A tutorial on Conformal Prediction

Pierre Humbert

January 24, 2025

Motivation

Overall goal behind CP: Quantifying the uncertainty of an algorithm in its predictions

Why ?

- Take a decision regarding these predictions
  $\longrightarrow$ Need to trust/have confidence in the algorithm

- Accuracy is not enough
  Global measure $\neq$ Instance-wise uncertainty quantification

  ex: Medical diagnostic $\longrightarrow$ Decision for THIS particular patient

Motivation

Overall goal behind CP: Quantifying the uncertainty of an algorithm in its predictions

Why ?

- Take a decision regarding these predictions
  $\longrightarrow$ Need to trust/have confidence in the algorithm

- Accuracy is not enough
  Global measure $\neq$ Instance-wise uncertainty quantification

  ex: Medical diagnostic $\longrightarrow$ Decision for THIS particular patient

Motivation

Overall goal behind CP: Quantifying the uncertainty of an algorithm in its predictions

Why ?

- Take a decision regarding these predictions
  $\longrightarrow$ Need to trust/have confidence in the algorithm

- Accuracy is not enough
  Global measure $\neq$ Instance-wise uncertainty quantification

  ex: Medical diagnostic $\longrightarrow$ Decision for THIS particular patient

Motivation

Overall goal behind CP: Quantifying the uncertainty of an algorithm in its predictions

Why ?

- Take a decision regarding these predictions
  $\longrightarrow$ Need to trust/have confidence in the algorithm

- Accuracy is not enough
  Global measure $\neq$ Instance-wise uncertainty quantification

  ex: Medical diagnostic $\longrightarrow$ Decision for THIS particular patient

Motivation

Overall goal behind CP: Quantifying the uncertainty of an algorithm in its predictions

Why ?

- Take a decision regarding these predictions
  $\longrightarrow$ Need to trust/have confidence in the algorithm

- Accuracy is not enough
  Global measure $\neq$ Instance-wise uncertainty quantification

  ex: Medical diagnostic $\longrightarrow$ Decision for THIS particular patient

Conformal Prediction (CP)
(Vovk et al., 2005)

CP is one way to provide uncertainty quantification

In a supervised problem

- Given a new observation
  $\longrightarrow$ Predict its associated response (point prediction)

In conformal prediction

- Given a new observation
  $\longrightarrow$ Construct a set containing the true response with high probability

Conformal Prediction (CP) (Vovk et al., 2005)

CP is one way to provide uncertainty quantification

In a supervised problem

- Given a new observation
  $\longrightarrow$ Predict its associated response (point prediction)

In conformal prediction

- Given a new observation
  $\longrightarrow$ Construct a set containing the true response with high probability

Conformal Prediction (CP) (Vovk et al., 2005)

CP is one way to provide uncertainty quantification

In a supervised problem

- Given a new observation
  $\longrightarrow$ Predict its associated response (point prediction)

In conformal prediction

- Given a new observation
  $\longrightarrow$ Construct a set containing the true response with high probability

Conformal Prediction (CP) (Vovk et al., 2005)

CP is one way to provide uncertainty quantification

In a supervised problem

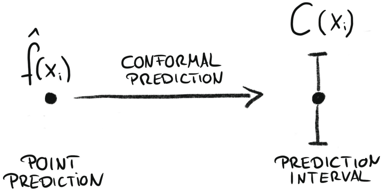
- Given a new observation
  $\longrightarrow$ Predict its associated response (point prediction)

In conformal prediction

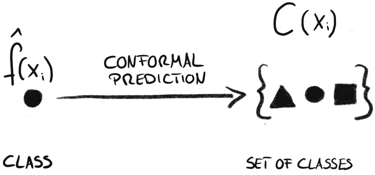
- Given a new observation
  $\longrightarrow$ Construct a set containing the true response with high probability

Conformal Prediction (CP)

Regression



Classification



Main objective

Setup: n i.i.d. (or exchangeable) random variables
$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \sim P$

Marginal guarantee:

For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha \quad (1)$$

for **any distribution** P and **any sample size** n.

How to do that? $\longrightarrow$ The split CP method (Papadopoulos et al., 2002)

Main objective

Setup: n i.i.d. (or exchangeable) random variables
$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \sim P$

Marginal guarantee:

For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha \tag{1}$$

for **any distribution** P and **any sample size** n.

How to do that? $\longrightarrow$ The split CP method (Papadopoulos et al., 2002)

Main objective

Setup: n i.i.d. (or exchangeable) random variables
$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \sim P$

Marginal guarantee:

For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha \tag{1}$$

for **any distribution** P and **any sample size** n.

How to do that? $\longrightarrow$ The split CP method (Papadopoulos et al., 2002)

Main objective

Setup: n i.i.d. (or exchangeable) random variables
$Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \sim P$

Marginal guarantee:

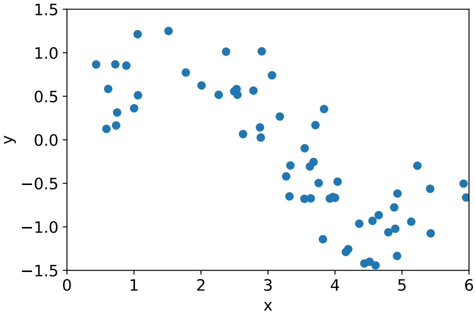
For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

$$\mathbb{P}(Y \in C(X)) \geq 1 - \alpha \tag{1}$$

for **any distribution** P and **any sample size** n.

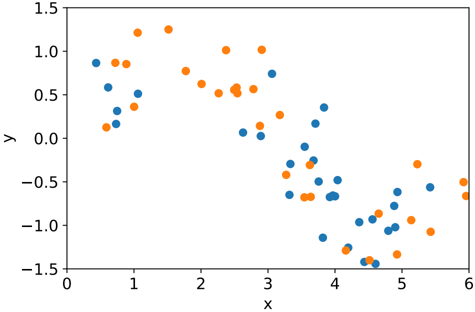
How to do that? $\longrightarrow$ The split CP method (Papadopoulos et al., 2002)

Split Conformal Prediction



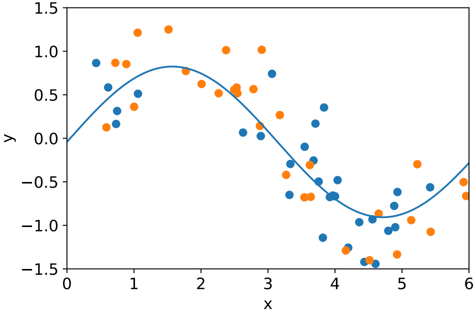
Input: $Z_1, \ldots, Z_n$ and $\alpha \in (0, 1)$

Split Conformal Prediction



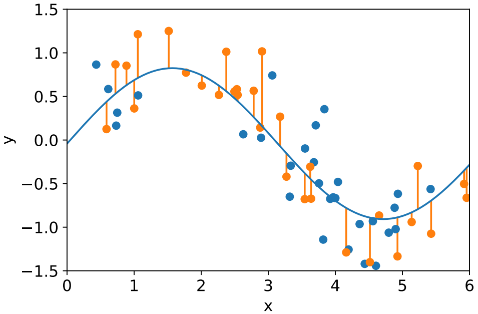
Randomly split $\{1, \ldots, n\}$ into two subsets $\mathcal{I}_1$ and $\mathcal{I}_2$

Split Conformal Prediction



Learn a predictor $\widehat{f}$ on $\{Z_i, i \in \mathcal{I}_1\}$ (blue)

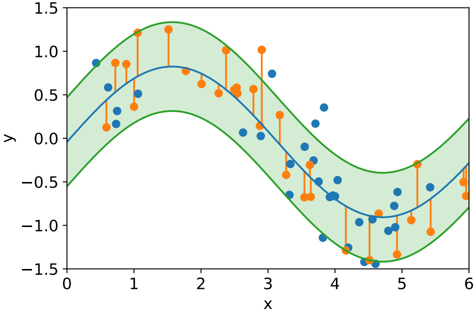
Split Conformal Prediction



Choose a **score function** $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ ex: $s(x, y) = |\widehat{f}(x) - y|$

