Aalto University

A practical introduction to conformal prediction for signal processing

Esa Ollila

Department of Information and Communications Engineering, Aalto University, Finland

Laboratoire TéSa Seminar, Oct. 20th, 2026

Outline

1. **Motivation**
2. **Conformal prediction**
   - Aims of conformal prediction
   - Key idea: order statistics and exchangeability
3. **Two conformal prediction schemes**
   - Full conformal prediction
   - Split conformal prediction
4. **SCP for regression**
   - Studentized split conformal regression
   - Conformalized quantile regression (CQP)
5. **SCP for classification**
6. **Online conformal prediction**

Motivation

A well-calibrated AI *knows when it knows and knows when it does not know*.

- ML models typically output a **hard decision** (a class with largest score), and sometimes a poorly calibrated confidence level $\Rightarrow$ often over- or under-confident predictions.

Motivation

A well-calibrated AI *knows when it knows and knows when it does not know.*

- ML models typically output a **hard decision** (a class with largest score), and sometimes a poorly calibrated confidence level $\Rightarrow$ often over- or under-confident predictions.

Uncertainty quantification asks ML model not just **point predictions**, but also a confidence band with coverage guarantees

- Difficult task for data-driven ML models (no distributional assumptions!)
- **Conformal prediction (CP)** [VGS05, LW14, PPVG02] has attained broad interest:
  - can provide reliable (exact finite sample and distribution free) coverage guarantees.
  - can be used as a wrapper to any prediction method.
- Many variations/extensions: full CP, split CP, online CP. and many applications [AB$^+$23]

Motivation

A well-calibrated AI *knows when it knows and knows when it does not know.*

- ML models typically output a **hard decision** (a class with largest score), and sometimes a poorly calibrated confidence level $\Rightarrow$ often over- or under-confident predictions.

Uncertainty quantification asks ML model not just **point predictions**, but also a confidence band with coverage guarantees

- Difficult task for data-driven ML models (no distributional assumptions!)
- **Conformal prediction (CP)** [VGS05, LW14, PPVG02] has attained broad interest:
  - can provide reliable (exact finite sample and distribution free) coverage guarantees.
  - can be used as a wrapper to any prediction method.
- Many variations/extensions: full CP, split CP, online CP. and many applications [AB$^+$23]

$\Rightarrow$ *This talk: how CP can be used in practise and in some SP applications*

From point predictions to prediction sets

Figure from [ABJM21]



Conformal prediction ensures **valid coverage**: the true label or value lies inside the set with a user-specified probability (e.g., 90%), *regardless of the model*.

Conformal prediction: goal

- $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$ i.i.d. data from distribution P on $\mathcal{X} \times \mathcal{Y}$.
- Feature space $\mathcal{X} = \mathbb{R}^p$. Regression $\mathcal{Y} = \mathbb{R}$. Classification: $y_i \in \mathcal{Y} = \{1, \ldots, K\}$.

Given a desired coverage rate $1 - \alpha \in (0, 1)$, construct a **prediction band/set**

$$\widehat{C} : \mathcal{X} \to \{\text{subsets of } \mathcal{Y}\},$$

s.t. for new i.i.d. point $(\mathbf{x}_{\text{test}}, y_{\text{test}})$ from P,

$$\Pr\big(y_{\text{test}} \in \widehat{C}(\mathbf{x}_{\text{test}})\big) \geq 1 - \alpha,$$

$$\text{or, } \Pr\big(y_{\text{test}} \in \widehat{C}(\mathbf{x}_{\text{test}})\big) \in \Big[1 - \alpha,\, 1 - \alpha + \frac{1}{n+1}\Big).$$

where the probability is taken over all data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \cup (\mathbf{x}_{\text{test}}, y_{\text{test}})$.

Model (and method) agnostic

We require $\Pr\big(y_{\text{test}} \in \widehat{C}(\mathbf{x}_{\text{test}})\big) \geq 1 - \alpha$, to hold

- with no assumptions whatsoever on the common distribution P
- for any algorithm (e.g., random forest regression/classifier), and for any tuning parameter value it may use.

Some remarks:

- the width of the confidence band will depend on the method used!
- strategy should adapt to the hardness of the problem: more easily we can predict y_{test} from $\mathbf{x}_{\text{test}}$, the smaller the set $\widehat{C}(\mathbf{x}_{\text{test}})$.

Model (and method) agnostic

We require $\Pr\big(y_{\text{test}} \in \widehat{C}(\mathbf{x}_{\text{test}})\big) \geq 1 - \alpha$, to hold

- with no assumptions whatsoever on the common distribution P
- for any algorithm (e.g., random forest regression/classifier), and for any tuning parameter value it may use.

Some remarks:

- the width of the confidence band will depend on the method used!
- strategy should adapt to the hardness of the problem: more easily we can predict y_{test} from $\mathbf{x}_{\text{test}}$, the smaller the set $\widehat{C}(\mathbf{x}_{\text{test}})$.

This all seems like a hard goal!

Prediction band via order statistics

- Consider the simplest setting: no features, we only observe $y_1, \ldots, y_n \in \mathbb{R}$
- Let find confidence band $\hat{C}$ s.t. $\Pr\big(y_{\text{test}} \in \widehat{C}\big) \geq 1 - \alpha$ with earlier goals in mind!

Prediction band via order statistics

- Consider the simplest setting: no features, we only observe $y_1, \ldots, y_n \in \mathbb{R}$
- Let find confidence band $\hat{C}$ s.t. $\Pr\big(y_{\text{test}} \in \widehat{C}\big) \geq 1 - \alpha$ with earlier goals in mind!
- As y_{test} is i.i.d. with $y_1, \ldots, y_n$, its ranking is uniformly distributed over $y_1, \ldots, y_n, y_{\text{test}}$.
- Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n+1)}$ be the ordered values of $y_1, \ldots, y_n, y_{\text{test}}$

Prediction band via order statistics

- Consider the simplest setting: no features, we only observe $y_1, \ldots, y_n \in \mathbb{R}$
- Let find confidence band $\hat{C}$ s.t. $\Pr\big(y_{\text{test}} \in \widehat{C}\big) \geq 1 - \alpha$ with earlier goals in mind!
- As y_{test} is i.i.d. with $y_1, \ldots, y_n$, its ranking is uniformly distributed over $y_1, \ldots, y_n, y_{\text{test}}$.
- Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n+1)}$ be the ordered values of $y_1, \ldots, y_n, y_{\text{test}}$
- Then $\Pr(y_{(k)} = y_{\text{test}}) = \frac{1}{n+1}$ for any $k \in \{1, \ldots, n+1\}$, and

$$\Pr\Big(y_{\text{test}} \text{ is among the } k = \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } y_1, \ldots, y_n, y_{\text{test}}\Big) \geq 1 - \alpha, \quad (1)$$



Prediction band via order statistics (2)

- But (1) is equivalent to

$$\Pr\Big(y_{\text{test}} \text{ is among the } k = \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } y_1, \ldots, y_n\Big) \geq 1 - \alpha.$$

  which is computable from *just the training data*.

- Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ denote the ordered values of $y_1, \ldots, y_n$, and set

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

- Then, defining $\hat{C} = (-\infty, y_{(k)}]$, it holds:

$$\Pr\left(y_{\text{test}} \in \hat{C}\right) = \Pr(y_{\text{test}} \leq y_{(k)}) \geq 1 - \alpha \;\Leftrightarrow \Pr\left(y_{\text{test}} \leq \hat{q}_n\right) \geq 1 - \alpha.$$

$$\hat{q}_n = y_{(k)} := \frac{k}{n} \textbf{ sample quantile } = k^{\text{th}} \text{ smallest of } y_1, \ldots, y_n.$$

Prediction band via order statistics (3)

- If there are almost surely no ties between $y_1, \ldots, y_n, y_{\text{test}}$ (or one use a tie-breaking rule):

$$\begin{aligned}&\Pr\Big(y_{\text{test}} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } y_1, \ldots, y_n, y_{\text{test}}\Big) \\ &= \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} < (1-\alpha) + \frac{1}{n+1}.\end{aligned}$$

- Which leads to the sharpened conclusion:

$$\Pr\left(y_{\text{test}} \leq \hat{q}_n\right) \in \left[1-\alpha,\, 1-\alpha+\frac{1}{n+1}\right).$$

where the lower bound always holds, and the upper bound holds if there are a.s. no ties.

Necessary assumption: exchangeability

- Key property: y_{test} can occupy *any* of the values of the ordered sample.
- For it to hold, we only need that $y_1, \ldots, y_n, y_{\text{test}}$ are exchangeable (weaker than i.i.d. assumption)

Exchangeability

A sample $y_1, \ldots, y_n$ is exchangeable if the joint distribution is unchanged under permutations:

$$(y_1, \ldots, y_n) =_d (y_{\sigma(1)}, \ldots, y_{\sigma(n)}), \quad \text{for all permutations } \sigma,$$

where $=_d$ reads "has the same distribution as".

Ful conformal prediction

Setting:
Training data: $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a test point $(\mathbf{x}_{\text{test}}, y) \in \mathcal{X}$, exchangeable, where $y \in \mathcal{Y}$ is a candidate reponse for $\mathbf{x}_{\text{test}}$.

Augmented data:
$\mathcal{D}^y = \mathcal{D} \cup \{(\mathbf{x}_{\text{test}}, y)\}$

Goal
Find $\hat{C}(\mathbf{x}_{\text{test}})$ s.t. $\Pr(y \in \hat{C}(\mathbf{x}_{\text{test}})) \geq 1 - \alpha$

Symmetry of the learning algorithm:
The model used to fit a predictor function $\hat{f}$ needs to be a symmetric function of the data, i.e., does not use knowledge of the order in which the data points were passed.

Nonconformity score function

- Define a negatively oriented **nonconformity score** function

$$S(\mathbf{x}, y) = S\big((\mathbf{x}, y); \hat{f}\big) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

  which assigns a score to the point $(\mathbf{x}, y)$ based on the predictor function $\hat{f} \colon \mathcal{X} \to \mathcal{Y}$

- Negatively-oriented $= S((\mathbf{x}, y); \hat{f})$ high when the model is boldly wrong

Nonconformity score function

- Define a negatively oriented **nonconformity score** function

$$S(\mathbf{x}, y) = S\big((\mathbf{x}, y); \hat{f}\big) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

  which assigns a score to the point $(\mathbf{x}, y)$ based on the predictor function $\hat{f} \colon \mathcal{X} \to \mathcal{Y}$

- Negatively-oriented $= S((\mathbf{x}, y); \hat{f})$ high when the model is boldly wrong
- **Example (regression):**

$$S(\mathbf{x}, y) = |y - \hat{f}(\mathbf{x})|$$

  where $\hat{f}$ can be any regressor, e.g., random forest regression, fit on(augmented) data $\mathcal{D}^y$.

Nonconformity score function

- Define a negatively oriented **nonconformity score** function

$$S(\mathbf{x}, y) = S\big((\mathbf{x}, y); \hat{f}\big) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

  which assigns a score to the point $(\mathbf{x}, y)$ based on the predictor function $\hat{f} \colon \mathcal{X} \to \mathcal{Y}$

- Negatively-oriented $= S((\mathbf{x}, y); \hat{f})$ high when the model is boldly wrong
- **Example (regression):**

$$S(\mathbf{x}, y) = |y - \hat{f}(\mathbf{x})|$$

  where $\hat{f}$ can be any regressor, e.g., random forest regression, fit on(augmented) data $\mathcal{D}^y$.

- ***Designing a good score is important***. Can have a big impact on performance!

Ful conformal prediction

For every candidate response $y \in \mathcal{Y}$:

1. **Fit a model** $\hat{f}^y : \mathcal{X} \to \mathcal{Y}$ to augmented data $\mathcal{D}^y$
2. **Compute nonconformity scores** for all $n+1$ points with the model:

$$S_i^y = S\big((\mathbf{x}_i, y_i); \hat{f}^y\big), i = 1, \ldots, n, \qquad S_{n+1}^y = S\big((\mathbf{x}_{\text{test}}, y); \hat{f}^y\big)$$

   and the $\frac{k}{n}$ sample quantile:

$$\hat{q}_n^y := \ k^{\text{th}} \text{ smallest of } S_1^y, \ldots, S_n^y, \qquad k := \lceil (1-\alpha)(n+1) \rceil$$

   with $\hat{q}_n^y := +\infty$ if $k > n$ (i.e. if $\alpha < 1/(n+1)$).
3. **Compare** S_{n+1}^y to the quantile $\hat{q}_n^y$,

$$\hat{C}(\mathbf{x}_{\text{test}}) = \{y \in \mathcal{Y} \ : S_{n+1}^y \leq \hat{q}_n^y\}$$

Ful conformal prediction

For every candidate response $y \in \mathcal{Y}$:

1. **Fit a model** $\hat{f}^y : \mathcal{X} \to \mathcal{Y}$ to augmented data $\mathcal{D}^y$
2. **Compute nonconformity scores** for all $n+1$ points with the model:

$$S_i^y = S\big((\mathbf{x}_i, y_i); \hat{f}^y\big), i = 1, \ldots, n, \qquad S_{n+1}^y = S\big((\mathbf{x}_{\text{test}}, y); \hat{f}^y\big)$$

   and the $\frac{k}{n}$ sample quantile:

$$\hat{q}_n^y := \ k^{\text{th}} \text{ smallest of } S_1^y, \ldots, S_n^y, \qquad k := \lceil (1-\alpha)(n+1) \rceil$$

   with $\hat{q}_n^y := +\infty$ if $k > n$ (i.e. if $\alpha < 1/(n+1)$).
3. **Compare** S_{n+1}^y to the quantile $\hat{q}_n^y$,

$$\hat{C}(\mathbf{x}_{\text{test}}) = \{y \in \mathcal{Y} \ : S_{n+1}^y \leq \hat{q}_n^y\}$$

Equivalently: include y in $\hat{\mathcal{C}}(\mathbf{x}_{\text{test}})$ iff at most $k-1$ of the calibration scores are smaller than $S_{n+1}^y \Leftrightarrow S_{n+1}^y$ ranks among the k smallest of *all* $n+1$ scores $S_1^y, \ldots, S_{n+1}^y$.

