha)(1 + 1/n))$;

return $\left\{y \in \mathcal{Y} : S_{n+1}^y \leq \hat{q}^y\right\}$;



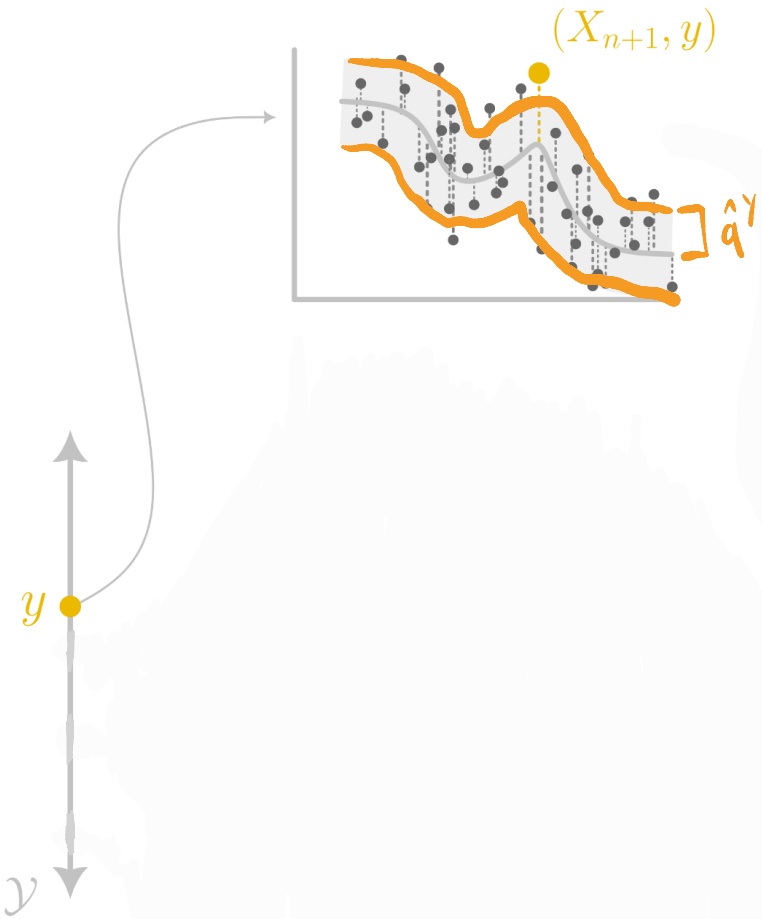
① compute predictive model

$\hat{f}(\cdot\,; D_{n+1}^{y})$

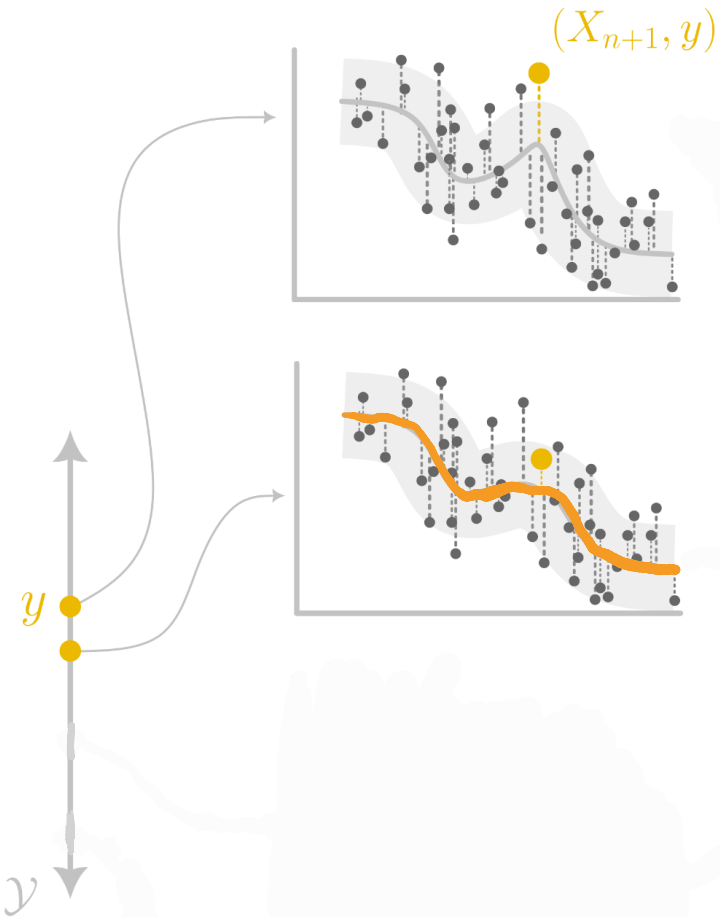


② compute scores

S_i^y

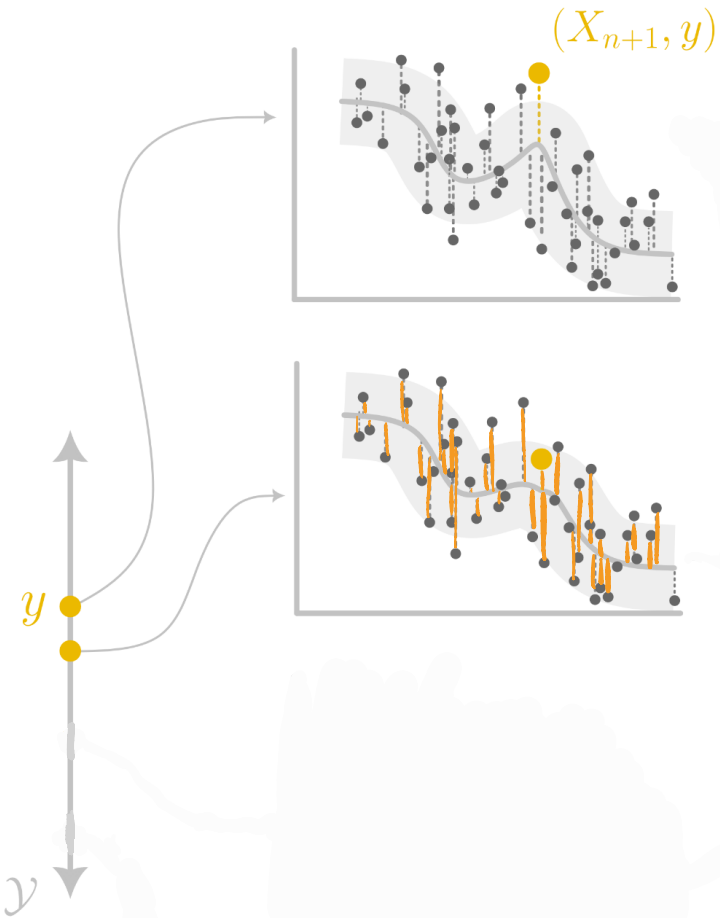


③ compute conformal quantile $\hat{q}^y$



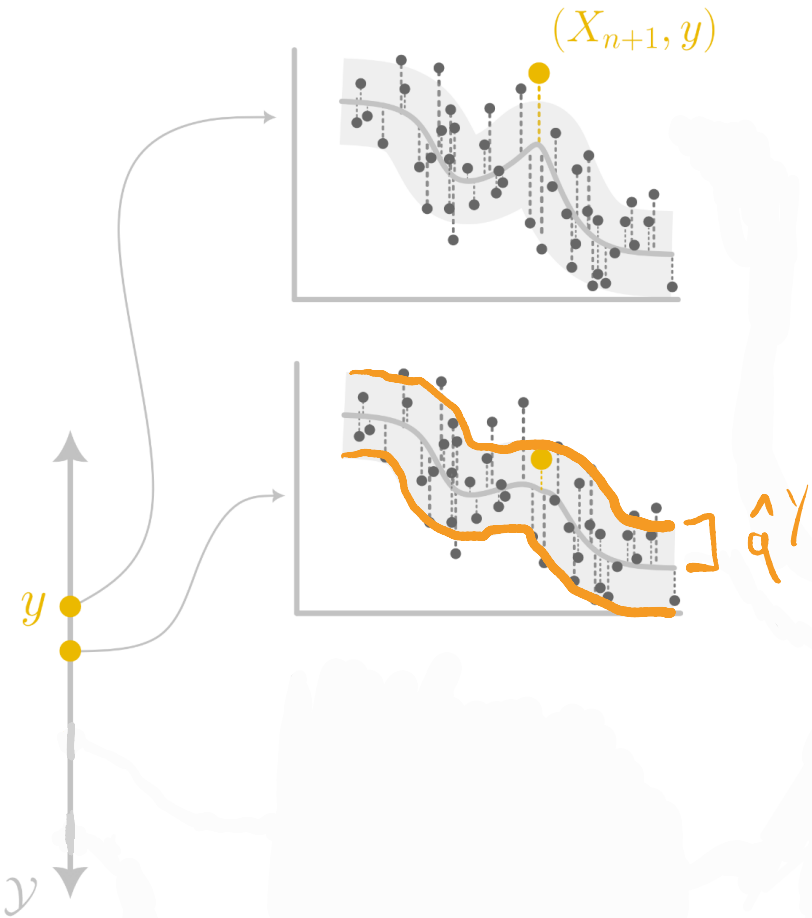
① compute predictive model

$\hat{f}(\cdot\,; D^{y}_{n+1})$

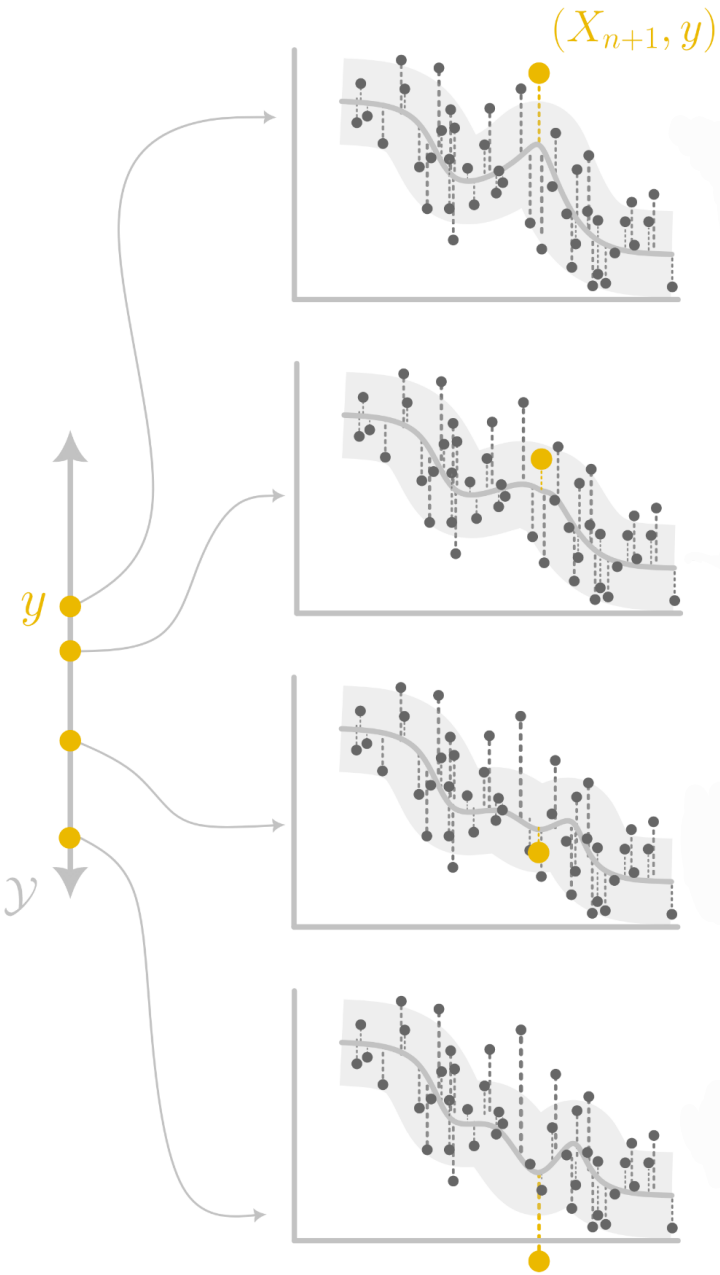