Compute scores $S_i = s(X, Y)$ on $\{Z_i, i \in \mathcal{I}_2\}$ (orange)

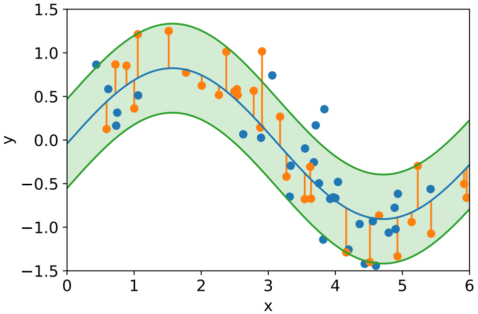
Split Conformal Prediction



Compute $S_{(r)} =$ the r-th smallest values in $\{S_i\}_{i \in \mathcal{I}_2}$ (quantile computation)

Return
$\widehat{C}_r(x) = \{y : s(x, y) \leq S_{(r)}\} \stackrel{\text{(ex)}}{=} [\widehat{f}(x) - S_{(r)}, \widehat{f}(x) + S_{(r)}]$

Split Conformal Prediction



$$\widehat{C}_r(x) = \{y : s(x,y) \le S_{(r)}\} \quad \overset{\text{(ex)}}{=} [\widehat{f}(x) - S_{(r)}, \widehat{f}(x) + S_{(r)}]$$

$\longrightarrow$ Which value for r to have: $\quad \mathbb{P}(Y \in \widehat{C}_r(X)) \ge 1 - \alpha \quad ?$

Main theorem in CP

Theorem

(Vovk et al., 2005; Lei et al., 2018)

- *If* $r^* = \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil$*, the set returns by the Split Conformal Prediction method satisfies*

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \geq 1 - \alpha \ , \tag{2}$$

for any distribution P*, any score function* $s(\cdot,\cdot)$*, and any sample size* n *(distribution-free).*

- *If we assume that the scores* $\{S_i\}_{i \in \mathcal{I}_2}$ *and* $S_{n+1} = s(X, Y)$ *are continuous, then*

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2|+1} \ , \tag{3}$$

with $|\mathcal{I}_2|$ *the size of the second subset.*

Main theorem in CP

Theorem

(Vovk et al., 2005; Lei et al., 2018)

- *If* $r^* = \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil$*, the set returns by the Split Conformal Prediction method satisfies*

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \geq 1-\alpha \ , \tag{2}$$

for any distribution P*, any score function* $s(\cdot,\cdot)$*, and any sample size* n *(distribution-free).*

- *If we assume that the scores* $\{S_i\}_{i\in\mathcal{I}_2}$ *and* $S_{n+1} = s(X,Y)$ *are continuous, then*

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \leq 1-\alpha + \frac{1}{|\mathcal{I}_2|+1} \ , \tag{3}$$

with $|\mathcal{I}_2|$ *the size of the second subset.*

Main theorem in CP

Theorem

(Vovk et al., 2005; Lei et al., 2018)

- *If* $r^* = \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil$, *the set returns by the Split Conformal Prediction method satisfies*

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \geq 1 - \alpha \ , \tag{2}$$

for any distribution P, *any score function* $s(\cdot,\cdot)$, *and any sample size* n *(distribution-free).*

- *If we assume that the scores* $\{S_i\}_{i \in \mathcal{I}_2}$ *and* $S_{n+1} = s(X, Y)$ *are continuous, then*

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2|+1} \ , \tag{3}$$

with $|\mathcal{I}_2|$ *the size of the second subset.*

Quick proof for intuition

- **Main argument:**

  By exchangeability, the rank of S_{n+1} among $\{S_i\}_{i\in\mathcal{I}_2}$ and S_{n+1} is uniformly distributed over the set $\{1,\ldots,|\mathcal{I}_2|+1\}$.

- In the continuous case, we obtain:

$$\begin{aligned}\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) &\overset{\text{(def)}}{=} \mathbb{P}(S_{n+1} \leq S_{(r^*)}) \\ &= \mathbb{P}\left(\text{rank}(S_{n+1}) \leq \lceil (1-\alpha)(|\mathcal{I}_2|+1)\rceil\right) \\ &= \frac{\lceil (1-\alpha)(|\mathcal{I}_2|+1)\rceil}{|\mathcal{I}_2|+1} \geq 1-\alpha\end{aligned}$$

Remainder: $\widehat{C}_r(x) = \{y : s(x,y) \leq S_{(r)}\}$

Quick proof for intuition

- **Main argument:**

  By exchangeability, the rank of S_{n+1} among $\{S_i\}_{i \in \mathcal{I}_2}$ and S_{n+1} is uniformly distributed over the set $\{1, \ldots, |\mathcal{I}_2| + 1\}$.

- In the continuous case, we obtain:

$$\begin{aligned}
\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) &\overset{(\text{def})}{=} \mathbb{P}(S_{n+1} \leq S_{(r^*)}) \\
&= \mathbb{P}\left(\text{rank}(S_{n+1}) \leq \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil\right) \\
&= \frac{\lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil}{|\mathcal{I}_2|+1} \geq 1-\alpha
\end{aligned}$$

Remainder: $\widehat{C}_r(x) = \{y : s(x,y) \leq S_{(r)}\}$

Quick proof for intuition

- **Main argument:**

  By exchangeability, the rank of S_{n+1} among $\{S_i\}_{i \in \mathcal{I}_2}$ and S_{n+1} is uniformly distributed over the set $\{1, \ldots, |\mathcal{I}_2| + 1\}$.

- In the continuous case, we obtain:

$$\begin{aligned}
\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) &\overset{\text{(def)}}{=} \mathbb{P}(S_{n+1} \leq S_{(r^*)}) \\
&= \mathbb{P}\left(\text{rank}(S_{n+1}) \leq \lceil (1-\alpha)(|\mathcal{I}_2| + 1) \rceil\right) \\
&= \frac{\lceil (1-\alpha)(|\mathcal{I}_2| + 1) \rceil}{|\mathcal{I}_2| + 1} \geq 1 - \alpha
\end{aligned}$$

Remainder: $\widehat{C}_r(x) = \{y : s(x, y) \leq S_{(r)}\}$

Summary of the Split CP method (Papadopoulos et al., 2002)

Input: $Z_1, \ldots, Z_n$, and $\alpha \in (0,1)$.

1. Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1$ and $\mathcal{I}_2$
2. Learn a predictor $\widehat{f}$ on $\{Z_i, i \in \mathcal{I}_1\}$
3. Compute scores $S_i = s(X_i, Y_i)$ for $i \in \mathcal{I}_2$
4. $S_{(r^*)}$ = the r^*-th smallest values in $\{S_i\}_{i \in \mathcal{I}_2}$ with $r^* = \lceil (1-\alpha)(|\mathcal{I}_2| + 1) \rceil$
5. Return the set $\widehat{C}_{r^*}(x) = \{y : s(y, x) \leq S_{(r^*)}\}$.