Comments

Interpretation via p-values:

- Compute p-**value** that represents the fraction of points in $\mathcal{D}^y$ that are at least as nonconforming as the test point:

$$p(y) = \frac{|i \in \{1, \ldots, n+1\} : S_i^y \geq S_{n+1}^y|}{n+1}$$

- Then $\hat{C}(\mathbf{x}_{\text{test}}) = \{y \in \mathcal{Y} \ : S_{n+1}^y \leq \hat{q}_n^y\} = \{y \in \mathcal{Y} \ : \ p(y) > \alpha\}$

Comments

Interpretation via p-values:

- Compute p-**value** that represents the fraction of points in $\mathcal{D}^y$ that are at least as nonconforming as the test point:

$$p(y) = \frac{|i \in \{1, \ldots, n+1\} : S_i^y \geq S_{n+1}^y|}{n+1}$$

- Then $\hat{C}(\mathbf{x}_{\text{test}}) = \{y \in \mathcal{Y} \ : S_{n+1}^y \leq \hat{q}_n^y\} = \{y \in \mathcal{Y} \ : \ p(y) > \alpha\}$

Remarks

- The refitted model $\hat{f}^y$, all $n+1$ scores, and hence the threshold $\hat{q}_n^y$ change with y.
- Provides exact coverage guarantee: At the true value y, the scores $S_1^y, \ldots, S_{n+1}^y$ are exchangeable and the rank of S_{n+1}^y is uniform on $\{1, \ldots, n+1\}$.

Comments

Interpretation via p-values:

- Compute p-**value** that represents the fraction of points in $\mathcal{D}^y$ that are at least as nonconforming as the test point:

$$p(y) = \frac{|i \in \{1, \ldots, n+1\} : S_i^y \geq S_{n+1}^y|}{n+1}$$

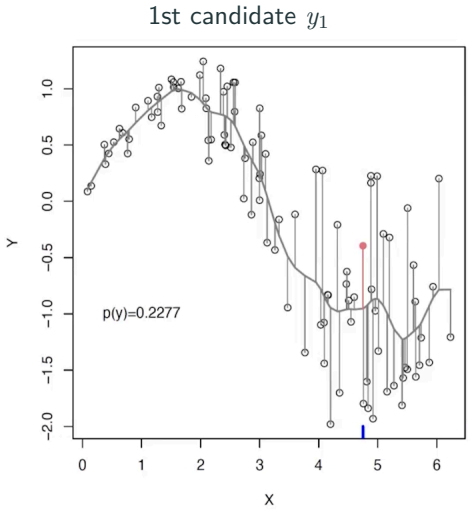
- Then $\hat{C}(\mathbf{x}_{\text{test}}) = \{y \in \mathcal{Y} \ : S_{n+1}^y \leq \hat{q}_n^y\} = \{y \in \mathcal{Y} \ : \ p(y) > \alpha\}$

Remarks

- The refitted model $\hat{f}^y$, all $n+1$ scores, and hence the threshold $\hat{q}_n^y$ change with y.
- Provides exact coverage guarantee: At the true value y, the scores $S_1^y, \ldots, S_{n+1}^y$ are exchangeable and the rank of S_{n+1}^y is uniform on $\{1, \ldots, n+1\}$.

Price: one model fit *per candidate* y

An example



For given level α, and each candidate y:

- Fit a model $\hat{f}^y$ on augmented data $\mathcal{D}^y$
- Calculate scores:

$$S_i^y = |y_i - \hat{f}^y(\mathbf{x}_i)|,\ 1 \le i \le n$$
$$S_{n+1}^y = |y - \hat{f}^y(\mathbf{x}_{\text{test}})|$$

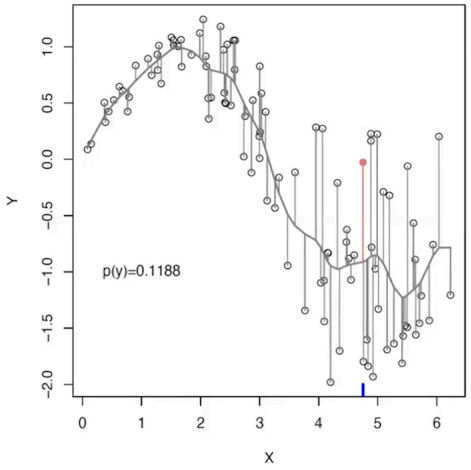
- $p(y) :=$ fraction of residuals with larger magnitudes than $|y - \hat{f}^y(\mathbf{x}_{\text{test}})|$:

$$p(y) = \frac{\#\{S_i^y \ge S_{n+1}^y\}}{n+1}$$

- y in $\hat{C}(\mathbf{x}_{\text{test}})$ if $p(y) \ge \alpha$

An example

2nd candidate y_2



For given level α, and each candidate y:

- Fit a model $\hat{f}^y$ on augmented data $\mathcal{D}^y$
- Calculate scores:

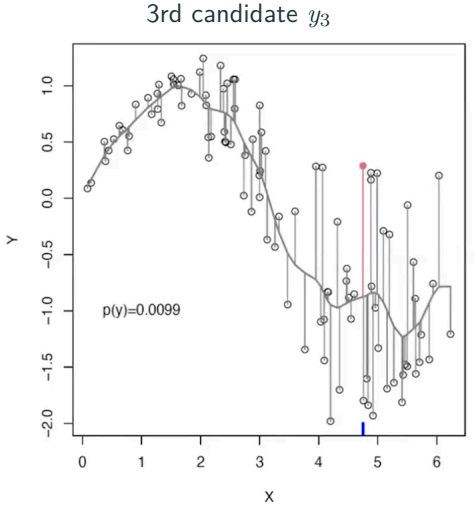
$$S_i^y = |y_i - \hat{f}^y(\mathbf{x}_i)|,\ 1 \le i \le n$$
$$S_{n+1}^y = |y - \hat{f}^y(\mathbf{x}_{\text{test}})|$$

- $p(y) :=$ fraction of residuals with larger magnitudes than $|y - \hat{f}^y(\mathbf{x}_{\text{test}})|$:

$$p(y) = \frac{\#\{S_i^y \ge S_{n+1}^y\}}{n+1}$$

- y in $\hat{C}(\mathbf{x}_{\text{test}})$ if $p(y) \ge \alpha$

An example



For given level α, and each candidate y:

- Fit a model $\hat{f}^y$ on augmented data $\mathcal{D}^y$
- Calculate scores:

$$S_i^y = |y_i - \hat{f}^y(\mathbf{x}_i)|,\ 1 \le i \le n$$
$$S_{n+1}^y = |y - \hat{f}^y(\mathbf{x}_{\text{test}})|$$

- $p(y) :=$ fraction of residuals with larger magnitudes than $|y - \hat{f}^y(\mathbf{x}_{\text{test}})|$:

$$p(y) = \frac{\#\{S_i^y \ge S_{n+1}^y\}}{n+1}$$

- y in $\hat{C}(\mathbf{x}_{\text{test}})$ if $p(y) \ge \alpha$

Remarks on full conformal

- Requires fitting a new model on augmented data for each candidate y
- Complexity scales with the cardinality of the response space $|\mathcal{Y}|$:
  - Classification: $\mathcal{Y}$ is discrete $\Rightarrow$ iterate through each value of y to construct the set $\hat{C}(\mathbf{x}_{\text{test}})$
  - Regression: $\mathcal{Y} = \mathbb{R} \Rightarrow$ discretize the response space
- Often practical only for **specialized algorithms** which do not require refitting the model

Split conformal prediction (SCP)

- Popular due to its simplicity, ease of implementation and computational efficiency.
- Starts by splitting the training set into two distinctive sets:
  - $D_1 :=$ **proper training set** (to fit a model)
  - $D_2 :=$ **calibration set** (to compute the scores and quantile)

$D_1 \cup D_2 = \{1, \ldots, n\}$, $|D_1| = n_1$, $|D_2| = n_2$.