② compute scores

S_i^y

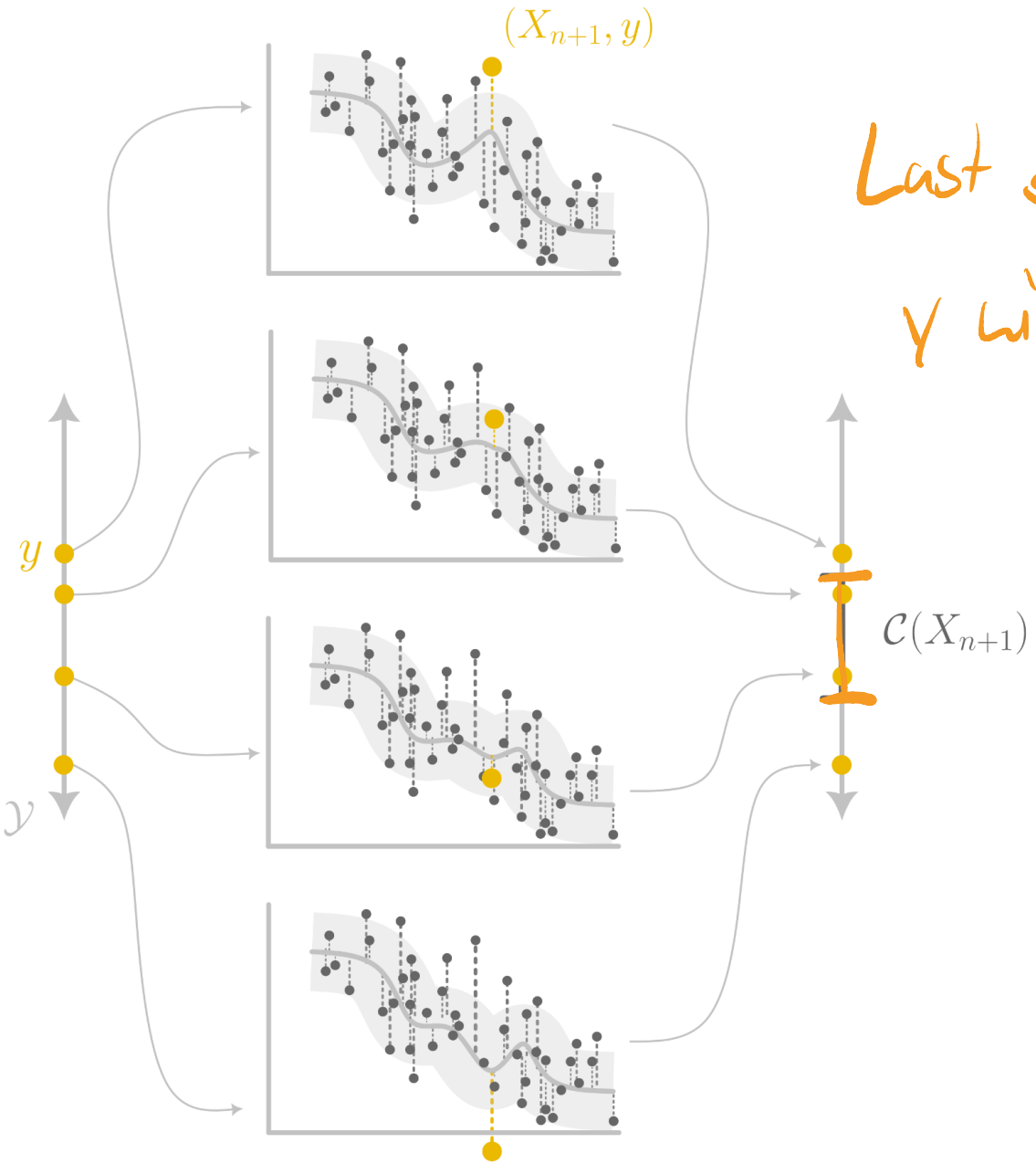


③ compute conformal quantile $\hat{q}^y$



... repeat

1. compute model
2. compute scores
3. compute $\hat{q}^y$



Last step: collect y with $S^y_{n+1} \leq \hat{q}^y$ in $C(X_{n+1})$

Next Up

- Naive approach to marginal coverage.
- Proof of marginal coverage for full conformal prediction.
- Split vs. full conformal prediction.

Naive approach to marginal coverage

If $\mathcal{D}_{n+1} = ((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}))$ is exchangeable, why not proceed as follows:

1. Train a model on $\mathcal{D}_n$ and calculate the score on all the available data pairs: $\{s(X_1, Y_1), \ldots, s(X_n, Y_n)\}$. This set is exchangeable.
2. Let $q^n = \text{Quantile}(s(X_1, Y_1), \ldots s(X_n, Y_n); (1-\alpha)(1+n))$.
3. Define the prediction set $\mathcal{C}(X_{n+1}) := \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq q^n\}$.