One example in regression:

- $s(x, y) = |y - \widehat{f}(x)|$
- $\widehat{C}_{r^*}(x)(x) = [\widehat{f}(x) - S_{(r^*)}, \widehat{f}(x) + S_{(r^*)}]$

Locally-Weighted Conformal Prediction

The split method with $s(x, y) = |y - \widehat{f}(x)|$ gives

$$\widehat{C}_r(x) = [\widehat{f}(x) - S_{(r)}, \widehat{f}(x) + S_{(r)}] \,.$$

$\longrightarrow$ fixed length! $\times$

Locally-Weighted CP (Papadopoulos et al., 2008):
A split conformal method with another score function $s(\cdot, \cdot)$:

$$s(x, y) = \frac{|y - \widehat{f}(x)|}{\hat{\rho}(x)} \,,$$

where $\hat{\rho}$ is an estimate of the conditional mean absolute deviation fitted one the samples in $\mathcal{I}_1$. The prediction set is now:

$$\widehat{C}_r(x) = [\widehat{f}(x) - \hat{\rho}(x) S_{(r)}, \widehat{f}(x) + \hat{\rho}(x) S_{(r)}] \,.$$

Locally-Weighted Conformal Prediction

The split method with $s(x, y) = |y - \widehat{f}(x)|$ gives

$$\widehat{C}_r(x) = [\widehat{f}(x) - S_{(r)}, \widehat{f}(x) + S_{(r)}] \ .$$

$\longrightarrow$ fixed length! $\times$

Locally-Weighted CP (Papadopoulos et al., 2008):
A split conformal method with another score function $s(\cdot, \cdot)$:

$$s(x, y) = \frac{|y - \widehat{f}(x)|}{\hat{\rho}(x)} \ ,$$

where $\hat{\rho}$ is an estimate of the conditional mean absolute deviation fitted one the samples in $\mathcal{I}_1$. The prediction set is now:

$$\widehat{C}_r(x) = [\widehat{f}(x) - \hat{\rho}(x) S_{(r)}, \widehat{f}(x) + \hat{\rho}(x) S_{(r)}] \ .$$

Locally-Weighted Conformal Prediction

The split method with $s(x, y) = |y - \widehat{f}(x)|$ gives

$$\widehat{C}_r(x) = [\widehat{f}(x) - S_{(r)}, \widehat{f}(x) + S_{(r)}] \ .$$

$\longrightarrow$ fixed length! $\times$

Locally-Weighted CP (Papadopoulos et al., 2008):
A split conformal method with another score function $s(\cdot, \cdot)$:

$$s(x, y) = \frac{|y - \widehat{f}(x)|}{\hat{\rho}(x)} \ ,$$

where $\hat{\rho}$ is an estimate of the conditional mean absolute deviation fitted one the samples in $\mathcal{I}_1$. The prediction set is now:

$$\widehat{C}_r(x) = [\widehat{f}(x) - \hat{\rho}(x) S_{(r)}, \widehat{f}(x) + \hat{\rho}(x) S_{(r)}] \ .$$

Standard split "v.s." Locally-Weighted

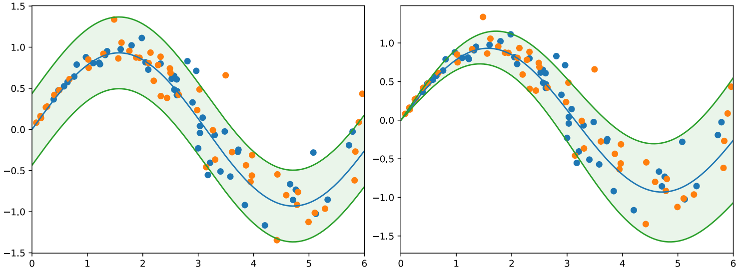


Figure: Left: Standard split CP. Right: Locally-Weighted CP.

Conformalized Quantile Regression (CQR)

Conformalized Quantile Regression (Romano et al., 2019):

Split CP method with another score function $s(\cdot, \cdot)$:

$$s(x, y) = \max\{\widehat{f}_{\alpha/2}(x) - y, y - \widehat{f}_{1-\alpha/2}(x)\} \ ,$$

where $(\widehat{f}_{\alpha/2}, \widehat{f}_{1-\alpha/2})$ are two quantile regressors fitted on $\{Z_i, i \in \mathcal{I}_1\}$.

The prediction set is now:

$$\widehat{C}_r(x) = [\widehat{f}_{\alpha/2}(x) - S_{(r)}, \widehat{f}_{1-\alpha/2}(x) + S_{(r)}] \ .$$

Locally-Weighted “v.s.” CQR

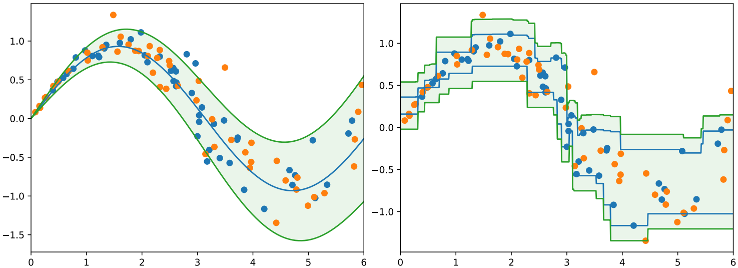


Figure: Left: Locally-Weighted CP. Right: CQR with random forest.

Comparison

- **Standard split CP Papadopoulos et al. (2002):**
  1. Works for any "black-box" predictor $\widehat{f}$
  2. Prediction set with constant size
- **Locally-weighted CP (Papadopoulos et al., 2008):**
  1. Works for any "black-box" predictor $\widehat{f}$
  2. Numerical instability
- **CQR (Romano et al., 2019):**
  1. Very adaptive
  2. **Does not** provide a set for a "black-box" predictor $\widehat{f}$

Comparison

- **Standard split CP Papadopoulos et al. (2002):**
  1. Works for any “black-box” predictor $\widehat{f}$
  2. Prediction set with constant size

- **Locally-weighted CP (Papadopoulos et al., 2008):**
  1. Works for any “black-box” predictor $\widehat{f}$
  2. Numerical instability

- **CQR (Romano et al., 2019):**
  1. Very adaptive
  2. **Does not** provide a set for a “black-box” predictor $\widehat{f}$

Comparison

- **Standard split CP Papadopoulos et al. (2002):**
  1. Works for any "black-box" predictor $\widehat{f}$
  2. Prediction set with constant size

- **Locally-weighted CP (Papadopoulos et al., 2008):**
  1. Works for any "black-box" predictor $\widehat{f}$
  2. Numerical instability

- **CQR (Romano et al., 2019):**
  1. Very adaptive
  2. **Does not** provide a set for a "black-box" predictor $\widehat{f}$

Comparison

- **Standard split CP Papadopoulos et al. (2002):**
  1. Works for any "black-box" predictor $\widehat{f}$
  2. Prediction set with constant size
- **Locally-weighted CP (Papadopoulos et al., 2008):**
  1. Works for any "black-box" predictor $\widehat{f}$
  2. Numerical instability
- **CQR (Romano et al., 2019):**
  1. Very adaptive
  2. **Does not** provide a set for a "black-box" predictor $\widehat{f}$

And in classification?

- $\mathcal{Y} = \{1, \cdots, K\}$
- $\widehat{\pi}_y(x)$ estimator of $\mathbb{P}(Y = y \mid X = x)$

Scores for classification

- **high-probability score:**

$$s(x, y) = -\widehat{\pi}_y(x)$$

- **(Romano et al., 2020):**

$$s(x, y) = \sum_{c=1}^{k} \widehat{\pi}_{(c)}(x)$$

where $\widehat{\pi}_{(1)} \geq \ldots \geq \widehat{\pi}_{(K)}$ and k is such that $\widehat{\pi}_{(k)} = \widehat{\pi}_y$

And in classification?

- $\mathcal{Y} = \{1, \cdots, K\}$
- $\widehat{\pi}_y(x)$ estimator of $\mathbb{P}(Y = y \mid X = x)$

Scores for classification

- **high-probability score:**

$$s(x, y) = -\widehat{\pi}_y(x)$$

- **(Romano et al., 2020):**

$$s(x, y) = \sum_{c=1}^{k} \widehat{\pi}_{(c)}(x)$$

where $\widehat{\pi}_{(1)} \geq \ldots \geq \widehat{\pi}_{(K)}$ and k is such that $\widehat{\pi}_{(k)} = \widehat{\pi}_y$

And in classification?