Split conformal prediction (SCP)

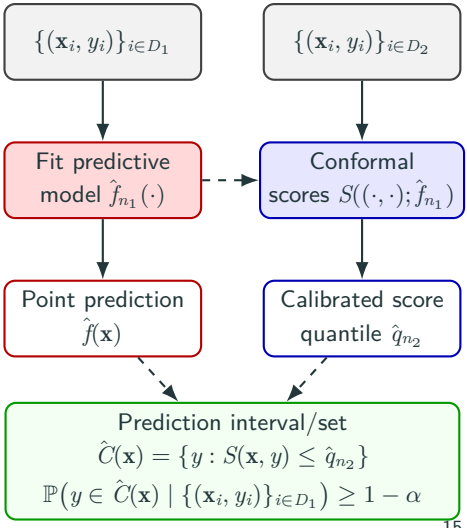
- Popular due to its simplicity, ease of implementation and computational efficiency.
- Starts by splitting the training set into two distinctive sets:
  - $D_1 :=$ **proper training set** (to fit a model)
  - $D_2 :=$ **calibration set** (to compute the scores and quantile)

  $D_1 \cup D_2 = \{1, \ldots, n\}$, $|D_1| = n_1$, $|D_2| = n_2$.
- **Choosing sample sizes n_1 and n_2**:
  - too small n_2 leads to a high variance of confidence (calibrating nonconformity scores not reliable)
  - too small n_1 leads to a downward bias in confidence ($\hat{f}$ does not produce confident predictions).

Split conformal prediction (SCP)

- Popular due to its simplicity, ease of implementation and computational efficiency.
- Starts by splitting the training set into two distinctive sets:
  - $D_1 :=$ **proper training set** (to fit a model)
  - $D_2 :=$ **calibration set** (to compute the scores and quantile)

  $D_1 \cup D_2 = \{1, \ldots, n\}$, $|D_1| = n_1$, $|D_2| = n_2$.
- **Choosing sample sizes n_1 and n_2**:
  - too small n_2 leads to a high variance of confidence (calibrating nonconformity scores not reliable)
  - too small n_1 leads to a downward bias in confidence ($\hat{f}$ does not produce confident predictions).



Algorithm 1: Split conformal prediction (SCP)

input : Data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, miscoverage level $\alpha \in (0,1)$, sets D_1 and D_2 s.t. $D_1 \cup D_2 = \{1, \ldots, n\}$, $|D_1| = n_1$, $|D_2| = n_2$, predictor $\hat{f}$.

1 Fit a predictor function $\hat{f}_{n_1}$ on the proper training set $\{(\mathbf{x}_i, y_i)\}_{i \in D_1}$.

2 Calculate nonconformity scores $S_i = S((\mathbf{x}_i, y_i); \hat{f}_{n_1})$ from the calibration set $(i \in D_2)$ and respective ordered scores

$$S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(n_2)}$$

3 Compute threshold $\hat{q}_{n_2} = S_{(k)}$, where $k = \lceil (1-\alpha)(n_2+1) \rceil$.

output : Conformal set $\hat{C}(\mathbf{x}) = \{y : S(\mathbf{x}, y) \leq \hat{q}_{n_2}\}$ that verifies

$$\Pr\Big(y \in \hat{C}(\mathbf{x}) \,\Big|\, (\mathbf{x}_i, y_i),\ i \in D_1\Big) \in \Big[1-\alpha,\ 1-\alpha+\frac{1}{n_2+1}\Big),$$

whereupper bound holds if scores are a.s. distinct.

SCP for regression

- Using abs. residual $S(\mathbf{x}, y) = |y - \hat{f}(\mathbf{x})|$ as the nonconformity score yields

$$\begin{aligned}\hat{C}(\mathbf{x}) &= \{y : S(\mathbf{x}, y) \leq \hat{q}_{n_2}\} \\ &= \{y : |y - \hat{f}_{n_1}(\mathbf{x})| \leq \hat{q}_{n_2}\} \\ &= \left[\hat{f}_{n_1}(\mathbf{x}) - \hat{q}_{n_2},\, \hat{f}_{n_1}(\mathbf{x}) + \hat{q}_{n_2}\right]\end{aligned}$$

  as the respective **conformal band**.

SCP for regression

- Using abs. residual $S(\mathbf{x}, y) = |y - \hat{f}(\mathbf{x})|$ as the nonconformity score yields

$$\begin{aligned}\hat{C}(\mathbf{x}) &= \{y : S(\mathbf{x}, y) \leq \hat{q}_{n_2}\} \\ &= \{y : |y - \hat{f}_{n_1}(\mathbf{x})| \leq \hat{q}_{n_2}\} \\ &= \left[\hat{f}_{n_1}(\mathbf{x}) - \hat{q}_{n_2},\, \hat{f}_{n_1}(\mathbf{x}) + \hat{q}_{n_2}\right]\end{aligned}$$

  as the respective **conformal band**.

- The width of the band is *exactly constant in* $\mathbf{x}$ regardless of how hard it is to predict at $\mathbf{x}$.

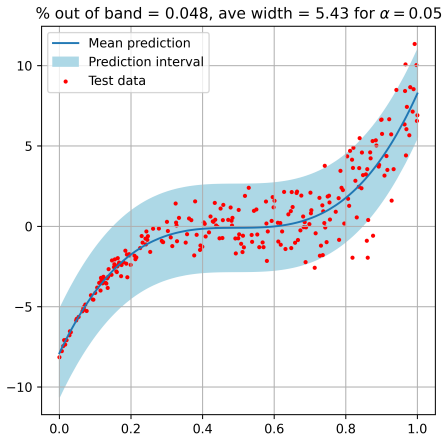
SCP for regression

- Using abs. residual $S(\mathbf{x}, y) = |y - \hat{f}(\mathbf{x})|$ as the nonconformity score yields

$$\begin{aligned}\hat{C}(\mathbf{x}) &= \{y : S(\mathbf{x}, y) \leq \hat{q}_{n_2}\} \\ &= \{y : |y - \hat{f}_{n_1}(\mathbf{x})| \leq \hat{q}_{n_2}\} \\ &= \left[\hat{f}_{n_1}(\mathbf{x}) - \hat{q}_{n_2},\, \hat{f}_{n_1}(\mathbf{x}) + \hat{q}_{n_2}\right]\end{aligned}$$

  as the respective **conformal band**.

- The width of the band is *exactly constant in* $\mathbf{x}$ regardless of how hard it is to predict at $\mathbf{x}$.



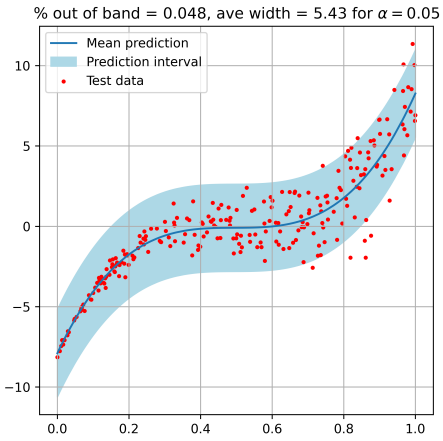
SCP for regression

- Using abs. residual $S(\mathbf{x}, y) = |y - \hat{f}(\mathbf{x})|$ as the nonconformity score yields

$$\begin{aligned}\hat{C}(\mathbf{x}) &= \{y : S(\mathbf{x}, y) \leq \hat{q}_{n_2}\} \\ &= \{y : |y - \hat{f}_{n_1}(\mathbf{x})| \leq \hat{q}_{n_2}\} \\ &= \left[\hat{f}_{n_1}(\mathbf{x}) - \hat{q}_{n_2},\, \hat{f}_{n_1}(\mathbf{x}) + \hat{q}_{n_2}\right]\end{aligned}$$

  as the respective **conformal band**.