One could think that since $\mathcal{D}_{n+1}$ is exchangeable, the scores of $\mathcal{D}_{n+1}$ would also be. In this case, by definition of q^n and exchangeability we would have marginal coverage:

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

Naive approach to marginal coverage

Issue: The score $s((X_{n+1}, Y_{n+1}); \mathcal{D}_n)$ is not exchangeable with the scores $\{s(X_1, Y_1), \ldots, s(X_n, Y_n)\}$.

Since we trained the model on $\mathcal{D}_n$, it might be overfitted to the training data therefore causing the score of the test pair to be higher than the others.

This leads to the $(1 - \alpha)(1 + n)$-Quantile of the scores $\{s(X_1, Y_1), \ldots, s(X_n, Y_n)\}$ being too low to ensure marginal coverage.

Naive Approach to Marginal Coverage

To solve this problem, we consider two options.

1. If we could train our model on the test data as well, that would reinstate exchangeability across all scores and we could get marginal coverage.
   This is the idea behind full conformal prediction: since we do not know Y_{n+1}, we instead train the model on all possible values that Y_{n+1} could take, and so we determine for every $y \in \mathcal{Y}$ a threshold $\hat{q}^y$.
2. Or we have a model and a score function that do not depend on the calibration data and (X_{n+1}, Y_{n+1}). This idea leads to split conformal prediction.

Next Up

- Naive approach to marginal coverage ✓
- Proof of marginal coverage for full conformal prediction.
- Split vs. full conformal prediction.

Lemma 3.4: Replacement Lemma

Lemma

Let $v_1, \ldots, v_{n+1} \in \mathbb{R}$. *Then for any* $t \in [0, 1]$,

$$v_{n+1} \leq \text{Quantile}(v_1, \ldots, v_{n+1}; t)$$
$$\iff v_{n+1} \leq \text{Quantile}(v_1, \ldots, v_n; t(1 + 1/n)).$$

Fact 2.14, Properties under exchangeability

Recall that a vector of random variables $(Z_1, \ldots, Z_n)$ is exchangeable if for every permutation $\sigma \in \mathcal{S}_n$ $(Z_1, \ldots, Z_N) \stackrel{d}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})$

Fact 2.14 ii)

Assume $Z \in \mathbb{R}^n$ is exchangeable, and fix any $i \in [n]$. Then we have that for all $\tau \in [0, 1]$,

$$\mathbb{P}(Z_i \leq \text{Quantile}(Z; \tau)) \geq \tau \quad \text{and, if } \tau > 0, \quad \mathbb{P}(Z_i < \text{Quantile}(Z; \tau)) < \tau.$$

Proof of marginal coverage for full conformal prediction

We will follow these 3 Steps to prove that the full conformal prediction algorithm satisfies marginal coverage:

1. Step 1: Reformuating the prediction set $\mathcal{C}(X_{n+1})$.
2. Step 2: Proving the exchangeability of the scores.
3. Step 3: Proving that marginal coverage holds.

Step 1: Reformulating the prediction set $\mathcal{C}(X_{n+1})$

By definition $\hat{q}^y = \text{Quantile}(S_1^y, \ldots, S_n^y; (1-\alpha)(1+1/n))$ and by direct application of Lemma 3.4 we get:

$$S_{n+1}^y \leq \hat{q}^y \iff S_{n+1}^y \leq \text{Quantile}(S_1^y, \ldots, S_{n+1}^y; 1-\alpha)$$

So

$$y \in \mathcal{C}(X_{n+1}) \iff S_{n+1}^y \leq \text{Quantile}(S_1^y, \ldots, S_{n+1}^y; 1-\alpha).$$

Since this holds for every $y \in \mathcal{Y}$ we can conclude that

$$Y_{n+1} \in \mathcal{C}(x_{n+1}) \iff S_{n+1} \leq \text{Quantile}(S_1, \ldots, S_{n+1}; 1-\alpha).$$

In the next steps we want to show that this holds with probability $1-\alpha$.

Step 2: Exchangeability of the scores

Let σ be a permutation on $\{1, \ldots, n+1\}$. Since the score function is symmetric, we have:

$$S_i = s((X_i, Y_i); \mathcal{D}_{n+1}) = s((X_i, Y_i); (\mathcal{D}_{n+1})_\sigma) \quad \forall i$$
$$S_{\sigma(i)} = s((X_{\sigma(i)}, Y_{\sigma(i)}); \mathcal{D}_{n+1}) = s((X_{\sigma(i)}, Y_{\sigma(i)}); (\mathcal{D}_{n+1})_\sigma) \quad \forall i$$

So for exchangeability we want to prove that

$$(S_1, \ldots, S_{n+1}) \stackrel{d}{=} (S_{\sigma(1)}, \ldots, S_{\sigma(n+1)})$$

which is equivalent to

$$[s((X_i, Y_i); \mathcal{D}_{n+1})]_{i \in [n+1]} \stackrel{d}{=} [s((X_{\sigma(i)}, Y_{\sigma(i)}); (\mathcal{D}_{n+1})_\sigma)]_{i \in [n+1]}$$

Step 2: Exchangeability of the scores

By exchangeability of the data $\mathcal{D}_{n+1} \stackrel{d}{=} (\mathcal{D}_{n+1})_\sigma$.
By applying the same function to these sets we get the following scores:

$$\mathcal{D}_{n+1} \to [s((X_i, Y_i); \mathcal{D}_{n+1})]_{i \in [n+1]} \quad \text{and}$$
$$(\mathcal{D}_{n+1})_\sigma \to [s((X_{\sigma(i)}, Y_{\sigma(i)}); (\mathcal{D}_{n+1})_\sigma)]_{i \in [n+1]}.$$

Thus

$$[s((X_i, Y_i); \mathcal{D}_{n+1})]_{i \in [n+1]} \stackrel{d}{=} [s((X_{\sigma(i)}, Y_{\sigma(i)}); (\mathcal{D}_{n+1})_\sigma)]_{i \in [n+1]}$$

which by the previous slide gives us exchangeability of the $n+1$ scores.