- $\mathcal{Y} = \{1, \cdots, K\}$
- $\widehat{\pi}_y(x)$ estimator of $\mathbb{P}(Y = y \mid X = x)$

Scores for classification

- **high-probability score:**

$$s(x, y) = -\widehat{\pi}_y(x)$$

- **(Romano et al., 2020):**

$$s(x, y) = \sum_{c=1}^{k} \widehat{\pi}_{(c)}(x)$$

where $\widehat{\pi}_{(1)} \geq \ldots \geq \widehat{\pi}_{(K)}$ and k is such that $\widehat{\pi}_{(k)} = \widehat{\pi}_y$

And in classification?

- $\mathcal{Y} = \{1, \cdots, K\}$
- $\widehat{\pi}_y(x)$ estimator of $\mathbb{P}(Y = y \mid X = x)$

Scores for classification

- **high-probability score:**

$$s(x, y) = -\widehat{\pi}_y(x)$$

- **(Romano et al., 2020):**

$$s(x, y) = \sum_{c=1}^{k} \widehat{\pi}_{(c)}(x)$$

where $\widehat{\pi}_{(1)} \geq \ldots \geq \widehat{\pi}_{(K)}$ and k is such that $\widehat{\pi}_{(k)} = \widehat{\pi}_y$

And in classification?

- $\mathcal{Y} = \{1, \cdots, K\}$
- $\widehat{\pi}_y(x)$ estimator of $\mathbb{P}(Y = y \mid X = x)$

Scores for classification

- **high-probability score:**

$$s(x, y) = -\widehat{\pi}_y(x)$$

- **(Romano et al., 2020):**

$$s(x, y) = \sum_{c=1}^{k} \widehat{\pi}_{(c)}(x)$$

where $\widehat{\pi}_{(1)} \geq \ldots \geq \widehat{\pi}_{(K)}$ and k is such that $\widehat{\pi}_{(k)} = \widehat{\pi}_y$

Another objective

Setup: n i.i.d. random variables $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n) \sim P$

For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

1. Marginal guarantee (previous slides)

$$\mathbb{P}(Y \in \widehat{C}(X)) \geq 1 - \alpha \ . \tag{4}$$

$\longrightarrow$ Probability taken on $\mathcal{D}_n = \{Z_1, \dots Z_n\}$ **and** $Z = (X, Y)$.

2. Training-conditional guarantee

Let $\alpha(\mathcal{D}_n) := \mathbb{P}(Y \notin C(X) \mid \mathcal{D}_n)$, given $\beta \in (0, 1)$ we want

$$\mathbb{P}\big(1 - \alpha(\mathcal{D}_n) \geq 1 - \alpha\big) \geq 1 - \beta \,. \tag{5}$$

Remark: $\mathbb{E}(1 - \alpha(\mathcal{D}_n)) = \mathbb{P}(Y \in C(X))$

Another objective

Setup: n i.i.d. random variables $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n) \sim P$

For $Z = (X, Y) \sim P$ and given $\alpha \in (0,1)$, construct $C(X)$ such that:

1. Marginal guarantee (previous slides)

$$\mathbb{P}(Y \in \widehat{C}(X)) \geq 1 - \alpha \,. \quad (4)$$

$\longrightarrow$ Probability taken on $\mathcal{D}_n = \{Z_1, \ldots Z_n\}$ **and** $Z = (X, Y)$.

2. Training-conditional guarantee

Let $\alpha(\mathcal{D}_n) := \mathbb{P}(Y \notin C(X) \mid \mathcal{D}_n)$, given $\beta \in (0,1)$ we want

$$\mathbb{P}\big(1 - \alpha(\mathcal{D}_n) \geq 1 - \alpha\big) \geq 1 - \beta \,. \quad (5)$$

Remark: $\mathbb{E}(1 - \alpha(\mathcal{D}_n)) = \mathbb{P}(Y \in C(X))$

Another objective

Setup: n i.i.d. random variables $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n) \sim P$

For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

1. Marginal guarantee (previous slides)

$$\mathbb{P}(Y \in \widehat{C}(X)) \geq 1 - \alpha \, . \tag{4}$$

$\longrightarrow$ Probability taken on $\mathcal{D}_n = \{Z_1, \dots Z_n\}$ **and** $Z = (X, Y)$.

2. Training-conditional guarantee

Let $\alpha(\mathcal{D}_n) := \mathbb{P}(Y \notin C(X) \mid \mathcal{D}_n)$, given $\beta \in (0, 1)$ we want

$$\mathbb{P}\big(1 - \alpha(\mathcal{D}_n) \geq 1 - \alpha\big) \geq 1 - \beta \, . \tag{5}$$

Remark: $\mathbb{E}(1 - \alpha(\mathcal{D}_n)) = \mathbb{P}(Y \in C(X))$

Another objective

Setup: n i.i.d. random variables $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n) \sim P$

For $Z = (X, Y) \sim P$ and given $\alpha \in (0, 1)$, construct $C(X)$ such that:

1. Marginal guarantee (previous slides)

$$\mathbb{P}(Y \in \widehat{C}(X)) \geq 1 - \alpha \; . \tag{4}$$

$\longrightarrow$ Probability taken on $\mathcal{D}_n = \{Z_1, \dots Z_n\}$ **and** $Z = (X, Y)$.

2. Training-conditional guarantee

Let $\alpha(\mathcal{D}_n) := \mathbb{P}(Y \notin C(X) \mid \mathcal{D}_n)$, given $\beta \in (0, 1)$ we want

$$\mathbb{P}\big(1 - \alpha(\mathcal{D}_n) \geq 1 - \alpha\big) \geq 1 - \beta \, . \tag{5}$$

Remark: $\mathbb{E}(1 - \alpha(\mathcal{D}_n)) = \mathbb{P}(Y \in C(X))$

Training-conditional coverage

Theorem

(Vovk, 2012) *In the i.i.d. setting, for any distribution P*

$$\mathbb{P}(1-\alpha_r(\mathcal{D}_n) \geq t) \geq \mathbb{P}(U_{(r)} \geq t) \tag{6}$$

where $\alpha_r(\mathcal{D}_n) = \mathbb{P}(Y \notin \widehat{C}_r(X) \mid \mathcal{D}_n)$ and $U_{(r)} \sim$ Beta$(r, |\mathcal{I}_2| - r + 1)$.

If the scores are continuous, $1-\alpha_r(\mathcal{D}_n) \sim$ Beta$(r, |\mathcal{I}_2| - r + 1)$.

Training-conditional coverage

Theorem

(Vovk, 2012) *In the i.i.d. setting, for any distribution* P

$$\mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq t) \geq \mathbb{P}(U_{(r)} \geq t) \tag{6}$$

where $\alpha_r(\mathcal{D}_n) = \mathbb{P}(Y \notin \widehat{C}_r(X) \mid \mathcal{D}_n)$ *and* $U_{(r)} \sim$ *Beta*$(r, |\mathcal{I}_2| - r + 1)$.

If the scores are continuous, $1 - \alpha_r(\mathcal{D}_n) \sim$ *Beta*$(r, |\mathcal{I}_2| - r + 1)$.

Training-conditional coverage

Theorem

(Vovk, 2012) *In the i.i.d. setting, for any distribution* P

$$\mathbb{P}(1-\alpha_r(\mathcal{D}_n) \geq t) \geq \mathbb{P}(U_{(r)} \geq t) \tag{6}$$

where $\alpha_r(\mathcal{D}_n) = \mathbb{P}(Y \notin \widehat{C}_r(X) \mid \mathcal{D}_n)$ *and* $U_{(r)} \sim$ *Beta*$(r, |\mathcal{I}_2| - r + 1)$.

If the scores are continuous, $1-\alpha_r(\mathcal{D}_n) \sim$ *Beta*$(r, |\mathcal{I}_2| - r + 1)$.