- The width of the band is *exactly constant in* $\mathbf{x}$ regardless of how hard it is to predict at $\mathbf{x}$.
- Better to adapt to hardness of the problem:
  - $\rightarrow$ studentized residuals scores
  - $\rightarrow$ conformal quantile regression



Studentized residuals

- On D_1 we fit both a predictor function $\hat{f}_{n_1}$ that predicts y given $\mathbf{x}$ as well as predictor for spread of responses, $\hat{\sigma}_{n_1}$, that predict (say) the standard deviation of $y - \hat{f}_{n_1}(\mathbf{x})$ given $\mathbf{x}$.

Studentized residuals

- On D_1 we fit both a predictor function $\hat{f}_{n_1}$ that predicts y given $\mathbf{x}$ as well as predictor for spread of responses, $\hat{\sigma}_{n_1}$, that predict (say) the standard deviation of $y - \hat{f}_{n_1}(\mathbf{x})$ given $\mathbf{x}$.
- Then on D_2, we compute studentized (or normalized) residuals [LGR+18]:

$$S_i = S((\mathbf{x}_i, y_i); \hat{f}_{n_1}) = \frac{|y_i - \hat{f}_{n_1}(\mathbf{x}_i)|}{\hat{\sigma}_{n_1}(\mathbf{x}_i)}, \quad i \in D_2,$$

and as before $\hat{q}_{n_2} = S_{(k)}$ where $k = \lceil (1-\alpha)(n_2+1) \rceil$.

Studentized residuals

- On D_1 we fit both a predictor function $\hat{f}_{n_1}$ that predicts y given $\mathbf{x}$ as well as predictor for spread of responses, $\hat{\sigma}_{n_1}$, that predict (say) the standard deviation of $y - \hat{f}_{n_1}(\mathbf{x})$ given $\mathbf{x}$.
- Then on D_2, we compute studentized (or normalized) residuals [LGR+18]:

$$S_i = S((\mathbf{x}_i, y_i); \hat{f}_{n_1}) = \frac{|y_i - \hat{f}_{n_1}(\mathbf{x}_i)|}{\hat{\sigma}_{n_1}(\mathbf{x}_i)}, \quad i \in D_2,$$

  and as before $\hat{q}_{n_2} = S_{(k)}$ where $k = \lceil (1-\alpha)(n_2+1) \rceil$.
- The conformal band is now

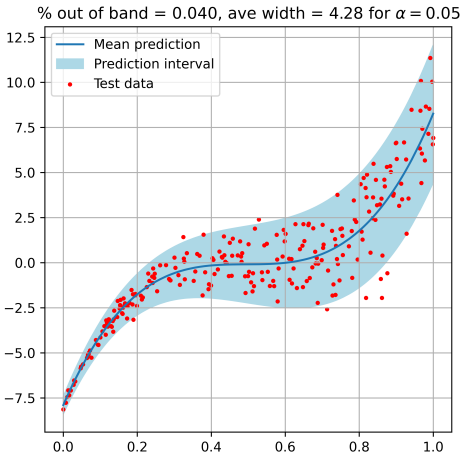
$$\hat{C}(\mathbf{x}) = \left[\hat{f}_{n_1}(\mathbf{x}) - \hat{\sigma}_{n_1}(\mathbf{x}) \hat{q}_{n_2}, \hat{f}_{n_1}(\mathbf{x}) + \hat{\sigma}_{n_1}(\mathbf{x}) \hat{q}_{n_2} \right].$$

  and its width adapts to $\mathbf{x}$ according to $\hat{\sigma}_{n_1}$.

Studentized conformal regression: an example

How to construct the spread predictor $\hat{\sigma}_{n_1}$:

- Let $r_i = y_i - \hat{f}_{n_1}(\mathbf{x}_i)$ denote residuals, $i \in D_1$.
- For 1D features, one can use $\hat{\sigma}_{\text{sd}}(x) = \text{spline}(x)$ where $\text{spline}(\cdot)$ is a spline of some proper degree k fitted to data x_i and $|r_i|$, $i \in D_1$.
- For higher dim. features, one can use Random Forest Regressor, or, some local estimate of spread of residuals.



Conformalized quantile regression (CQP)

Problems of studentized residuals:

1. **overfitting**: If $\hat{f}_{n_1}$ is complex (e.g., NN), then little information is left on proper training set to fit $\hat{\sigma}_{n_1}$ because the proper training residuals will be close to zero.
2. **symmetric confidence band**: $\hat{C}(\mathbf{x})$ is symmetric around $\hat{f}_{n_1}(\mathbf{x})$.

Conformalized quantile regression (CQP)

Problems of studentized residuals:

1. **overfitting**: If $\hat{f}_{n_1}$ is complex (e.g., NN), then little information is left on proper training set to fit $\hat{\sigma}_{n_1}$ because the proper training residuals will be close to zero.
2. **symmetric confidence band**: $\hat{C}(\mathbf{x})$ is symmetric around $\hat{f}_{n_1}(\mathbf{x})$.

Solution: change the predictor function

- instead of using $\hat{f}(\mathbf{x})$ to estimate $\mathbb{E}[y \mid \mathbf{x}]$, estimate both the lower and upper $\frac{\alpha}{2}$-quantiles of the conditional distribution $y \mid \mathbf{x}$:

$$\hat{L}(\mathbf{x}) := \hat{f}_{n_1}^{\alpha/2}(\mathbf{x}) \quad \text{(lower)}, \qquad \hat{U}(\mathbf{x}) := \hat{f}_{n_1}^{1-\alpha/2}(\mathbf{x}) \quad \text{(upper)}.$$

  $\hat{f}_{n_1}^{\alpha/2}$:= quantile regressor fitted on the proper training set using $\alpha/2$ quantile
- If they were perfect, $\hat{C}(\mathbf{x}) = [\hat{L}(\mathbf{x}), \hat{U}(\mathbf{x})]$ would already cover y with prob. $1 - \alpha$ *given* $\mathbf{x}$

The nonconformity score: a *signed* distance to the band

$$S_i \;=\; S\big((\mathbf{x}_i, y_i); \hat{L}, \hat{U}\big) \;:=\; \max\Big\{ \underbrace{\hat{L}(\mathbf{x}_i) - y_i}_{\text{how far } \textit{below} \text{ the band}} \;,\; \underbrace{y_i - \hat{U}(\mathbf{x}_i)}_{\text{how far } \textit{above} \text{ the band}} \Big\}, \qquad i \in D_2.$$

Where is y_i?	S_i	**Meaning of $\lvert S_i \rvert$**
above the band, $y_i > \hat{U}(\mathbf{x}_i)$	> 0	how much the band missed it
below the band, $y_i < \hat{L}(\mathbf{x}_i)$	> 0	how much the band missed it
inside the band	< 0	margin: distance to the nearer edge
exactly on an edge	$= 0$	—

$|S_i|$:= distance from y_i to the **nearest edge**. $\operatorname{sign}(S_i)$ tells if we are outside $(+)$ or inside $(-)$.

$$S_i \le 0 \quad \Longleftrightarrow \quad y_i \in [\hat{L}(\mathbf{x}_i),\, \hat{U}(\mathbf{x}_i)].$$