Step 3: Completing the proof

By fact 2.14 ii)

$$\mathbb{P}(S_{n+1} \leq \text{Quantile}(S_1, \ldots, S_{n+1}; \tau)) \geq \tau \quad \forall \tau \in [0, 1]$$

Choosing $\tau = 1 - \alpha$ gives us:

$$\mathbb{P}(S_{n+1} \leq \text{Quantile}(S_1, \ldots, S_{n+1}; 1 - \alpha)) \geq 1 - \alpha$$

which by step one was what we needed to show to complete the proof.

□

Next Up

- Naive approach to marginal coverage ✓
- Proof of marginal coverage for full conformal prediction ✓
- Split vs. full conformal prediction

Split Conformal Prediction as a Special Case

Consider a dataset $\mathcal{D}_{pre}$, called the pretraining set, with $\mathcal{D}_{pre} \cap \mathcal{D}_n = \emptyset$. $\mathcal{D}_{pre}$ is only used for model training while $\mathcal{D}_n$ is used for calculating a threshold to define the prediction set $\mathcal{C}(X_{n+1})$
By design, any score function s constructed on the pretraining set will be symmetric, since s does not depend on $\mathcal{D}_n$.

$$s((X, Y); \mathcal{D}_n) = s(X, Y) = s((X, Y), (\mathcal{D}_n)_\sigma).$$

Example: The residual score for split conformal is:

$$s_{\text{res}}((x, y); \mathcal{D}_n) = |y - \hat{f}(x; \mathcal{D}_{pre})| = s_{\text{res}}(x, y).$$

Split Conformal Prediction as a Special Case

Definition (Definition 3.5)

Split conformal prediction refers to the case where the score function $s((x, y); \mathcal{D})$ does not depend on $\mathcal{D}$. In this case, we write $s(x, y)$ as a shorthand for $s((x, y); \mathcal{D})$

The prediction set can then be constructed as follows:

$$\mathcal{C}(X_{n+1}) := \{y : s(X_{n+1,}, y) \leq \hat{q}\}$$
$$\hat{q} := \text{Quantile}(S_1, \ldots, S_n; (1 - \alpha)(1 + 1/n))$$

Remark: Note that $\hat{q}$ is equivalent to $\hat{q}^y$ (the threshold defined for full conformal prediction) for all y in $\mathcal{Y}$ since the score does not depend on any other data than $\mathcal{D}_{pre}$.

Algorithm for Split Conformal Prediction

Algorithm for Split Conformal

1. Input $\mathcal{D}_{pre}$, $(X_1, Y_1), \ldots, (X_n, Y_n)$, X_{n+1}, α.
2. Using the pretraining dataset $\mathcal{D}_{pre}$ construct a conformal score function $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$.
3. Compute the conformal scores $S_i = s(X_i, Y_i)$ on the calibration set $\mathcal{D}$.
4. Compute the quantile $\hat{q} = \text{Quantile}(S_1, \ldots, S_n; (1-\alpha)(1+1/n))$.
5. Return the prediction set $\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1},, y) \leq \hat{q}\}$

Since this is a special case of Algorithm 3.3, the marginal coverage guarantee still holds:

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

Conditions for Split Conformal

The most straightforward condition to justify the split conformal prediction is to have $\mathcal{D}_{pre}$ independent of the calibration data and the test pair, and $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ exchangeable.

- This allows us to view the score function as a fixed function.
- Under these conditions we have: $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ exchangeable $\Rightarrow$ $s(X_1, Y_1), \ldots, s(X_{n+1}, Y_{n+1})$ exchangeable.

But a weaker conditions would also suffice:

$$((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}))|\mathcal{D}_{pre} \quad \text{is exchangeable.}$$

This means that $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ has an exchangeable conditional distribution when conditioning on $\mathcal{D}_{pre}$.
Example: This holds when the entire dataset $(\mathcal{D}_{pre}, D_n, (X_{n+1}, Y_{n+1}))$ is exchangeable.

Differences between Full and Split Conformal

Full Conformal	**Split Conformal**
• All data is used for both training and calibration • Retrains the model for each value $y \in \mathcal{Y}$ • Requires a symmetric score function s	• Disjoint datasets for training and calibration • No model retraining (requires only one model fit) • Works for any (pretrained) score function s

Statistical vs. Computational Efficiency

Split conformal predictions are more computationally efficient.

- They only require the model $\hat{f}$ to be trained once.
- For full conformal predictions, $\hat{f}$ needs to be retrained for all $y \in \mathcal{Y}$.

Full conformal predictions are more statistically efficient.

- All of the data available is used to train $\hat{f}$ and determine $\mathcal{C}(X_{n+1})$.
- With the split conformal method only part of the data can be used for training and the other part for calibrating. In general, this leads to a bigger prediction set $\mathcal{C}(X_{n+1})$.

Next Up

- Naive approach to marginal coverage ✓
- Proof of marginal coverage for full conformal prediction ✓
- Split vs. full conformal prediction ✓

Last part

1. Permutation tests & exchangeability testing.
2. Example: testing if a new data point is an outlier.
3. Conformal prediction as a permutation test.
4. Optionally: tuning based on a plug-in estimate of the error rate.
5. Can conformal predictions be overly conservative?

Permutation tests

Often, a permutation test can be expressed as testing the null hypothesis of exchangeability.