Quick proof for intuition

- $U_1, \cdots, U_{|\mathcal{I}_2|} \sim \mathcal{U}(0,1)$ then $U_{(r)} \sim \text{Beta}(r, |\mathcal{I}_2| - r + 1)$ where $U_{(1)} \leq \ldots \leq U_{(|\mathcal{I}_2|)}$

- **Main argument of the proof:**

  $S_1, \cdots, S_{|\mathcal{I}_2|} \sim F_S$ then $S_{(r)} \overset{d}{=} F_S^{-1}(U_{(r)})$

- $1 - \alpha_r(\mathcal{D}_n) = \mathbb{P}(S \leq S_{(r)} \mid \mathcal{D}_n) = F_S(S_{(r)})$

- $\mathbb{P}(F_S(S_{(r)}) \geq t) \geq \mathbb{P}(U_{(r)} \geq t)$

Quick proof for intuition

- $U_1, \cdots, U_{|\mathcal{I}_2|} \sim \mathcal{U}(0,1)$ then $U_{(r)} \sim \text{Beta}(r, |\mathcal{I}_2| - r + 1)$ where $U_{(1)} \leq \ldots \leq U_{(|\mathcal{I}_2|)}$

- **Main argument of the proof:**

  $S_1, \cdots, S_{|\mathcal{I}_2|} \sim F_S$ then $S_{(r)} \overset{d}{=} F_S^{-1}(U_{(r)})$

- $1 - \alpha_r(\mathcal{D}_n) = \mathbb{P}(S \leq S_{(r)} \mid \mathcal{D}_n) = F_S(S_{(r)})$

- $\mathbb{P}(F_S(S_{(r)}) \geq t) \geq \mathbb{P}(U_{(r)} \geq t)$

Quick proof for intuition

- $U_1, \cdots, U_{|\mathcal{I}_2|} \sim \mathcal{U}(0,1)$ then $U_{(r)} \sim \text{Beta}(r, |\mathcal{I}_2| - r + 1)$ where $U_{(1)} \leq \ldots \leq U_{(|\mathcal{I}_2|)}$

- **Main argument of the proof:**

  $S_1, \cdots, S_{|\mathcal{I}_2|} \sim F_S$ then $S_{(r)} \stackrel{d}{=} F_S^{-1}(U_{(r)})$

- $1 - \alpha_r(\mathcal{D}_n) = \mathbb{P}(S \leq S_{(r)} \mid \mathcal{D}_n) = F_S(S_{(r)})$

- $\mathbb{P}(F_S(S_{(r)}) \geq t) \geq \mathbb{P}(U_{(r)} \geq t)$

Quick proof for intuition

- $U_1, \cdots, U_{|\mathcal{I}_2|} \sim \mathcal{U}(0,1)$ then $U_{(r)} \sim \text{Beta}(r, |\mathcal{I}_2| - r + 1)$ where $U_{(1)} \leq \ldots \leq U_{(|\mathcal{I}_2|)}$
- **Main argument of the proof:**

  $S_1, \cdots, S_{|\mathcal{I}_2|} \sim F_S$ then $S_{(r)} \overset{d}{=} F_S^{-1}(U_{(r)})$
- $1 - \alpha_r(\mathcal{D}_n) = \mathbb{P}(S \leq S_{(r)} \mid \mathcal{D}_n) = F_S(S_{(r)})$
- $\mathbb{P}(F_S(S_{(r)}) \geq t) \geq \mathbb{P}(U_{(r)} \geq t)$

Quick proof for intuition

- $U_1, \cdots, U_{|\mathcal{I}_2|} \sim \mathcal{U}(0,1)$ then $U_{(r)} \sim \text{Beta}(r, |\mathcal{I}_2| - r + 1)$ where $U_{(1)} \leq \ldots \leq U_{(|\mathcal{I}_2|)}$

- **Main argument of the proof:**

  $S_1, \cdots, S_{|\mathcal{I}_2|} \sim F_S$ then $S_{(r)} \overset{d}{=} F_S^{-1}(U_{(r)})$

- $1 - \alpha_r(\mathcal{D}_n) = \mathbb{P}(S \leq S_{(r)} \mid \mathcal{D}_n) = F_S(S_{(r)})$

- $\mathbb{P}(F_S(S_{(r)}) \geq t) \geq \mathbb{P}(U_{(r)} \geq t)$

Training-conditional coverage

Theorem

(Vovk, 2012) *In the i.i.d. setting, for any distribution* P

$$\mathbb{P}(1-\alpha_r(\mathcal{D}_n) \geq t) \geq \mathbb{P}(U_{(r)} \geq t) \tag{7}$$

where $\alpha_r(\mathcal{D}_n) = \mathbb{P}(Y \notin \widehat{C}_r(X) \mid \mathcal{D}_n)$ *and* $U_{(r)} \sim$ *Beta*$(r, |\mathcal{I}_2| - r + 1)$.

$\longrightarrow \quad \mathbb{P}(1-\alpha_r(\mathcal{D}_n) \geq F^{-1}_{U_{(r)}}(\beta)) \geq \mathbb{P}(U_{(r)} \geq F^{-1}_{U_{(r)}}(\beta)) \geq 1-\beta$
with $F^{-1}_{U_{(r)}}$ the quantile function of Beta$(r, |\mathcal{I}_2| - r + 1)$.

If r is such that $F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha$, then

$$\mathbb{P}(1-\alpha_r(\mathcal{D}_n) \geq 1-\alpha) \geq 1-\beta \tag{8}$$

Training-conditional coverage

Theorem

(Vovk, 2012) *In the i.i.d. setting, for any distribution* P

$$\mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq t) \geq \mathbb{P}(U_{(r)} \geq t) \tag{7}$$

where $\alpha_r(\mathcal{D}_n) = \mathbb{P}(Y \notin \widehat{C}_r(X) \mid \mathcal{D}_n)$ *and* $U_{(r)} \sim$ *Beta*$(r, |\mathcal{I}_2| - r + 1)$.

$$\longrightarrow \quad \mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq F^{-1}_{U_{(r)}}(\beta)) \geq \mathbb{P}(U_{(r)} \geq F^{-1}_{U_{(r)}}(\beta)) \geq 1 - \beta$$

with $F^{-1}_{U_{(r)}}$ the quantile function of Beta$(r, |\mathcal{I}_2| - r + 1)$.

If r is such that $F^{-1}_{U_{(r)}}(\beta) \geq 1 - \alpha$, then

$$\mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq 1 - \alpha) \geq 1 - \beta \tag{8}$$

Training-conditional coverage

Theorem

(Vovk, 2012) In the i.i.d. setting, for any distribution P

$$\mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq t) \geq \mathbb{P}(U_{(r)} \geq t) \tag{7}$$

where $\alpha_r(\mathcal{D}_n) = \mathbb{P}(Y \notin \widehat{C}_r(X) \mid \mathcal{D}_n)$ and $U_{(r)} \sim$ Beta$(r, |\mathcal{I}_2| - r + 1)$.

$\longrightarrow \quad \mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq F^{-1}_{U_{(r)}}(\beta)) \geq \mathbb{P}(U_{(r)} \geq F^{-1}_{U_{(r)}}(\beta)) \geq 1 - \beta$
with $F^{-1}_{U_{(r)}}$ the quantile function of Beta$(r, |\mathcal{I}_2| - r + 1)$.

If r is such that $F^{-1}_{U_{(r)}}(\beta) \geq 1 - \alpha$, then

$$\mathbb{P}(1 - \alpha_r(\mathcal{D}_n) \geq 1 - \alpha) \geq 1 - \beta \tag{8}$$

Training-conditional Split CP

- Find r_c such that

$$r_c = \arg\min_r \left\{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \right\}$$

- Construct $\widehat{C}_{r_c}(x)$ using the split CP method:

$$\widehat{C}_{r_c}(x) = \{y : s(x,y) \leq S_{(r_c)}\} \overset{\text{(ex)}}{=} [\widehat{f}(x) - S_{(r_c)}, \widehat{f}(x) + S_{(r_c)}]$$

$\longrightarrow$ By construction, $\widehat{C}_{r_c}(x)$ is training-conditionally valid.