The conformalized quantile regression (CQR) bands

- Set $\hat{q}_{n_2} = S_{(k)} := k^{\text{th}}$ smallest of the n_2 calibration scores, $k = \lceil (1-\alpha)(n_2+1) \rceil$.
- The SCP principle leads to **conformalized quantile regression (CQR)** bands [RPC19]

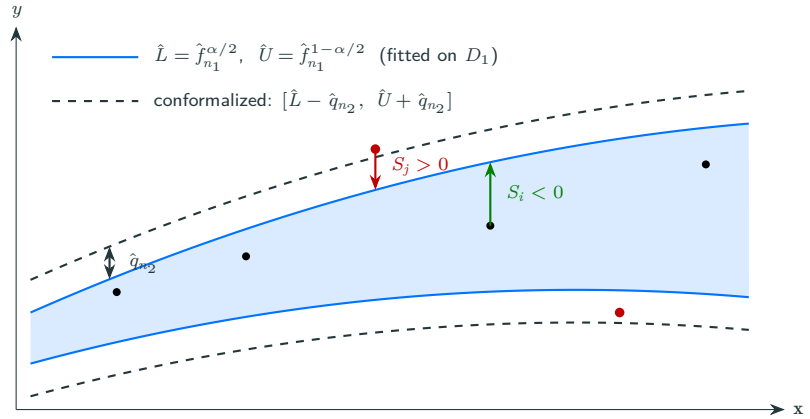
$$\begin{aligned}\hat{C}(\mathbf{x}) &= \{y : S((\mathbf{x}, y); \hat{L}, \hat{U}) \leq \hat{q}_{n_2}\} \\ &= \{y : \max\left\{\hat{L}(\mathbf{x}) - y,\ y - \hat{U}(\mathbf{x})\right\} \leq \hat{q}_{n_2}\} \\ &= [\hat{L}(\mathbf{x}) - \hat{q}_{n_2},\ \hat{U}(\mathbf{x}) + \hat{q}_{n_2}]\end{aligned}$$

  where we used the elementary fact: $\max\{a, b\} \leq q \iff a \leq q$ **and** $b \leq q$.
- Testing every candidate y is unnecessary: the set is available in closed form!

$\hat{q}_{n_2}$:= correction needed s.t. approx. $(1-\alpha)$ of the calibration points would have been covered

- $\hat{q}_{n_2} > 0$: too many calibration points fell outside the band $\Rightarrow$ **Inflate** the band.
- $\hat{q}_{n_2} < 0$: the band already over-covered: most scores were negative $\Rightarrow$ **Shrink** the band.

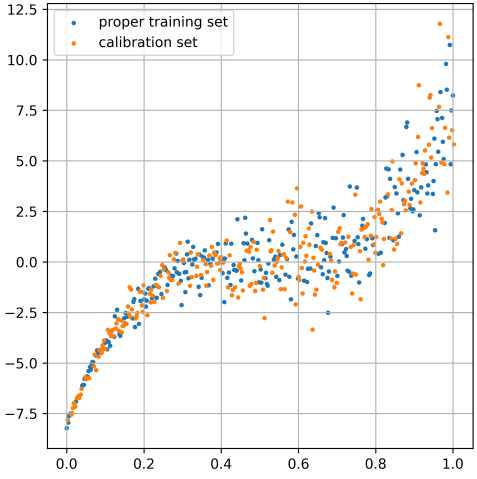
Same but explained in picture

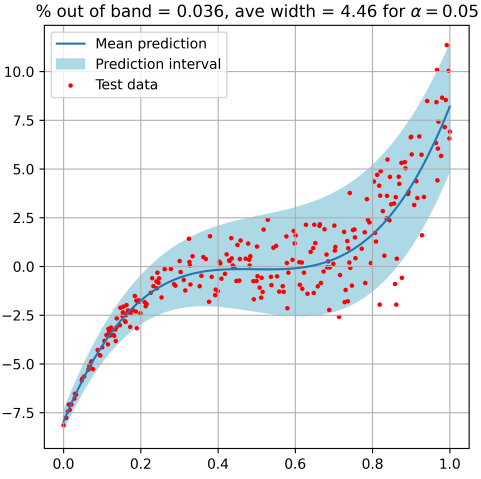


Inflate the band: both edges are pushed out by the *same* $\hat{q}_{n_2} > 0$.

The width varies with $\mathbf{x}$: the model (quantile regressors) supply the *shape*.

Conformalized quantile regression: an example



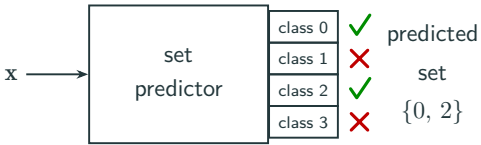


SCP for classification

$y \in \mathcal{Y} = \{1, \ldots, K\}$, $K := \#$ of classes

Classifier yields a prob. prediction

$\hat{f}_{n_1}(\mathbf{x}; y) \approx \Pr(Y = y \mid X = \mathbf{x}), \forall y \in \mathcal{Y}$



- Consider nonconformity score

$$S_i = 1 - \hat{f}_{n_1}(\mathbf{x}_i; y_i), \quad i \in D_2.$$

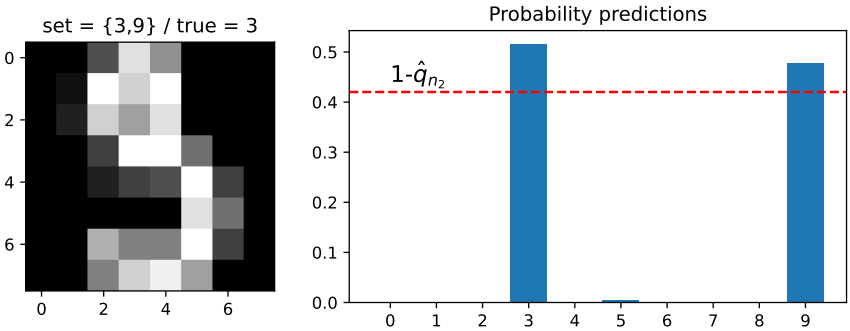
which equal one minus the **softmax output** of the true class of $(\mathbf{x}_i, y_i)$, $i \in D_2$.

- Respective conformal set is

$$\hat{C}(\mathbf{x}) = \left\{k : 1 - \hat{f}_{n_1}(\mathbf{x}; k) \leq \hat{q}_{n_2}\right\} = \left\{k : \hat{f}_{n_1}(\mathbf{x}; k) \geq 1 - \hat{q}_{n_2}\right\}$$

- Has the usual SCP guarantee $\Pr\Big(y \in \hat{C}(\mathbf{x}) \,\Big|\, (\mathbf{x}_i, y_i),\, i \in D_1\Big) \in \Big[1 - \alpha,\, 1 - \alpha + \frac{1}{n_2+1}\Big)$

Digits data set example



Left: an image (with true label "3") from the the test set which has conformal set {"1", "3"} with significance level $\alpha = 0.05$. Model used is LDA classifier.

Right: probability predictions for the given image as well as the threshold $1 - \hat{q}_{n_2}$.