Exchangeability — Reminder

Let $Z_1, ..., Z_n \in \mathcal{Z}$ be random variables with a joint distribution. We say that the random vector $(Z_1, ..., Z_n)$ is *exchangeable* if

$$(Z_1, ..., Z_n) \stackrel{d}{=} (Z_{\sigma(1)}, ..., Z_{\sigma(n)}) \quad \forall \sigma \in S_n.$$

- $\mathcal{P}$ is the set of all distributions on $\mathcal{Z}^n$.
- $\mathcal{P}_{\text{exch}} \subseteq \mathcal{P}$ is the subset of distributions for which exchangeability satisfied.
- $(Z_1, ..., Z_n) \sim P$ (P is some joint distribution).
- Hypothesis test of

$$H_0 : P \in \mathcal{P}_{\text{exch}} \quad \text{vs.} \quad H_1 : P \in \mathcal{P} \backslash \mathcal{P}_{\text{exch}}.$$

Exchangeability test — Procedure

- Before observing the data, fix any function $T : \mathcal{Z}^n \to \mathbb{R}$. *Intuition*: a large value of a test statistic $T(Z_1, ..., Z_n)$ indicates the evidence against exchangeability.

$$p = \frac{\sum_{\sigma \in S_n} \mathbb{I}\{T(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}) \geq T(Z_1, \ldots, Z_n)\}}{n!}$$

- Compares test statistic T of original data with all possible data **permutations**.
- Measures how extreme the observed statistic is **under the null hypothesis**.
- **Statement:** p is a valid p-value for a permutation test.

Exchangeability test — Procedure

- To avoid the computational burden of computing all $n!$ permutations, sample with replacement a smaller number of permutations uniformly at random, to obtain the p-value

$$p = \frac{1 + \sum_{m=1}^{M} \mathbb{I}\{T(Z_{\sigma_m(1)}, \ldots, Z_{\sigma_m(n)}) \geq T(Z_1, \ldots, Z_n)\}}{1 + M}.$$

- The term "$1 +$" makes sure the event $p = 0$ doesn't have a nonzero probability under the null hypothesis.
- It's a valid p-value.

Example — Testing if a new data point is an outlier

- **Goal**: test whether the last data point Z_n is an outlier relative to the rest of the sequence. Consider the test statistic:

$$T(z_1, ..., z_n) = \sum_{i=1}^{n} \mathbb{I}\{z_n > z_i\}.$$

- $T(z_1, ..., z_n)$ captures whether Z_n is more likely to be unusually large relative to the other Z_i's.
- The permutation test p-value can be simplified because

$$T(Z_{\sigma(1)}, ..., Z_{\sigma(n)}) = \sum_{i=1}^{n} \mathbb{I}\{Z_{\sigma(n)} > Z_{\sigma(i)}\} = \sum_{i=1}^{n} \mathbb{I}\{Z_{\sigma(n)} > Z_i\}$$

$$T(Z_{\sigma(1)}, ..., Z_{\sigma(n)}) \geq T(Z_1, ..., Z_n) \iff Z_{\sigma(n)} \geq Z_n$$

and so

$$p = \frac{\sum_{\sigma \in S_n} \mathbb{I}\{T(Z_{\sigma(1)}, ..., Z_{\sigma(n)}) \geq T(Z_1, ..., Z_n)}{n!} = \frac{1}{n!} \sum_{\sigma \in S_n} \mathbb{I}\{Z_{\sigma(n)} \geq Z_n\}$$

$$= \frac{1}{n!} \sum_{i=1}^{n} \sum_{\sigma \in S_n, \sigma(n)=i} \mathbb{I}\{Z_i \geq Z_n\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i \geq Z_n\}.$$

Example — continuation

- The last step holds true $\forall i \in [n]$ since there are $(n-1)! = \frac{n!}{n}$ permutations $\sigma \in S_n$ for which $\sigma(n) = i$.
- So,

$$p = \frac{\sum_{i=1}^{n} \mathbb{I}_{Z_i \geq Z_n}}{n}$$

is a valid p-value under the assumption that $Z_1, \ldots, Z_n$ are exchangeable, i.e., $\mathbb{P}(p \leq \tau) \leq \tau \quad \forall \tau \in [0, 1]$.

Conformal prediction as a permutation test

Idea: conformal prediction can be viewed as a test of whether (X_{n+1}, y) is an outlier to the other data points $(X_1, Y_1), ..., (X_n, Y_n)$.

The conformal p-value

Given training data $(X_1, Y_1), ..., (X_n, Y_n)$, a test feature X_{n+1}, and a score function s, the conformal p-value is defined as

$$p^y = \frac{1 + \sum_{i=1}^{n} \mathbb{I}\{S_i^y \geq S_{n+1}^y\}}{n+1}$$

for each $y \in \mathcal{Y}$, where as before,

$$S_i^y = s((X_i, Y_i); \mathcal{D}_{n+1}^y), \quad i \in [n]$$
$$S_{n+1}^y = s((X_{n+1}, y); \mathcal{D}_{n+1}^y)$$

Intuition

- **Intuition**: this p-value is asking whether the hypothesized test point (X_{n+1}, y) appears to follow the same distribution as the training data $(X_1, Y_1), ..., (X_n, Y_n)$ — if not, its score $s((X_{n+1}, y); \mathcal{D}^y_{n+1})$ might be substantially larger that the other scores, and consequently its p-value p^y will likely be small.
- **Interpretation**: p^y is a p-value of testing exchangeability of the points $(X_1, Y_1), ..., (X_n, Y_n), (X_{n+1}, y)$.
- **Algorithm**: we can repeat this reasoning for every possible $y \in \mathcal{Y}$, and collect the plausible values (i.e., y for which p^y is not too small) into a prediction set.