Training-conditional Split CP

- Find r_c such that

$$r_c = \arg\min_r \left\{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \right\}$$

- Construct $\widehat{C}_{r_c}(x)$ using the split CP method:

$$\widehat{C}_{r_c}(x) = \{y : s(x,y) \leq S_{(r_c)}\} \overset{\text{(ex)}}{=} [\widehat{f}(x) - S_{(r_c)}, \widehat{f}(x) + S_{(r_c)}]$$

$\longrightarrow$ By construction, $\widehat{C}_{r_c}(x)$ is training-conditionally valid.

Training-conditional Split CP

- Find r_c such that

$$r_c = \arg\min_r \left\{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \right\}$$

- Construct $\widehat{C}_{r_c}(x)$ using the split CP method:

$$\widehat{C}_{r_c}(x) = \{y : s(x,y) \leq S_{(r_c)}\} \overset{\text{(ex)}}{=} [\widehat{f}(x) - S_{(r_c)}, \widehat{f}(x) + S_{(r_c)}]$$

$\longrightarrow$ By construction, $\widehat{C}_{r_c}(x)$ is training-conditionally valid.

Training-conditional Split CP

- Find r_c such that

$$r_c = \arg\min_r \left\{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \right\}$$

- Construct $\widehat{C}_{r_c}(x)$ using the split CP method:

$$\widehat{C}_{r_c}(x) = \{y : s(x,y) \leq S_{(r_c)}\} \overset{\text{(ex)}}{=} [\widehat{f}(x) - S_{(r_c)}, \widehat{f}(x) + S_{(r_c)}]$$

$\longrightarrow$ By construction, $\widehat{C}_{r_c}(x)$ is training-conditionally valid.

Upper bound

Theorem
(Vovk, 2012)

In the i.i.d. setting, for any distribution P and any $\beta \in [0, 0.5)$, if the scores are continuous

$$\mathbb{P}\left(1-\alpha \leq 1-\alpha_{r_c}(\mathcal{D}_N) \leq 1-\alpha+\sqrt{\frac{\log(1/\beta)}{2|\mathcal{I}_2|}}\right) \geq 1-2\beta\ , \qquad (9)$$

where $\widehat{C}_{r_c}(X)$ is returned by the training-conditional split CP method.

Upper bound

Theorem

(Vovk, 2012)

In the i.i.d. setting, for any distribution P and any $\beta \in [0, 0.5)$, if the scores are continuous

$$\mathbb{P}\left(1 - \alpha \leq 1 - \alpha_{r_c}(\mathcal{D}_N) \leq 1 - \alpha + \sqrt{\frac{\log(1/\beta)}{2|\mathcal{I}_2|}}\right) \geq 1 - 2\beta \ , \qquad (9)$$

where $\widehat{C}_{r_c}(X)$ is returned by the training-conditional split CP method.

In summary

With split CP:

- If $r^* = \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil$, then

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \geq 1-\alpha \ , \tag{10}$$

$\longrightarrow$ Marginal guarantee

- If $r_c = \arg\min_r \left\{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \right\}$, then

$$\mathbb{P}(1-\alpha_{r_c}(\mathcal{D}_n) \geq 1-\alpha) \geq 1-\beta \tag{11}$$

$\longrightarrow$ Training-conditional guarantee

In summary

With split CP:

- If $r^* = \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil$, then

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \geq 1-\alpha \; , \tag{10}$$

$\longrightarrow$ Marginal guarantee

- If $r_c = \arg\min_r \{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \}$, then

$$\mathbb{P}(1-\alpha_{r_c}(\mathcal{D}_n) \geq 1-\alpha) \geq 1-\beta \tag{11}$$

$\longrightarrow$ Training-conditional guarantee

In summary

With split CP:

- If $r^* = \lceil (1-\alpha)(|\mathcal{I}_2|+1) \rceil$, then

$$\mathbb{P}(Y \in \widehat{C}_{r^*}(X)) \geq 1-\alpha \ , \tag{10}$$

$\longrightarrow$ Marginal guarantee

- If $r_c = \arg\min_r \left\{ F^{-1}_{U_{(r)}}(1-\beta) : F^{-1}_{U_{(r)}}(\beta) \geq 1-\alpha \right\}$, then

$$\mathbb{P}(1-\alpha_{r_c}(\mathcal{D}_n) \geq 1-\alpha) \geq 1-\beta \tag{11}$$

$\longrightarrow$ Training-conditional guarantee

A conditional guarantee

Overall, we want in fact a set such that:

$$\mathbb{P}\left(Y \in C(X) \mid X = x\right) \geq 1 - \alpha \ , \tag{12}$$

for all P and almost all x.

Impossibility result (Vovk, 2012)

If $\widehat{C}$, constructed from a finite sample, satisfies the above equation, then for all distributions P , it holds that

$$\mathbb{E}(\text{leb}(\widehat{\mathcal{C}}(x))) = \infty$$

at almost all points x. Here, $\text{leb}(\cdot)$ is the Lebesgue measure.

$\longrightarrow$ distribution-free conditional guarantee is impossible to attain in any meaningful sense.

A conditional guarantee

Overall, we want in fact a set such that:

$$\mathbb{P}\left(Y \in C(X) \mid X = x\right) \geq 1 - \alpha \ , \tag{12}$$

for all P and almost all x.

Impossibility result (Vovk, 2012)

If $\widehat{C}$, constructed from a finite sample, satisfies the above equation, then for all distributions P , it holds that

$$\mathbb{E}(\text{leb}(\widehat{\mathcal{C}}(x))) = \infty$$

at almost all points x. Here, leb$(\cdot)$ is the Lebesgue measure.

$\longrightarrow$ distribution-free conditional guarantee is impossible to attain in any meaningful sense.

A conditional guarantee

Overall, we want in fact a set such that:

$$\mathbb{P}\left(Y \in C(X) \mid X = x\right) \geq 1 - \alpha \ , \tag{12}$$

for all P and almost all x.

Impossibility result (Vovk, 2012)

If $\widehat{C}$, constructed from a finite sample, satisfies the above equation, then for all distributions P , it holds that

$$\mathbb{E}(\text{leb}(\widehat{\mathcal{C}}(x))) = \infty$$

at almost all points x. Here, leb$(\cdot)$ is the Lebesgue measure.

$\longrightarrow$ distribution-free conditional guarantee is impossible to attain in any meaningful sense.

A conditional guarantee

Overall, we want in fact a set such that:

$$\mathbb{P}\left(Y \in C(X) \mid X = x\right) \geq 1 - \alpha \ , \tag{12}$$

for all P and almost all x.

Impossibility result (Vovk, 2012)

If $\widehat{C}$, constructed from a finite sample, satisfies the above equation, then for all distributions P , it holds that

$$\mathbb{E}(\text{leb}(\widehat{\mathcal{C}}(x))) = \infty$$

at almost all points x. Here, leb$(\cdot)$ is the Lebesgue measure.

$\longrightarrow$ distribution-free conditional guarantee is impossible to attain in any meaningful sense.