Binary case: the conformal set is a threshold rule

- Two classes: $\mathcal{K} = \{0, 1\}$, score $S\big((\mathbf{x}, y)\big) = 1 - \hat{f}_{n_1}(\mathbf{x}; y) :=$ prob. of the *false* label
- Write $\hat{p}(\mathbf{x}) := \hat{f}_{n_1}(\mathbf{x}; 1)$, so that $\hat{f}_{n_1}(\mathbf{x}; 0) = 1 - \hat{p}(\mathbf{x})$.
- The split-CP rule "include y iff $S \leq \hat{q}_{n_2}$" reads

$$y \in \hat{C}(\mathbf{x}) \iff 1 - \hat{f}_{n_1}(\mathbf{x}; y) \leq \hat{q}_{n_2} \iff \boxed{\hat{f}_{n_1}(\mathbf{x}; y) \geq 1 - \hat{q}_{n_2}}$$

- **Each label is admitted on its own**

$$1 \in \hat{C}(\mathbf{x}) \iff \hat{p}(\mathbf{x}) \geq 1 - \hat{q}_{n_2}, \qquad 0 \in \hat{C}(\mathbf{x}) \iff \hat{p}(\mathbf{x}) \leq \hat{q}_{n_2}.$$

Note: the two conditions are *separate*, i.e., neither implies the other.

The conformal step:

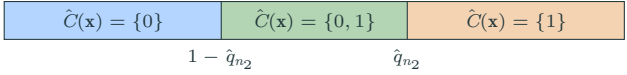
- replaces the arbitrary cut-off 0.5 by the data-driven threshold $\tau = 1 - \hat{q}_{n_2}$
- allows the answer to have 0, 1 or 2 labels

Two regimes, decided by $\hat{q}_{n_2}$ vs. $1/2$

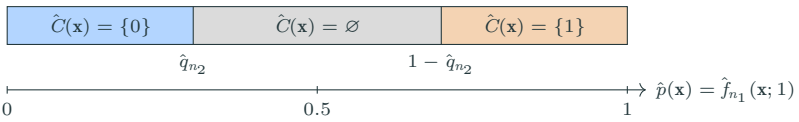
The two cut points are always $\hat{q}_{n_2}$ and $1 - \hat{q}_{n_2}$, i.e., mirror images about $1/2$.

Which one lies on the left decides what the middle region means.

A. $\hat{q}_{n_2} \geq \frac{1}{2}$ (threshold $\tau = 1 - \hat{q}_{n_2} \leq \frac{1}{2}$): the middle is **both labels** $\hat{C}(\mathbf{x}) = \{0, 1\}$ iff $\hat{p}(\mathbf{x}) \in [1 - \hat{q}_{n_2},\ \hat{q}_{n_2}]$
(the classifier is weak / the calibration scores are large)



B. $\hat{q}_{n_2} < \frac{1}{2}$ (threshold $\tau = 1 - \hat{q}_{n_2} > \frac{1}{2}$): the middle is the **empty set** $\hat{C}(\mathbf{x}) = \varnothing$ iff $\hat{p}(\mathbf{x}) \in (\hat{q}_{n_2},\ 1 - \hat{q}_{n_2})$



Note: $\{0, 1\}$ and $\varnothing$ can *never* both occur for the same $\hat{q}_{n_2}$,

How to read the output

- $|\hat{C}(\mathbf{x})| = 1$: a confident decision. One label clears the threshold $1 - \hat{q}_{n_2}$.
- $|\hat{C}(\mathbf{x})| = 2$: **ambiguity**. $\hat{p}(\mathbf{x})$ near $1/2$ and the calibration set says the classifier is not reliable enough to break the tie.
- $|\hat{C}(\mathbf{x})| = 0$: **both labels look abnormal**. $\hat{p}(\mathbf{x})$ is in a region where calibration points of *either* class were rare. A useful outlier / distribution-shift flag, not a bug.

Coverage $\Pr(y_{\text{test}} \in \hat{C}(\mathbf{x}_{\text{test}})) \geq 1 - \alpha$ holds regardless of empty set.

If a decision is required, output $\arg\max_y \hat{f}_{n_1}(\mathbf{x}; y)$ when $\hat{C}(\mathbf{x}) = \varnothing$: enlarging a set can only *increase* coverage, so validity survives.

Online conformal prediction

- We observe covariate-response pairs $\{(\mathbf{x}_t, y_t)\}_{t \in \mathbb{N}} \subseteq \mathcal{X} \times \mathcal{Y}$ in a sequential fashion.
- Split and full conformal require exchangeability. What if this does not hold?
- Ex: distribution shifts, temporal dependence and non-stationarity
- **Online CP (OCP)** [GC21] overcomes this limitation by adapting the miscoverage level α
- Provides a deterministic long-run coverage guarantee without exchangeability assumption

Online conformal prediction

- We observe covariate-response pairs $\{(\mathbf{x}_t, y_t)\}_{t\in\mathbb{N}} \subseteq \mathcal{X} \times \mathcal{Y}$ in a sequential fashion.
- Split and full conformal require exchangeability. What if this does not hold?
- Ex: distribution shifts, temporal dependence and non-stationarity
- **Online CP (OCP)** [GC21] overcomes this limitation by adapting the miscoverage level α
- Provides a deterministic long-run coverage guarantee without exchangeability assumption

Key Idea:

- Under iid assumption observed and true conformal score distribution will approx. align.
- Distribution shift can cause the true distribution of conformal scores to move to the right or to the left.
- $1 - \alpha_t^\star$ quantile of obs. dist. $\triangleq$ $1 - \alpha$ quantile of true dist. $\rightarrow$ learn $\alpha_t^\star$ (track it online)

Adapting conformal to distribution shifts

1. At time t, use historical data $\{(\mathbf{x}_r, y_r)\}_{t-W \le r < t}$ where W is some time window to form the fitted score function $S_t((\mathbf{x}, y))$ and the corresponding quantile $q_t(1-\alpha_t)$.
2. The prediction set $\hat{C}_t$ at time t is $\hat{C}_t = \{y \in \mathcal{Y} : S_t(\mathbf{x}_t, y) \le q_t(1-\alpha_t)\}$
3. Update α via simple rule

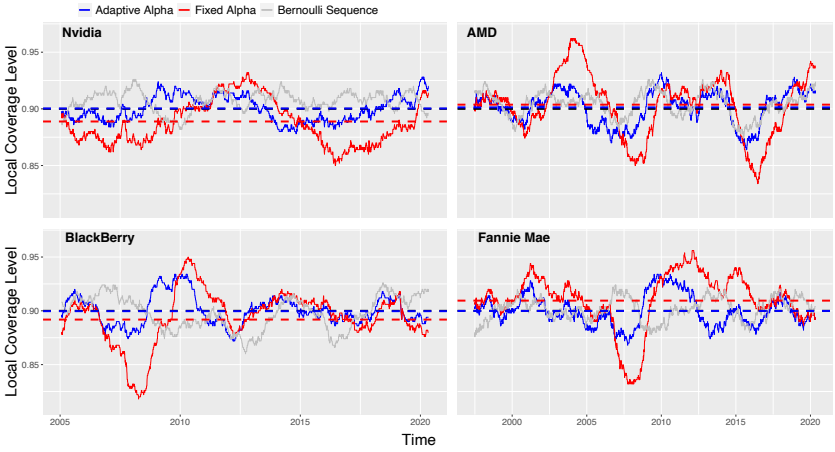
$$\alpha_{t+1} := \alpha_t + \gamma(\alpha - \mathsf{err}_t), \quad \mathsf{err}_t := \begin{cases} 1, \text{ if } y_t \notin \hat{C}_t, \\ 0, \text{ otherwise,} \end{cases}$$

where $\{\alpha_t\}$ is initialized with $\alpha_W = \alpha$ and $\gamma > 0$ a fixed step size parameter.

Not covered: $1-\alpha_t \uparrow$ (interval gets larger), covered: $1-\alpha_t \downarrow$ (interval gets shorter)

Under no assumptions whatsoever , $\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathsf{err}_t \overset{a.s.}{=} \alpha.$

An example from [GC21]



- Local coverage frequencies for the prediction of stock market volatility via Garch$(1,1)$ model
- Col. dotted lines: aver. coverage across all time points, black line: target level of $1-\alpha = 0.9$.

Conformal prediction in my research group

Recent titles:

Vision-Aided LLM-Empowered Online Conformal Beam Prediction, submitted to Wireless Communications Letters. Aug. 14, 2026

Uncertainty Quantification for LLM-Empowered Channel Prediction via Online Conformal Prediction, submitted to ICASSP-2027

Credits to Lei Wang 

Beam prediction

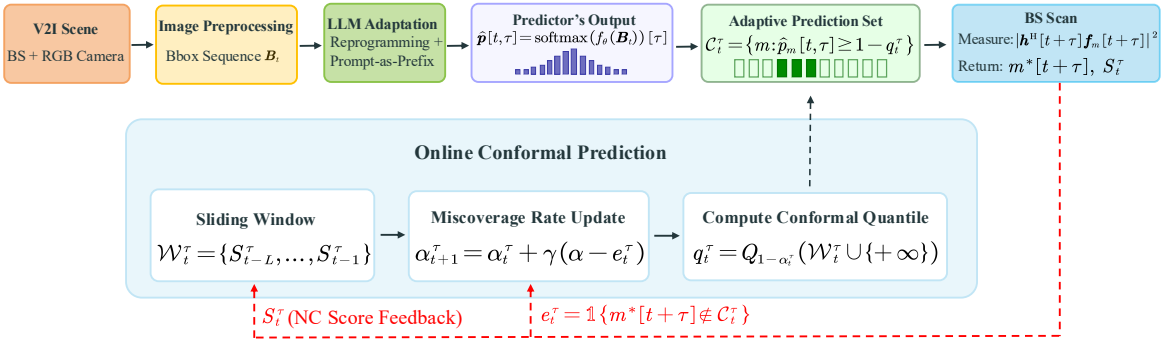
- a single-antenna UE communicating with a BS with N antennas and an RGB camera.
- BS uses predefined beamforming codebook $\mathcal{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_M\}$, where M is the codebook size, $\mathbf{f}_m \in \mathbb{C}^{N\times 1}$ is the m^{th} beamforming vector ($\|\mathbf{f}_m\|^2 = 1$).
- The instantaneous optimal beam index is

$$m^\star[t] = \underset{m\in\{1,\ldots,M\}}{\arg\max}\ |\mathbf{h}^{\mathsf{H}}[t]\mathbf{f}_m[t]|^2, \qquad \mathbf{h}[t] := \text{channel at time } t$$

- BS captures an RGB image at time t, from which a YOLO-based detector extracts the bounding-box feature $\mathbf{b}[t]$ of the target UE
- T_{hist} camera observations are then collected as $\mathbf{B}_t = [\mathbf{b}[t - T_{\text{hist}}], \ldots, \mathbf{b}[t-1]]$ and fed into the pre-trained LLM-empowered beam predictor $f_\theta(\cdot)$.
- For each prediction horizon τ the predictor outputs a probability vector for beams:

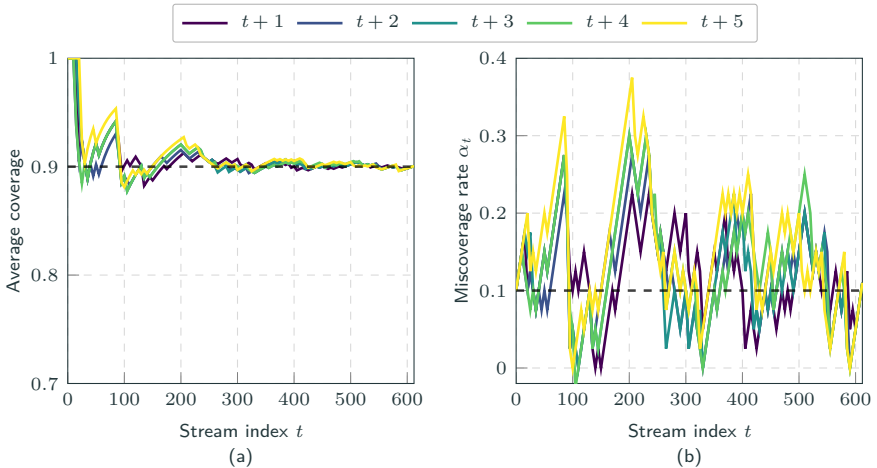
$$\hat{\mathbf{p}}[t,\tau] = \text{softmax}\big(f_\theta(\mathbf{B}_t)\big)[\tau] \in [0,1]^{M\times 1}, \quad \sum_{m=1}^{M} \hat{p}_m[t,\tau] = 1.$$

LLM-empowered beam prediction using CP

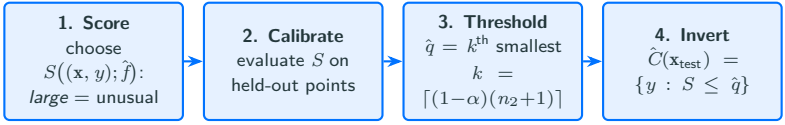


Performance

DeepSense 6G Scenario 8 V2I dataset ($\alpha = 0.1$, $\gamma = 0.05$, $L = 100$ recent scores for q_t calibration)



Conclusions



- Coverage is **finite-sample**, **distribution-free** and holds for *any* underlying predictor. A useless $\hat{f}$ still yields valid sets, they will just be very wide.
- **Efficiency**: short intervals / small sets is the job of the model and the score.
- **Exchangeability**: Split and Full CP fail under distribution shift and serial correlation. Then coverage fails too.
- **Randomness of the split.** set depends on points in D_2; re-splitting gives a different band.
- A wide interval or a two-label set is the method **telling you the model is unsure**.
- Online conformal prediction does not require exchangeability, and adaptively calibrate the quantile (via α_t)

References i

- Anastasios N Angelopoulos, Stephen Bates, et al., *Conformal prediction: A gentle introduction*, Foundations and Trends® in Machine Learning **16** (2023), no. 4, 494–591.
- Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik, *Uncertainty sets for image classifiers using conformal prediction*, Proc. ICLR-2021 (Vienna, Austria (virtual)), 2021.
- Isaac Gibbs and Emmanuel Candes, *Adaptive conformal inference under distribution shift*, Advances in Neural Information Processing Systems **34** (2021), 1660–1672.
- Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman, *Distribution-free predictive inference for regression*, J. Amer. Stat. Assoc. **113** (2018), no. 523, 1094–1111.

References ii

- Jing Lei and Larry Wasserman, *Distribution-free prediction bands for non-parametric regression*, J. Royal Stat. Soc., Ser. B **76** (2014), no. 1, 71–96.
- Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman, *Inductive confidence machines for regression*, Proc. ECML-2022 (Helsinki, Finland), 2002, pp. 345–356.
- Yaniv Romano, Evan Patterson, and Emmanuel Candes, *Conformalized quantile regression*, Adv. Neural Inf. Process. Syst. **32** (2019).
- Yaniv Romano, Matteo Sesia, and Emmanuel Candes, *Classification with valid and adaptive coverage*, Advances in Neural Information Processing Systems **33** (2020), 3581–3591.

References iii


Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, *Algorithmic learning in a random world*, vol. 29, Springer, 2005.