Conformal prediction as a permutation test

Before, we defined the **full conformal prediction set** $\mathcal{C}(X_{n+1})$ as

$$\mathcal{C}(X_{n+1}) = \{y : S^y_{n+1} \leq \hat{q}^y\},$$

where

$$\hat{q}^y = \text{Quantile}(S^y_1, ..., S^y_n; (1-\alpha)(1+1/n)).$$

Proposition — equivalence of full conformal prediction set definitions

The full conformal prediction set satisfies

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : p^y > \alpha\}.$$

Marginal coverage guarantee of conformal prediction can be also proved using this interpretation.

Marginal coverage guarantee of conformal prediction — Reminder

Suppose that $(X_1, Y_1), ..., (X_{n+1}, Y_{n+1})$ are exchangeable and that s is a symmetric score function. Then, the prediction set $\mathcal{C}(X_{n+1})$ satisfies the marginal coverage guarantee,

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

Marginal coverage guarantee — another proof

- Denote $Z_i = (X_i, Y_i)$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- Given any test function $T : \mathcal{Z}^{n+1} \to \mathbb{R}$, permutation test p-value:

$$p_{\text{perm}} = \frac{\sum_{\sigma \in S_{n+1}} \mathbb{I}\left\{T(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) \geq T(Z_1, \ldots, Z_{n+1})\right\}}{(n+1)!}$$

- To complete the proof, we verify:

$$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff p_{\text{perm}} > \alpha$$

when the test function T is chosen appropriately. We set:

$$T(z_1, \ldots, z_{n+1}) = s(z_{n+1}; z_1, \ldots, z_n)$$

- By Proposition, it's enough to show $p_{\text{perm}} = p^{Y_{n+1}}$. Then,

$$T(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) = s(Z_{\sigma(n+1)}; \mathcal{D}^{\sigma}_{n+1}) = s(Z_{\sigma(n+1)}; \mathcal{D}_{n+1}) = S_{\sigma(n+1)}$$

$$p_{\text{perm}} = \frac{1}{(n+1)!} \sum_{\sigma \in S_{n+1}} \mathbb{I}\{S_{\sigma(n+1)} \geq S_{n+1}\} = \sum_{i=1}^{n+1} \frac{n!}{(n+1)!} \mathbb{I}\{S_i \geq S_{n+1}\}$$

$$= \frac{1 + \sum_{i=1}^{n} \mathbb{I}\{S_i \geq S_{n+1}\}}{n+1} = p^{Y_{n+1}} \quad (S_i = S_i^{Y_{n+1}} \text{ by def})$$

Tuning Based on a Plug-in Estimate of the Error Rate

Consider prediction sets of the form

$$\mathcal{C}(X_{n+1}; \lambda) = \{y : s(X_{n+1}, y) \leq \lambda\},$$

where $\lambda \in \mathbb{R}$ is a parameter controlling the size of the set. Why we choose $\lambda = \hat{q} = \text{Quantile}(S_1, ..., S_n; (1-\alpha)(1+1/n))$ for split conformal prediction?

- Goal: choose threshold λ to control error rate in prediction set.
- We estimate the miscoverage rate on the calibration set, where $S_i = s(X_i, Y_i)$:

$$\widehat{R}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{S_i > \lambda\}.$$

- Naively choosing λ such that $\widehat{R}(\lambda) \leq \alpha$ may not guarantee coverage ($\hat{R}(\hat{\lambda})$ is a noisy estimate of this true risk).
- Instead, define a more conservative threshold:

Adjusted threshold

$$\hat{\lambda} = \inf\left\{\lambda \in \mathbb{R} : \widehat{R}(\lambda) \leq \alpha'\right\}, \quad \text{where } \alpha' = \alpha - \frac{1-\alpha}{n}$$

Tuning Based on a Plug-in Estimate of the Error Rate — result

Split conformal prediction set

The split conformal prediction set $\mathcal{C}(X_{n+1})$ satisfies

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{\lambda}\}$$

Can conformal prediction be overly conservative?

- We discussed that the prediction sets will cover the ground truth with *at least* probability $1 - \alpha$.
- **Issue**: the sets can be unnecessarily large. How much the coverage probability could potentially exceed the target level $1 - \alpha$?

Probability upper bound

Under the conditions of Marginal coverage guarantee, we have that

$$\begin{aligned}\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) &\leq \frac{[(1-\alpha)(n+1)]}{n+1} + \epsilon_{\text{tie}} \\ &\leq 1 - \alpha + \frac{1}{n+1} + \epsilon_{\text{tie}},\end{aligned}$$

where ϵ_{tie} captures the likelihood of the score of the $(n+1)$-st data point being tied with any other data point:

$$\epsilon_{\text{tie}} = \mathbb{P}(\exists j \in [n] \colon S_{n+1} = S_j).$$

Can conformal prediction be overly conservative? — Conclusions

- The coverage of conformal prediction isn't too far from $1-\alpha$ as long as the distribution of the scores is unlikely to produce ties.
- This doesn't directly translate to a bound on the size of prediction set $\mathcal{C}(X_{n+1})$.
- But says that this set isn't conservative on the scale of coverage.
- For example, for $s((x,y);\mathcal{D})) = |y - \hat{f}(x;\mathcal{D})|$, a model $\hat{f}(\cdot;\mathcal{D})$ that is a very poor fit to the data distribution, will necessarily lead to wide prediction intervals. This result tells us the prediction intervals are no wider than is needed to compensate for the errors in $\hat{f}(\cdot;\mathcal{D})$.

If the joint distribution of the scores is continuous

Under the conditions of Marginal coverage guarantee, further assume that the scores $S_1, ..., S_{n+1}$ have a continuous joint distribution. Then,

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

Thank you!

Questions or Comments?

Split Conformal Prediction, Special Case

Algorithm 2: Split conformal prediction, special case

Input : data (X_i, Y_i) for $i = 1, \ldots, n$ with even $n \geq 2$,
test point X_{n+1},
target coverage level $1 - \alpha$

Output: prediction interval $\mathcal{C}(X_{n+1})$

Partition the indices $\{1, \ldots, n\}$ into a training set $\{1, \ldots, n/2\}$ and a calibration set $\{n/2 + 1, \ldots, n\}$.;

Fit predictive model $\hat{f} : \mathcal{X} \to \mathbb{R}$ on $\{(X_i, Y_i) : i = 1, \ldots, n/2\}$;

for $i = n/2 + 1, \ldots, n$ **do**

  $S_i \leftarrow \left| Y_i - \hat{f}(X_i) \right|$;

$Q \leftarrow \texttt{sort\_ascending}(S_n, \ldots S_{n/2+1})$;

$\hat{q} \leftarrow Q_{\lceil (1-\alpha)(n/2+1) \rceil}$;

return $\mathcal{C}(X_{n+1}) = \left[\hat{f}(X_{n+1}) - \hat{q},\ \hat{f}(X_{n+1}) + \hat{q} \right]$;

Theorem (Split conformal coverage guarantee, special case)

Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ *are i.i.d., and let* $\mathcal{C}(X_{n+1})$ *be the output of Algorithm 2. Then*

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

Split Conformal Prediction, General Case

Algorithm 3: Split conformal prediction, general case

Input : data (X_i, Y_i) for $i = 1, \ldots, n$ with even $n \geq 2$,
test point X_{n+1},
target coverage level $1 - \alpha$

Output: prediction set $\mathcal{C}(X_{n+1})$

Partition the indices $\{1, \ldots, n\}$ into a training set $\{1, \ldots, n/2\}$ and a calibration set $\{n/2 + 1, \ldots, n\}$.;

Use the training set to construct a conformal score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$;

for $i = n/2 + 1, \ldots, n$ **do**

  $S_i \leftarrow s(X_i, Y_i)$;

$Q \leftarrow \texttt{sort\_ascending}(S_n, \ldots S_{n/2+1})$;

$\hat{q} \leftarrow Q_{\lceil (1-\alpha)(n/2+1) \rceil}$;

return $\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq \hat{q} \right\}$;

Theorem (Split conformal coverage guarantee, general case)

Suppose $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ *are i.i.d., and let* $\mathcal{C}(X_{n+1})$ *be the output of Algorithm 3. Then*

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

Symmetry of the Joint Density or PMF

- **Characterization (discrete case)**: Let $\mathcal{Z}$ be a countable space and p be a joint PMF over $\mathcal{Z}^n$. $(Z_1, \ldots, Z_n)$ is exchangeable if and only if for all $\sigma \in S_n$ and $z_1, \ldots, z_n \in \mathcal{Z}$:

$$p(z_1, \ldots, z_n) = p\big(z_{\sigma(1)}, \ldots, z_{\sigma(n)}\big).$$

- **Characterization (continuous case):** Analogous. Let f be a joint density over $\mathbb{R}^n$. $(Z_1, \ldots, Z_n)$ is exchangeable if and only if for all $\sigma \in S_n$ and *almost every* $z_1, \ldots, z_n \in \mathcal{Z}$:

$$f(z_1, \ldots, z_n) = f\big(z_{\sigma(1)}, \ldots, z_{\sigma(n)}\big).$$

Conditioning on the Order Statistics

- Assume real-valued RVs, i.e. $\mathcal{Z} = \mathbb{R}$ and order statistics $Z_{(1)} \leq \cdots \leq Z_{(n)}$.
- If **all values are distinct a.s.** then every permutation is equally likely, i.e. has occurence probability $\frac{1}{n!}$.
- **Characterization**: If values are not distinct a.s., then

$$(Z_1, \ldots, Z_n) \mid \{Z_{(1)}, \ldots, Z_{(n)}\} \sim \frac{1}{n!} \sum_{\sigma \in S_n} \delta_{(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})}.$$

- Sum counts the number of permutations that produce $(Z_1, \ldots, Z_n)$.
- Example: Unordered collection is $\{1, 2, 2\}$, then correct probabilities are $1/3$ (not $1/6$) since 2 permutations can produce each possible sequence.
- Fact: $\mathbb{P}\big(Z_i \leq Z_{(k)}\big) = k/n$ for each index $i \in [n]$ and rank $k \in [n]$.

Conditioning on the Empirical Distribution

- Define the empirical measure

$$\widehat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i}.$$

- **Characterization**: $(Z_1, \ldots, Z_n)$ is exchangeable if and only if conditional on $\widehat{P}_n$, the variables $Z_1, \ldots, Z_n$ have common distribution $\widehat{P}_n$, i.e. for all $i \in [n]$,

$$Z_i \mid \widehat{P}_n \sim \widehat{P}_n.$$

- Can be extended beyond $\mathcal{Z} = \mathbb{R}$.

Residual Score

- If we use the residual score, symmetry of s is the same as saying that **learning algorithm** is invariant to the data point order as well.
- Easy to see. (1) Use the definition of the residual score, (2) choose $y = \hat{f}(x; \mathcal{D})$:

$$
\begin{aligned}
& s((x,y);\mathcal{D}) = s((x,y);\mathcal{D}_\sigma) && \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \\
\overset{(1)}{\iff} & |y - \hat{f}(x;\mathcal{D})| = |y - \hat{f}(x;\mathcal{D}_\sigma)| && \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \\
\overset{(2)}{\iff} & \hat{f}(x;\mathcal{D}) = \hat{f}(x;\mathcal{D}_\sigma) && \forall x \in \mathcal{X}.
\end{aligned}
$$