Other conditional guarantees

Two lines of research

- Approximate conditional guarantee:

$$\mathbb{P}\left(Y \in C(X) \mid X \in \mathcal{A}\right) \geq 1 - \alpha \; , \tag{13}$$

  e.g. (Lei and Wasserman, 2014; Gibbs et al., 2023)

- Asymptotic conditional coverage:

$$\mathbb{P}\left(Y \in \widehat{C}(X) \mid X = x\right) \xrightarrow[|\mathcal{I}_2| \to \infty]{} 1 - \alpha \; , \tag{14}$$

  e.g. (Chernozhukov et al., 2021; Sesia and Romano, 2021; Izbicki et al., 2022)

Other conditional guarantees

Two lines of research

- Approximate conditional guarantee:

$$\mathbb{P}\left(Y \in C(X) \mid X \in \mathcal{A}\right) \geq 1 - \alpha \ , \quad (13)$$

  e.g. (Lei and Wasserman, 2014; Gibbs et al., 2023)

- Asymptotic conditional coverage:

$$\mathbb{P}\left(Y \in \widehat{C}(X) \mid X = x\right) \xrightarrow[|\mathcal{I}_2| \to \infty]{} 1 - \alpha \ , \quad (14)$$

  e.g. (Chernozhukov et al., 2021; Sesia and Romano, 2021; Izbicki et al., 2022)

Other conditional guarantees

Two lines of research

- Approximate conditional guarantee:

$$\mathbb{P}\left(Y \in C(X) \mid X \in \mathcal{A}\right) \geq 1 - \alpha \ , \tag{13}$$

e.g. (Lei and Wasserman, 2014; Gibbs et al., 2023)

- Asymptotic conditional coverage:

$$\mathbb{P}\left(Y \in \widehat{C}(X) \mid X = x\right) \xrightarrow[|\mathcal{I}_2| \to \infty]{} 1 - \alpha \ , \tag{14}$$

e.g. (Chernozhukov et al., 2021; Sesia and Romano, 2021; Izbicki et al., 2022)

Avoiding data-splitting

Issue with split CP: Split the data = loss in accuracy for $\widehat{f}$.

Other CP methods:

1. Full conformal prediction
2. Jackknife+
3. CV+

Avoiding data-splitting

Issue with split CP: Split the data = loss in accuracy for $\widehat{f}$.

Other CP methods:

1. Full conformal prediction
2. Jackknife+
3. CV+

Avoiding data-splitting

Issue with split CP: Split the data = loss in accuracy for $\widehat{f}$.

Other CP methods:

1. Full conformal prediction
2. Jackknife+
3. CV+

Full Conformal Prediction (Vovk et al., 2005)

Input: $Z_1, \ldots, Z_n$, and $\alpha \in (0,1)$.

1. For any $y \in \mathcal{Y}$, construct $\widehat{f}_y$ with $Z_1, \ldots, Z_n$, **and** (X, y)
2. $S_i^y = |Y_i - \widehat{f}_y(X_i)|$ and $S_{n+1}^y = |y - \widehat{f}_y(X)|$
3. If $S_{n+1}^y \leq S_{(k^*)}$ with $k^* = \lceil (1-\alpha)(n+1) \rceil$ then add y to the set.

$\longrightarrow$ These steps must be repeated for each value of y.

$\longrightarrow$ In practice, we must restrict us to a discrete grid of trial values y.

Full Conformal Prediction

Theorem

(Vovk et al., 2005)

If data are exchangeable and $\widehat{f}$ is symmetric, the set returns by the Full CP method satisfies

$$\mathbb{P}(Y \in \widehat{C}(X)) \geq 1 - \alpha \,. \tag{15}$$

Moreover, if we assume that the scores $\{S_i^y\}_i$ are continuous, then

$$\mathbb{P}(Y \in \widehat{C}(X)) \leq 1 - \alpha + \frac{1}{n+1} \,. \tag{16}$$

No training-conditional coverage guarantee for full CP

Theorem

(Bian and Barber, 2022)

For any sample size $n \geq 2$ and any distribution P for which the marginal P_X is nonatomic, there exists a symmetric and deterministic regression algorithm $\widehat{f}$ such that the ***full conformal method*** *satisfies*

$$\mathbb{P}\left(1 - \alpha(\mathcal{D}_n) \leq n^{-2}\right) \geq \alpha - 6\sqrt{\frac{\log n}{n}}\ . \qquad (17)$$

$\longrightarrow$ Without additional assumptions on P and/or on $\widehat{f}$, we cannot avoid the worst-case scenario.

Jackknife+
(Barber et al., 2021)

Input: $Z_1, \ldots, Z_n$, and $\alpha \in (0, 1)$.

1. For i in $\{1, \ldots, n\}$:

   Learn $\widehat{f}_{[n]\setminus\{i\}}$, the model fitted to the training data with the i-th point removed

   $$S_i^- = \left(\widehat{f}_{[n]\setminus\{i\}}(X_{n+1}) - R_i\right)$$

   $$S_i^+ = \left(\widehat{f}_{[n]\setminus\{i\}}(X_{n+1}) + R_i\right),$$

   where $R_i = |\widehat{f}_{[n]\setminus\{i\}}(X_i) - Y_i|$

2. Return $\widehat{C}(X_{n+1}) = [S^-_{(n+1-k^*)}, S^+_{(k^*)}]$ with $k^* = \lceil (n+1)(1-\alpha) \rceil$

CV+
(Barber et al., 2021)

Input: $Z_1, \ldots, Z_n$, $A_1 \cap \ldots \cap A_K = [n]$ a partition of the training data into K subsets of size n/K, and $\alpha \in (0,1)$.

1. For k in $\{1, \ldots, K\}$:

   Learn $\widehat{f}_{[n]\setminus A_k}$, the model fitted to the training data with the k-th fold A_k removed

   $S^-_{k,i} = \widehat{f}_{[n]\setminus A_k}(X_{n+1}) - R_i$

   $S^+_{k,i} = \widehat{f}_{[n]\setminus A_k}(X_{n+1}) + R_i,$

   with $i \in A_k$ and $R_i = |\widehat{f}_{[n]\setminus A_k}(X_i) - Y_i|$

2. Return $\widehat{C}(X_{n+1}) = [S^-_{(n+1-k^*)}, S^+_{(k^*)}]$ with $k^* = \lceil (n+1)(1-\alpha) \rceil$

A summary of all the results

Marginal guarantee:

- Ok for all the methods (if $\widehat{f}$ symmetric and data exchangeable)

Training conditional coverage guarantee:

- Ok for Split CP and CV+ method (if data i.i.d.)
- Not possible for full CP or jackknife+ methods without additional assumptions

Conditional guarantee:

- Not possible without additional assumptions

A summary of all the results

Marginal guarantee:

- Ok for all the methods (if $\widehat{f}$ symmetric and data exchangeable)

Training conditional coverage guarantee:

- Ok for Split CP and CV+ method (if data i.i.d.)
- Not possible for full CP or jackknife+ methods without additional assumptions

Conditional guarantee:

- Not possible without additional assumptions

A summary of all the results

Marginal guarantee:

- Ok for all the methods (if $\widehat{f}$ symmetric and data exchangeable)

Training conditional coverage guarantee:

- Ok for Split CP and CV+ method (if data i.i.d.)
- Not possible for full CP or jackknife+ methods without additional assumptions

Conditional guarantee:

- Not possible without additional assumptions

A summary of all the results

Marginal guarantee:

- Ok for all the methods (if $\widehat{f}$ symmetric and data exchangeable)

Training conditional coverage guarantee:

- Ok for Split CP and CV+ method (if data i.i.d.)
- Not possible for full CP or jackknife+ methods without additional assumptions

Conditional guarantee:

- Not possible without additional assumptions

A final important question

Are coverage guarantees enough? $\longrightarrow$ No

- Take $\widehat{C}(X) = \mathbb{R}$ with probability $1-\alpha$ and $\widehat{C}(X) = \emptyset$ with probability α
  $\longrightarrow \mathbb{P}(Y \in \widehat{C}(X)) = 1 - \alpha$

We must look at the **size** of $\widehat{C}(x)$

Size of $\widehat{C}(x)$

- Asymptotic results, under strong assumptions
- Empirical evaluations

A final important question

Are coverage guarantees enough? $\longrightarrow$ No

- Take $\widehat{C}(X) = \mathbb{R}$ with probability $1-\alpha$ and $\widehat{C}(X) = \emptyset$ with probability α
$\longrightarrow \mathbb{P}(Y \in \widehat{C}(X)) = 1-\alpha$

We must look at the **size** of $\widehat{C}(x)$

Size of $\widehat{C}(x)$

- Asymptotic results, under strong assumptions
- Empirical evaluations

A final important question

Are coverage guarantees enough? $\longrightarrow$ No

- Take $\widehat{C}(X) = \mathbb{R}$ with probability $1 - \alpha$ and $\widehat{C}(X) = \emptyset$ with probability α
$\longrightarrow \mathbb{P}(Y \in \widehat{C}(X)) = 1 - \alpha$

We must look at the **size** of $\widehat{C}(x)$

Size of $\widehat{C}(x)$

- Asymptotic results, under strong assumptions
- Empirical evaluations

A final important question

Are coverage guarantees enough? $\longrightarrow$ No

- Take $\widehat{C}(X) = \mathbb{R}$ with probability $1-\alpha$ and $\widehat{C}(X) = \emptyset$ with probability α
  $\longrightarrow \mathbb{P}(Y \in \widehat{C}(X)) = 1-\alpha$

We must look at the **size** of $\widehat{C}(x)$

Size of $\widehat{C}(x)$

- Asymptotic results, under strong assumptions
- Empirical evaluations

A final important question

Are coverage guarantees enough? $\longrightarrow$ No

- Take $\widehat{C}(X) = \mathbb{R}$ with probability $1-\alpha$ and $\widehat{C}(X) = \emptyset$ with probability α
$\longrightarrow \mathbb{P}(Y \in \widehat{C}(X)) = 1 - \alpha$

We must look at the **size** of $\widehat{C}(x)$

Size of $\widehat{C}(x)$

- Asymptotic results, under strong assumptions
- Empirical evaluations

CP in practice

Setup

- Evaluation on 5 regression data sets
- Split CP
- score function is $s(x, y) =$ "CQR with Quantile Random Forest"
- $\alpha = 0.1$ and $\beta = 0.2$

CP in practice

Setup

- Evaluation on 5 regression data sets
- Split CP
- score function is $s(x, y) =$ "CQR with Quantile Random Forest"
- $\alpha = 0.1$ and $\beta = 0.2$

CP in practice

Metrics

On 50 random training-test split, we compute:

- Coverage (on the test set)
- Length of the returned set

CP in practice

Metrics

On 50 random training-test split, we compute:

- Coverage (on the test set)
- Length of the returned set

Results on all the data sets

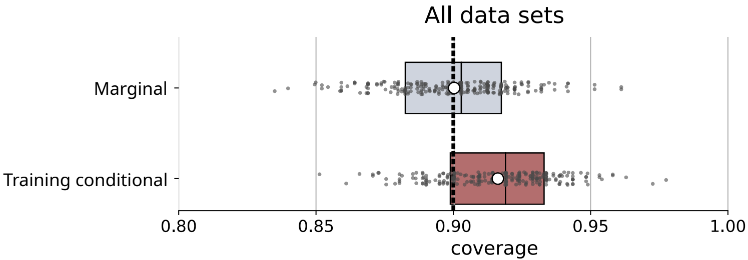


Figure: Empirical coverages of prediction intervals ($\alpha = 0.1$). The white circle represents the mean.

Before left of the box: 20% of the points
After right of the box: 20% of the points

Result on one data set

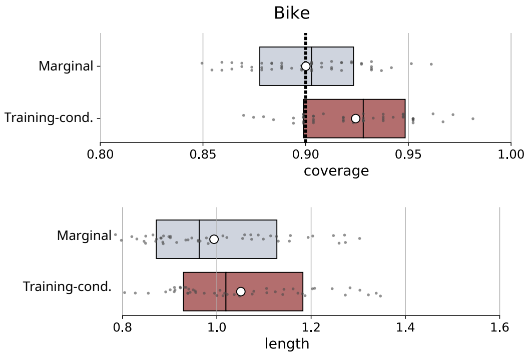


Figure: Coverage (top) and average length (bottom) of prediction intervals. The white circle represents the mean.

Other important topics

Beyond the standard setting

1. Online setting
2. Weighted CP
3. Decentralized setting

Other important topics

Beyond the standard setting

1. Online setting
2. Weighted CP
3. Decentralized setting

Online CP

Setup

- Sequentially observe pairs $\{(X_t, Y_t), t \geq 1\}$
- No assumption on the data

Objective

- Control of the False Coverage Proportion (FCP)

$$1/t \cdot \sum_{k=1}^{t} 1\{Y_t \notin \widehat{C}_t(X_t)\} - \alpha \tag{18}$$

e.g. (Gibbs and Candes, 2021; Zaffran et al., 2022; Angelopoulos et al., 2024)

Weighted CP

Setup

- $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{Y|X} \times P_X$
- $(X, Y) \sim P_{Y|X} \times Q_X$ (covariate shift)

Objective

- Construct a marginally valid set for Y

How?

- Give more importance to calibration points that are closer in distribution to the test point:
  1. Estimate the likelihood ratio dQ_X/dP_X
  2. Use a "weighted empirical quantile" to construct the set

e.g. (Tibshirani et al., 2019)

Decentralized CP

Setup

- m agents and a central server
- n_j i.i.d. random variables per agent
  $\longrightarrow$ i-th data of agent j: $Z_i^j = (X_i^j, Y_i^j) \sim P_j$

Objectives

Construct a set with guarantees when:

1. Only one round of communication
2. Heterogeneous data

e.g. (Humbert et al., 2023; Lu et al., 2023; Plassier et al., 2023; Humbert et al., 2024)

Take home messages

1. Conformal prediction works (both in theory and in practice)
2. Easy to implement on top of any ML methods
3. Coverage is not all you need
   $\longrightarrow$ You have to look at the size of the sets

Nice recent reference

Theoretical Foundations of Conformal Prediction (2024)
by A N. Angelopoulos, R F Barber, and S Bates

Take home messages

1. Conformal prediction works (both in theory and in practice)
2. Easy to implement on top of any ML methods
3. Coverage is not all you need
   $\longrightarrow$ You have to look at the size of the sets

Nice recent reference

Theoretical Foundations of Conformal Prediction (2024)
by A N. Angelopoulos, R F Barber, and S Bates

Take home messages

1. Conformal prediction works (both in theory and in practice)
2. Easy to implement on top of any ML methods
3. Coverage is not all you need
   $\longrightarrow$ You have to look at the size of the sets

Nice recent reference

Theoretical Foundations of Conformal Prediction (2024)
by A N. Angelopoulos, R F Barber, and S Bates

Take home messages

1. Conformal prediction works (both in theory and in practice)

2. Easy to implement on top of any ML methods

3. Coverage is not all you need
   $\longrightarrow$ You have to look at the size of the sets

Nice recent reference

Theoretical Foundations of Conformal Prediction (2024)
by A N. Angelopoulos, R F Barber, and S Bates

Take home messages

1. Conformal prediction works (both in theory and in practice)
2. Easy to implement on top of any ML methods
3. Coverage is not all you need
   $\longrightarrow$ You have to look at the size of the sets

Nice recent reference

Theoretical Foundations of Conformal Prediction (2024)
by A N. Angelopoulos, R F Barber, and S Bates

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024). Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1).

Bian, M. and Barber, R. F. (2022). Training-conditional coverage for distribution-free predictive inference. *arXiv preprint arXiv:2205.03647*.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*.

Humbert, P., Bars, B. L., Bellet, A., and Arlot, S. (2024). Marginal and training-conditional guarantees in one-shot federated conformal prediction. *arXiv preprint arXiv:2405.12567*.

Humbert, P., Le Bars, B., Bellet, A., and Arlot, S. (2023). One-shot federated conformal prediction. In *International Conference on Machine Learning*, pages 14153–14177. PMLR.

Izbicki, R., Shimizu, G., and Stern, R. B. (2022). Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96.

Lu, C., Yu, Y., Karimireddy, S. P., Jordan, M., and Raskar, R. (2023). Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR.

Papadopoulos, H., Gammerman, A., and Vovk, V. (2008). Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer.

Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., and Panov, M. (2023). Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, pages 27907–27947. PMLR.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR.