Distribution-Free Predictive Uncertainty Quantification: Strengths and Limits of Conformal Prediction

Aymeric Dieuleveut & Margaux Zaffran

July 15th, 2024

40th Conference on Uncertainty in Artifical Intelligence (UAI)

Why are we all here today?

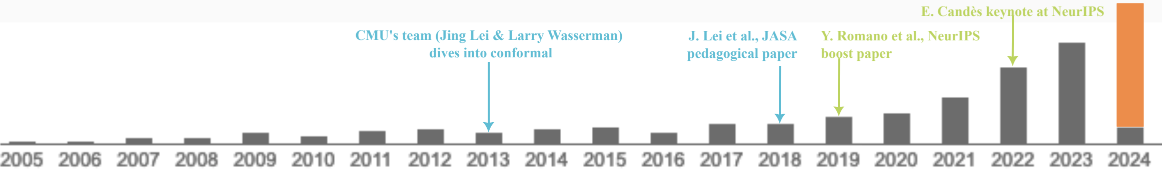
Figure 1: us

Why are we all here today?

(Slides available on our webpages)

(Slides available on our webpages)

- Because Conformal Prediction has been a **popular** topic recently.



Vovk et al. (2005) algorithmic learning in a random world cite count.

Why are we all here today?

(Slides available on our webpages)

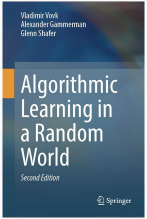
- Because Conformal Prediction has been a **popular** topic recently.
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.

  Successfully applied to
  - Medical applications
  - Markets / demand forecasting
  - Computer Vision

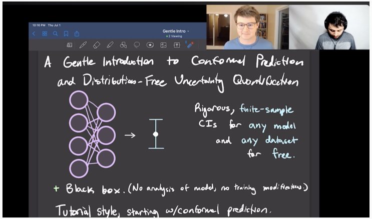
Why are we all here today?

(Slides available on our webpages)

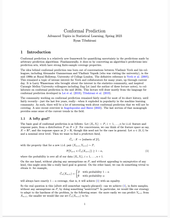
- Because Conformal Prediction has been a **popular** topic recently.
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.
- To be part of the **diffusion** effort that many colleagues are making.



Book reference: Vovk et al. (2005)
(new edition in 2022)



A gentle tutorial: Angelopoulos and Bates (2023)
+ Videos playlist

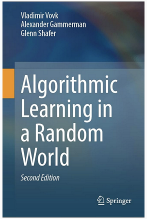

Conformal Prediction
Advanced Topics in Statistical Learning, Spring 2023
Ryan Tibshirani
1 Introduction
1.1 A lofty goal?

R. J. Tibshirani
introductive lecture's notes

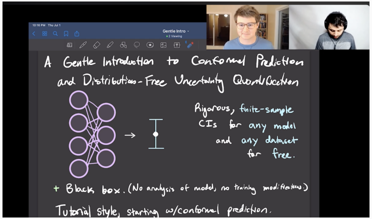
Why are we all here today?

(Slides available on our webpages)

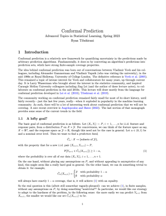
- Because Conformal Prediction has been a **popular** topic recently.
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.
- To be part of the **diffusion** effort that many colleagues are making.



Book reference: Vovk et al. (2005)
(new edition in 2022)



A gentle tutorial: Angelopoulos and Bates (2023)
+ Videos playlist


Conformal Prediction
Advanced Topics in Statistical Learning, Spring 2023
Ryan Tibshirani

R. J. Tibshirani
introductive lecture's notes

→ **Based on material freely accessible on this webpage, including sources.**
Feel free to reuse these contents for presentations or teaching!

Goals and disclaimers

Goals

- Provide a detailed introduction to the basics
- Demystify the results: fair introduction with limits
- Give you tools to leverage those techniques in your own fields

Disclaimers

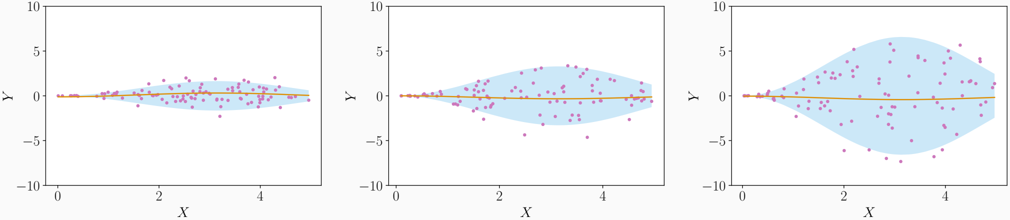
- Many people contributed to the domain - list of references may not be exhaustive
- Multiple other excellent resources

On the importance of quantifying uncertainty

- **Obvious in most applications - weather, medical, markets**

On the importance of quantifying uncertainty

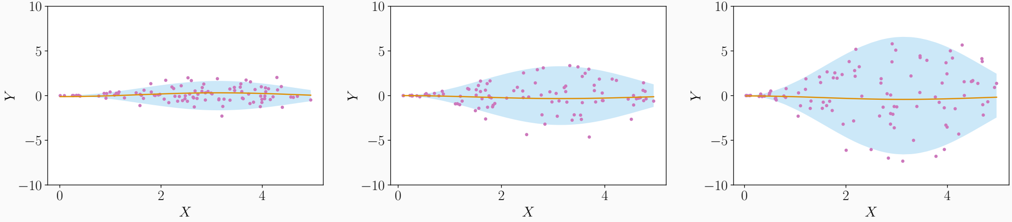
- **Obvious in most applications - weather, medical, markets**
- **Mathematically**



↪ Same "best" predictor, yet 3 distinct underlying phenomena!

On the importance of quantifying uncertainty

- **Obvious in most applications - weather, medical, markets**
- **Mathematically**

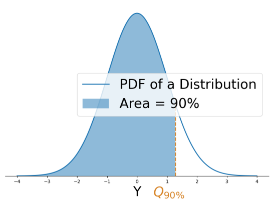


$\hookrightarrow$ Same “best” predictor, yet 3 distinct underlying phenomena!

$\Longrightarrow$ Quantifying uncertainty conveys this information.

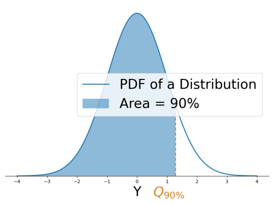
Reminder about quantiles

- Quantile level $\beta \in [0, 1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \leq t) \geq \beta\}$



Reminder about quantiles

- Quantile level $\beta \in [0, 1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \leq t) \geq \beta\}$
- Empirical quantile
  $q_\beta(Y_1, \dots, Y_n)$
  $:= \lceil \beta \times n \rceil$ smallest value of $(Y_1, \dots, Y_n)$



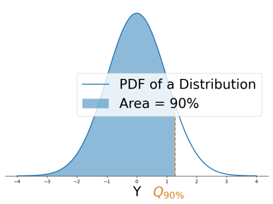
Reminder about quantiles

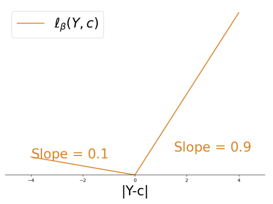
- Quantile level $\beta \in [0,1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \leq t) \geq \beta\}$
- Empirical quantile

  $q_\beta(Y_1, \ldots, Y_n)$

  $:= \lceil \beta \times n \rceil$ smallest value of $(Y_1, \ldots, Y_n)$
- Pinball loss

$$\begin{aligned}\ell_\beta(Y, Y') =& \beta|Y - Y'|\mathbb{1}_{\{Y-Y' \geq 0\}} \\ &+ (1-\beta)|Y - Y'|\mathbb{1}_{\{Y-Y' \leq 0\}}\end{aligned}$$





Reminder about quantiles

- Quantile level $\beta \in [0,1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \leq t) \geq \beta\}$
- Empirical quantile

  $q_\beta(Y_1, \ldots, Y_n)$

  $:= \lceil \beta \times n \rceil$ smallest value of $(Y_1, \ldots, Y_n)$
- Pinball loss

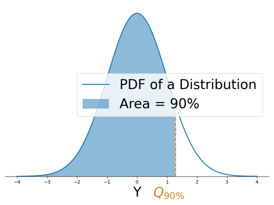
$$\ell_\beta(Y, Y') = \beta |Y - Y'| \mathbb{1}_{\{Y - Y' \geq 0\}} + (1-\beta)|Y - Y'| \mathbb{1}_{\{Y - Y' \leq 0\}}$$

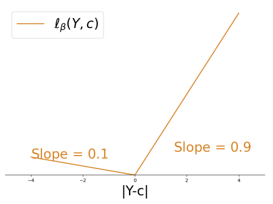
Associated risk:

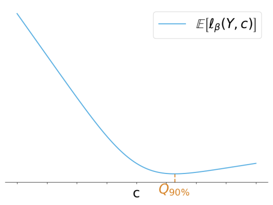
$$\mathrm{Risk}_{\ell_\beta}(c) = \mathbb{E}\left[\ell_\beta(Y, c)\right]$$

Link to quantile:

$$Q_Y(\beta) = \arg\min_{c \in \mathbb{R}} \mathrm{Risk}_{\ell_\beta}(c)$$







Reminder about quantiles

- Quantile level $\beta \in [0,1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \le t) \ge \beta\}$
- Empirical quantile

  $q_\beta(Y_1, \dots, Y_n)$

  $:= \lceil \beta \times n \rceil$ smallest value of $(Y_1, \dots, Y_n)$
- Pinball loss

  $$\begin{aligned}\ell_\beta(Y, Y') = &\beta |Y - Y'| \mathbb{1}_{\{Y - Y' \ge 0\}} \\ &+ (1-\beta)|Y - Y'| \mathbb{1}_{\{Y - Y' \le 0\}}\end{aligned}$$

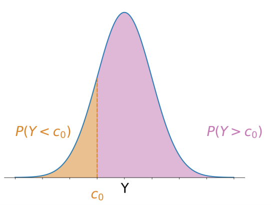
  Associated risk:

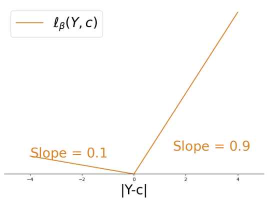
  $$\mathrm{Risk}_{\ell_\beta}(c) = \mathbb{E}\left[\ell_\beta(Y, c)\right]$$

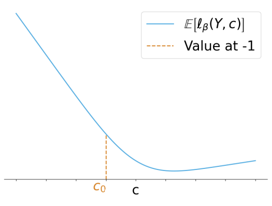
  Link to quantile:

  $$Q_Y(\beta) = \arg\min_{c \in \mathbb{R}} \mathrm{Risk}_{\ell_\beta}(c)$$

  Proof: sub-differential







Reminder about quantiles

- Quantile level $\beta \in [0,1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \leq t) \geq \beta\}$
- Empirical quantile

  $q_\beta(Y_1, \ldots, Y_n)$

  $:= \lceil \beta \times n \rceil$ smallest value of $(Y_1, \ldots, Y_n)$
- Pinball loss

$$\ell_\beta(Y, Y') = \beta |Y - Y'| \mathbb{1}_{\{Y - Y' \geq 0\}} + (1-\beta)|Y - Y'| \mathbb{1}_{\{Y - Y' \leq 0\}}$$

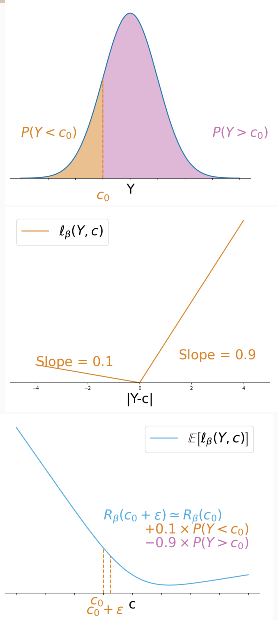
  Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(c) = \mathbb{E}\left[\ell_\beta(Y, c)\right]$$

  Link to quantile:

$$Q_Y(\beta) = \arg\min_{c \in \mathbb{R}} \mathrm{Risk}_{\ell_\beta}(c)$$

  Proof: sub-differential



Median

Example (a special quantile: the median)**.**

$\beta = 0.5$

$\hookrightarrow$ $Q_Y(0.5)$ represents the median of the distribution of Y.

$\hookrightarrow$ $Q_Y(0.5) = argmin_c \mathbb{E}\left[|Y - c|\right]$.

Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$

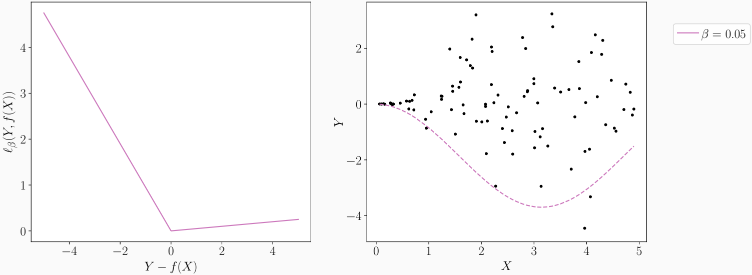
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



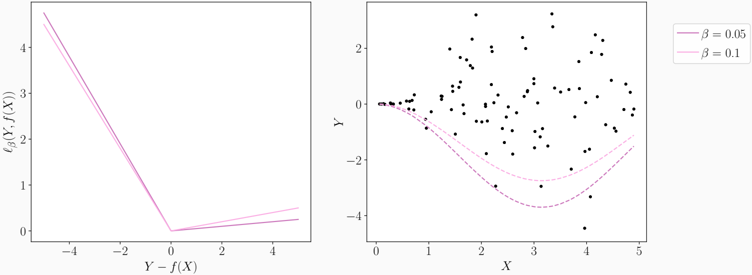
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \underset{f \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



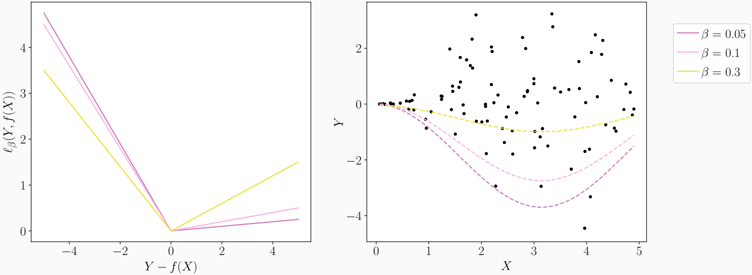
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



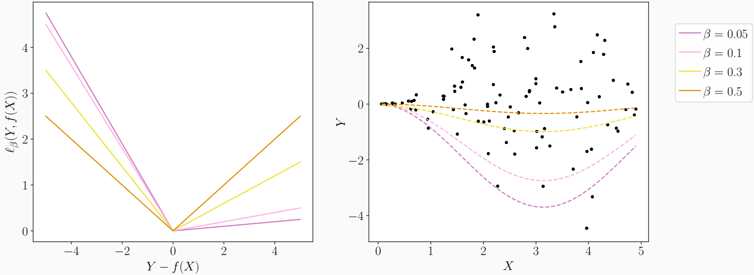
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\text{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \underset{f \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}} \text{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



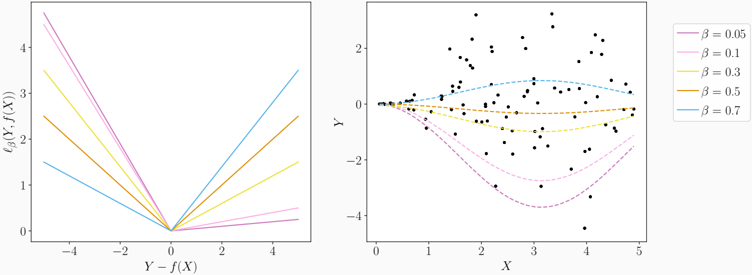
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



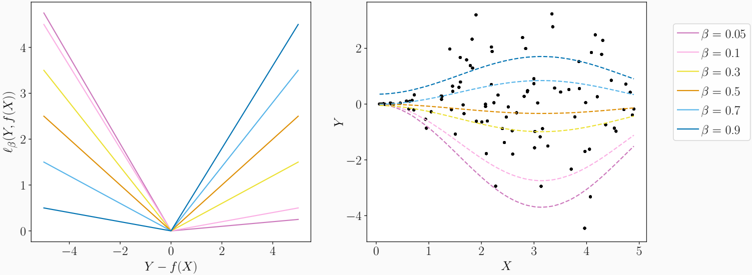
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



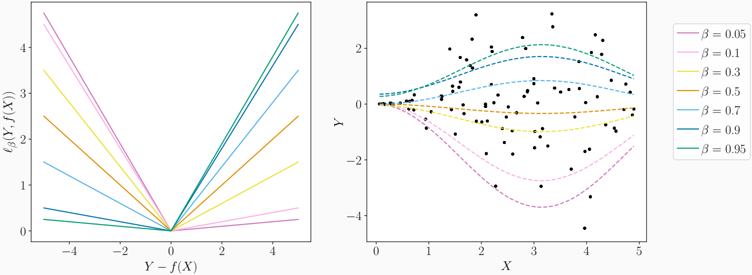
Quantile regression

- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\mathrm{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \operatorname*{argmin}_{f \in \mathbb{R}^{\mathcal{X}}} \mathrm{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$



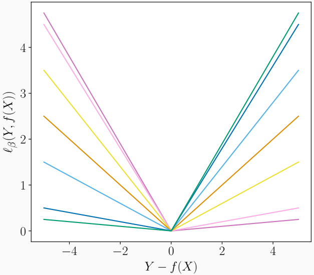
Quantile regression

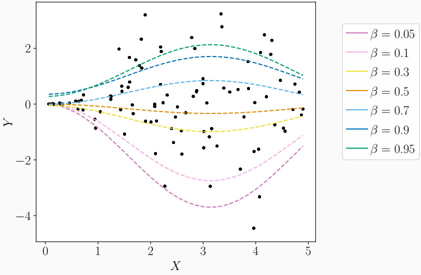
- Goal : approximate $Q_{Y|X}(\beta)$ – Quantile level β – Pinball loss $\ell_\beta(Y, Y')$.
- Associated risk:

$$\text{Risk}_{\ell_\beta}(f) = \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$$

- Bayes predictor:

$$f^\star \in \underset{f \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}} \text{Risk}_{\ell_\beta}(f) \quad \Rightarrow \quad f^\star(X) = Q_{Y|X}(\beta)$$





Warning - No theoretical guarantee with a finite sample!

$$\mathbb{P}\left(Y \in \left[\hat{Q}_{Y|X}(\beta/2); \hat{Q}_{Y|X}(1-\beta/2)\right]\right) \neq 1 - \beta$$

Quantifying predictive uncertainty

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $(X_i, Y_i)_{i=1}^n$
- Goal: predict an unseen point Y_{n+1} at X_{n+1} with **confidence**
- How? Given a miscoverage level $\alpha \in [0, 1]$, build a predictive set $\mathcal{C}_\alpha$ such that:

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\alpha\left(X_{n+1}\right)\right\} \geq 1 - \alpha, \tag{1}$$

  and $\mathcal{C}_\alpha$ should be as small as possible, in order to be informative

  For example: $\alpha = 0.1$ and obtain a 90% coverage interval

Quantifying predictive uncertainty

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $(X_i, Y_i)_{i=1}^n$
- Goal: predict an unseen point Y_{n+1} at X_{n+1} with **confidence**
- How? Given a miscoverage level $\alpha \in [0, 1]$, build a predictive set $\mathcal{C}_\alpha$ such that:
$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\alpha\left(X_{n+1}\right)\right\} \geq 1 - \alpha, \tag{1}$$
and $\mathcal{C}_\alpha$ should be as small as possible, in order to be informative
For example: $\alpha = 0.1$ and obtain a 90% coverage interval
- Construction of the predictive intervals should be
    - agnostic to the model
    - agnostic to the data distribution
- *Validity* should be ensured
    - in finite samples
    - for all data distribution and underlying model

Quantile Regression

Split Conformal Prediction (SCP)

- Standard regression case
- Conformalized Quantile Regression (CQR)
- Generalization of SCP: going beyond regression

On the design choices of conformity scores and (empirical) conditional guarantees

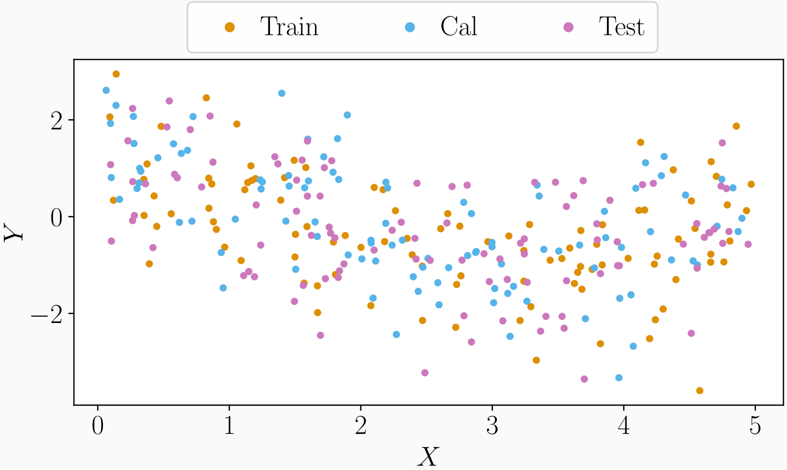
Avoiding data splitting: full conformal and out-of-bags approaches

Beyond exchangeability

Some case studies

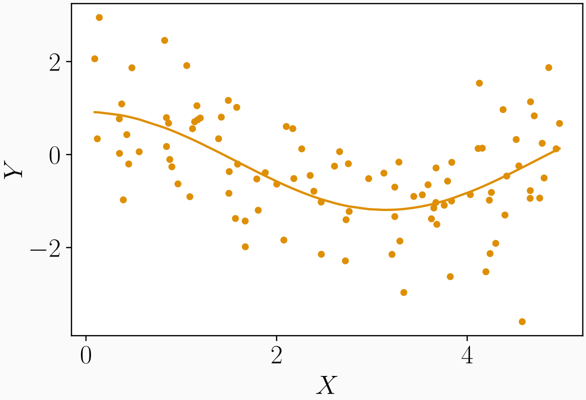
Concluding remarks

Split Conformal Prediction (SCP)[1,2,3]: toy example



[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

Split Conformal Prediction (SCP)[1,2,3]: toy example

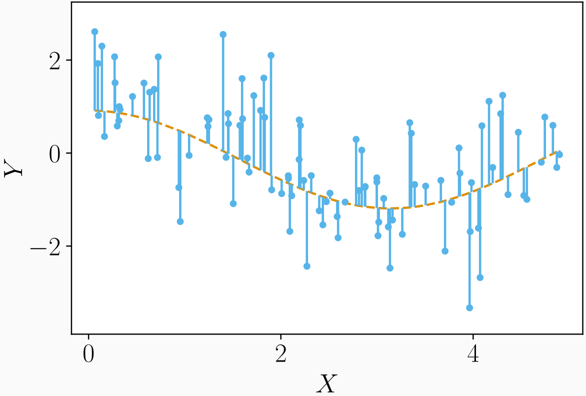


- Learn (or get) $\hat{\mu}$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*

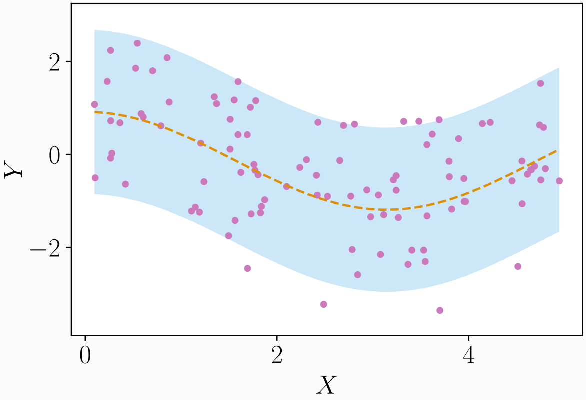
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML

[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B



- Predict with $\hat{\mu}$
- Get the $|\text{residuals}|$, a.k.a. conformity scores
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S} = \{|\text{residuals}|\}_{\text{Cal}} \cup \{+\infty\}$, noted $q_{1-\alpha}(\mathcal{S})$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B



- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)

SCP: implementation details

V0 with $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm $\mathcal{A}$ on the proper training set*

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$

SCP: implementation details

V0 with $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal}+1$ conformity scores:

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

($+$ worst-case scenario)

SCP: implementation details — V0 with $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal} + 1$ conformity scores:

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

($+$ worst-case scenario)

5. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$

SCP: implementation details — V0 with $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal} + 1$ conformity scores:

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

(+ worst-case scenario)

5. Compute the $1 - \alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
6. For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = [\hat{\mu}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \hat{\mu}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$$

SCP: implementation details — V1 without $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal}$ conformity scores:

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\}$$

SCP: implementation details — V1 without $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal}$ conformity scores :

$$\mathcal{S} = \{ S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal} \}$$

5. Compute the $(1-\alpha)\left(\frac{1}{\#\mathrm{Cal}} + 1\right)$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$

SCP: implementation details — V1 without $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal}$ conformity scores:
$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\}$$
5. Compute the $(1-\alpha)\left(\frac{1}{\#\mathrm{Cal}} + 1\right)$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
6. For a new point X_{n+1}, return
$$\widehat{C}_\alpha(X_{n+1}) = [\hat{\mu}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \hat{\mu}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$$

SCP: theoretical foundation

SCP: theoretical foundation

Definition (Exchangeability)**.**

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \ldots, (X_n, Y_n)) \overset{d}{=} \left(\left(X_{\sigma(1)}, Y_{\sigma(1)}\right), \ldots, \left(X_{\sigma(n)}, Y_{\sigma(n)}\right)\right) .$$

SCP: theoretical foundation

Definition (Exchangeability).

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \ldots, (X_n, Y_n)) \overset{d}{=} \left((X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n)}, Y_{\sigma(n)})\right).$$

Toy case: Z_1 and Z_2 are exchangeable if $(Z_1, Z_2) \overset{d}{=} (Z_2, Z_1)$.

Definition (Exchangeability).

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \ldots, (X_n, Y_n)) \overset{d}{=} \left(\left(X_{\sigma(1)}, Y_{\sigma(1)}\right), \ldots, \left(X_{\sigma(n)}, Y_{\sigma(n)}\right)\right).$$

Example (exchangeable sequences).

- i.i.d. samples

Definition (Exchangeability).

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \ldots, (X_n, Y_n)) \stackrel{d}{=} \left(\left(X_{\sigma(1)}, Y_{\sigma(1)}\right), \ldots, \left(X_{\sigma(n)}, Y_{\sigma(n)}\right)\right).$$

Example (exchangeable sequences).

- i.i.d. samples
- The components of $\mathcal{N}\left(\begin{pmatrix} m \\ \vdots \\ \vdots \\ m \end{pmatrix}, \begin{pmatrix} \sigma^2 & & & \\ & \ddots & \gamma^2 & \\ & \gamma^2 & \ddots & \\ & & & \sigma^2 \end{pmatrix}\right)$

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Theorem (Marginal validity).

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Theorem (Marginal validity).

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

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

[a] Only the calibration and test data need to be exchangeable.

Lemma (Quantile lemma)**.**

If $(U_1, \ldots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0,1[$:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty)\right) \geq \beta.$$

Additionally, if $U_1, \ldots, U_n, U_{n+1}$ are almost surely distinct, then:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty)\right) \leq \beta + \frac{1}{n+1}.$$

Lemma (Quantile lemma).

If $(U_1, \ldots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0, 1[$:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty)\right) \geq \beta.$$

Additionally, if $U_1, \ldots, U_n, U_{n+1}$ are almost surely distinct, then:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty)\right) \leq \beta + \frac{1}{n+1}.$$

When $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are exchangeable.
$\hookrightarrow$ applying the quantile lemma to the scores concludes the proof.

$$\begin{aligned}
\left\{Y_{n+1} \in \widehat{C}_{n,\alpha}\left(X_{n+1}\right)\right\} &= \left\{\hat{\mu}\left(X_{n+1}\right) - q_{1-\alpha}\left(\mathcal{S}\right) \leq Y_{n+1} \leq \hat{\mu}\left(X_{n+1}\right) + q_{1-\alpha}\left(\mathcal{S}\right)\right\} \\
&= \left\{\left|Y_{n+1} - \hat{\mu}\left(X_{n+1}\right)\right| \leq q_{1-\alpha}\left(\mathcal{S}\right)\right\} \\
&= \left\{S_{n+1} \leq q_{1-\alpha}\left(\mathcal{S}\right)\right\}.
\end{aligned}$$

First note that $U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})$.

First note that $U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) \overset{d}{=} \mathbb{P}\left(U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right)$. Thus:

$$
\begin{aligned}
\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right\}\right] \\
&\geq \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil\right] \\
&= \frac{\lceil \beta(n+1) \rceil}{n+1} \\
&\geq \beta,
\end{aligned}
$$

proving the first statement.

First note that $U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) \overset{d}{=} \mathbb{P}\left(U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$. Thus:

$$
\begin{aligned}
\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right\}\right] \\
&= \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil\right] \quad \text{if all } (U_i) \text{ are distinct} \\
&= \frac{\lceil \beta(n+1) \rceil}{n+1} \\
&\leq \beta + \frac{1}{n+1},
\end{aligned}
$$

proving the second statement.

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Theorem (Marginal validity Vovk et al. (2005)).

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[d]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

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

[d] Only the calibration and test data need to be exchangeable.

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Theorem (Marginal validity Vovk et al. (2005)).

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[d]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

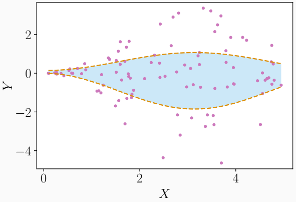
Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

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

[d] Only the calibration and test data need to be exchangeable.

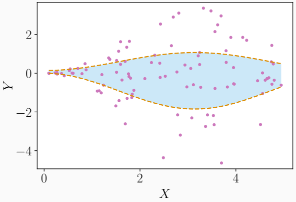
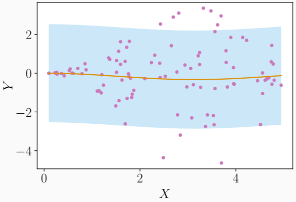
✗ Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha$

Conditional coverage implies adaptiveness

no coverage		
		

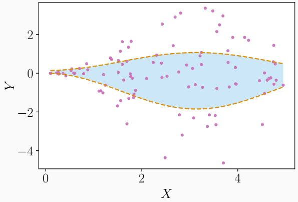
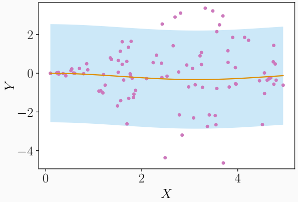
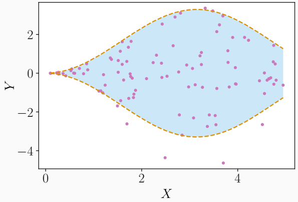
Conditional coverage implies adaptiveness

- Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right)\right\}$ the errors may differ across regions of the input space (i.e. non-adaptive)

no coverage	marginal	
		

Conditional coverage implies adaptiveness

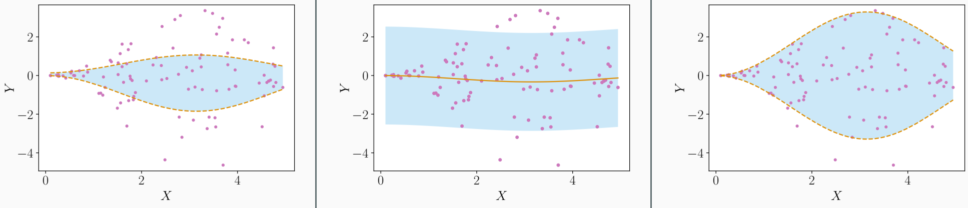
- Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right)\right\}$ the errors may differ across regions of the input space (i.e. non-adaptive)
- Conditional coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \mid X_{n+1}\right\}$ errors are evenly distributed (i.e. fully adaptive)

no coverage	marginal	conditional
		

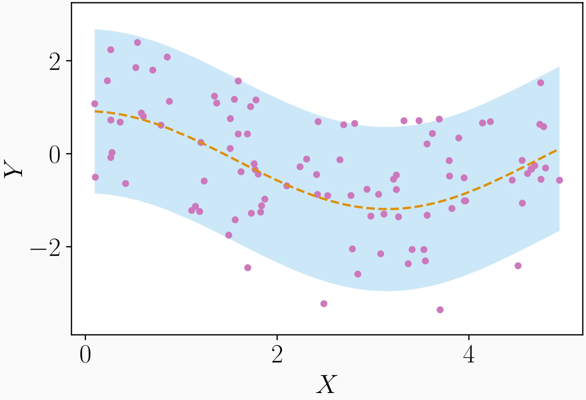
Conditional coverage implies adaptiveness

- Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\}$ the errors may differ across regions of the input space (i.e. non-adaptive)
- Conditional coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,|X_{n+1}\right\}$ errors are evenly distributed (i.e. fully adaptive)
- Conditional coverage is stronger than marginal coverage

no coverage	marginal	conditional

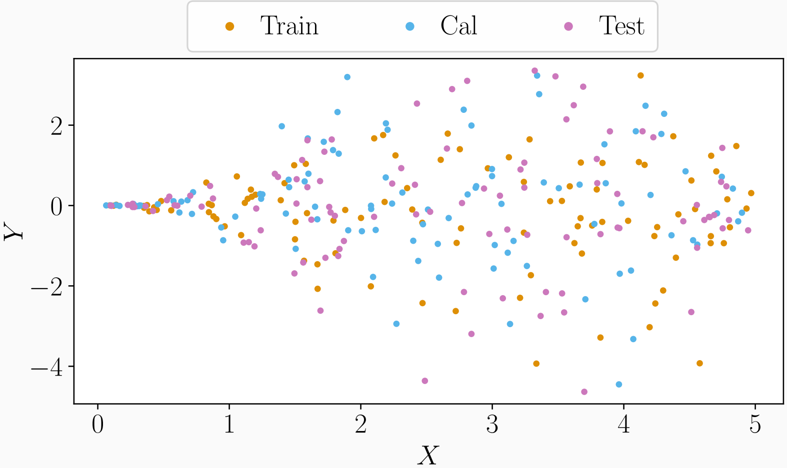


Standard mean-regression SCP is not adaptive

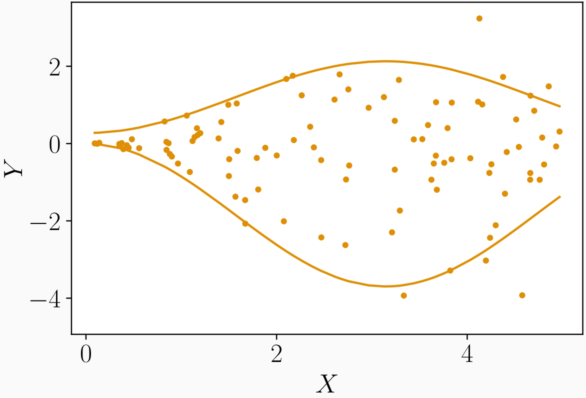


- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Conformalized Quantile Regression (CQR)[5]

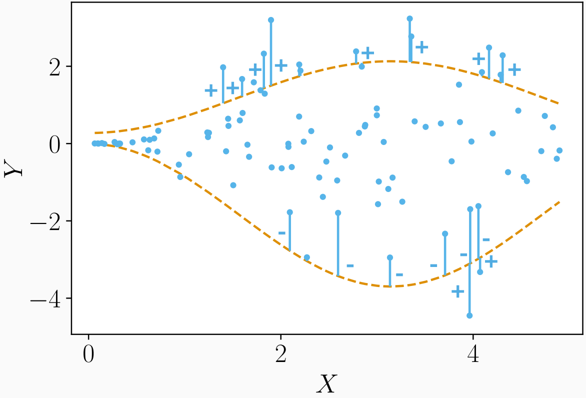


[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



- Learn (or get) $\widehat{\text{QR}}_{\text{lower}}$ and $\widehat{\text{QR}}_{\text{upper}}$

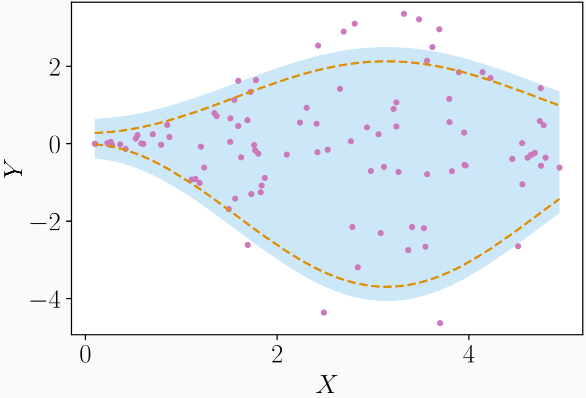
[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



- Predict with $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$
- Get the scores $\mathcal{S} = \{S_i\}_{\mathrm{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}(\mathcal{S})$

$$\hookrightarrow \quad S_i := \max\left\{\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right\}$$

[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



- Predict with $\widehat{\text{QR}}_{\text{lower}}$ and $\widehat{\text{QR}}_{\text{upper}}$

- Build

$$\widehat{C}_{\alpha}(x) = [\widehat{\text{QR}}_{\text{lower}}(x) - q_{1-\alpha}(\mathcal{S}); \widehat{\text{QR}}_{\text{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$$

[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$ *by training the algorithm $\mathcal{A}$ on the proper training set*

CQR: implementation details

V0 with $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. Obtain a set of $\#\mathrm{Cal}+1$ conformity scores $\mathcal{S}$:

$$\mathcal{S} = \{S_i = \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

CQR: implementation details V0 with $+\infty$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. Obtain a set of $\#\mathrm{Cal}+1$ conformity scores $\mathcal{S}$:

   $$\mathcal{S} = \{S_i = \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

4. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. Obtain a set of $\#\mathrm{Cal}+1$ conformity scores $\mathcal{S}$:

   $$\mathcal{S} = \{S_i = \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right), i \in \mathrm{Cal}\} \cup \{+\infty\}$$
4. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point X_{n+1}, return

   $$\widehat{C}_\alpha(X_{n+1}) = [\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$$

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\widehat{\mathsf{QR}}_{\mathsf{lower}}$ and $\widehat{\mathsf{QR}}_{\mathsf{upper}}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. Obtain a set of $\#\mathrm{Cal}$ conformity scores $\mathcal{S}$:
   $$\mathcal{S} = \{S_i = \max\left(\widehat{\mathsf{QR}}_{\mathsf{lower}}(X_i) - Y_i, Y_i - \widehat{\mathsf{QR}}_{\mathsf{upper}}(X_i)\right), i \in \mathrm{Cal}\}$$
4. Compute the $(1-\alpha)\left(\frac{1}{\#\mathrm{Cal}} + 1\right)$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point X_{n+1}, return
   $$\widehat{C}_\alpha(X_{n+1}) = [\widehat{\mathsf{QR}}_{\mathsf{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \widehat{\mathsf{QR}}_{\mathsf{upper}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$$

CQR: theoretical guarantees

This procedure enjoys the finite sample guarantee proposed and proved in Romano et al. (2019).

Theorem (Marginal validity of CQR Romano et al. (2019))**.**

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. CQR on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

If, in addition, the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are almost surely distinct, then

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

[a] Only the calibration and test data need to be exchangeable.

Proof: quantile lemma again $Y_{n+1} \in \widehat{C}_{n,\alpha}(X_{n+1}) \Leftrightarrow S_{n+1} \leq q_{1-\alpha}(\mathcal{S})$.

CQR: theoretical guarantees

This procedure enjoys the finite sample guarantee proposed and proved in Romano et al. (2019).

Theorem (Marginal validity of CQR Romano et al. (2019))**.**

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. CQR on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

If, in addition, the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are almost surely distinct, then

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

[a]Only the calibration and test data need to be exchangeable.

Proof: quantile lemma again $Y_{n+1} \in \widehat{C}_{n,\alpha}\left(X_{n+1}\right) \Leftrightarrow S_{n+1} \leq q_{1-\alpha}\left(\mathcal{S}\right)$.

✗ Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathsf{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S}=\{S_i=\mathrm{s}(\hat{A}(X_i),Y_i),i\in\mathrm{Cal}\}\cup\{+\infty\}$$

Ex 1: $\mathrm{s}(\hat{A}(X_i),Y_i):=|\hat{\mu}(X_i)-Y_i|$ in regression with standard scores

Ex 2: $\mathrm{s}(\hat{A}(X_i),Y_i):=\max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i)-Y_i,Y_i-\widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

   Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

   Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

4. Compute the $1 - \alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

   Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

   Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

4. Compute the $1 - \alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

   Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

   Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

4. Compute the $1 - \alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

   Ex 1: $\widehat{C}_\alpha(X_{n+1}) = [\hat{\mu}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S})]$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal}+1$ conformity scores

$$\mathcal{S} = \{S_i = \mathrm{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathrm{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

Ex 2: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ in CQR

4. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

Ex 2: $\widehat{C}_\alpha(X_{n+1}) = [\widehat{QR}_{\mathrm{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S});$

$\widehat{QR}_{\mathrm{upper}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$

SCP is defined by the conformity score function

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, obtain $\#\mathrm{Cal} + 1$ conformity scores

$$\mathcal{S} = \{S_i = \mathsf{s}(\hat{A}(X_i), Y_i), i \in \mathrm{Cal}\} \cup \{+\infty\}$$

Ex 1: $\mathsf{s}(\hat{A}(X_i), Y_i) := |\hat{\mu}(X_i) - Y_i|$ in regression with standard scores

Ex 2: $\mathsf{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathsf{QR}}_{\mathsf{lower}}(X_i) - Y_i, Y_i - \widehat{\mathsf{QR}}_{\mathsf{upper}}(X_i)\right)$ in CQR

4. Compute the $1 - \alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$
5. For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathsf{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

$\hookrightarrow$ The definition of the conformity scores is crucial, as they incorporate almost all the information: data + underlying model

SCP: theoretical guarantees

This procedure enjoys the finite sample guarantee proposed and proved in Vovk et al. (2005).

Theorem (Marginal validity of SCP Vovk et al. (2005)).

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. SCP on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

If, in addition, the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are almost surely distinct, then

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

[a] Only the calibration and test data need to be exchangeable.

Proof: application of the quantile lemma.

SCP: theoretical guarantees

This procedure enjoys the finite sample guarantee proposed and proved in Vovk et al. (2005).

Theorem (Marginal validity of SCP Vovk et al. (2005)).

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. SCP on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

If, in addition, the scores $\{S_i\}_{i\in\mathrm{Cal}} \cup \{S_{n+1}\}$ are almost surely distinct, then

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

[a] Only the calibration and test data need to be exchangeable.

Proof: application of the quantile lemma.

✗ Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha$

SCP: standard classification

- $Y \in \{1, \ldots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \ldots, \hat{p}_C(X))$ (estimated probabilities)

SCP: standard classification

- $Y \in \{1, \ldots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \ldots, \hat{p}_C(X))$ (estimated probabilities)
- $\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

SCP: standard classification

- $Y \in \{1, \dots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \dots, \hat{p}_C(X))$ (estimated probabilities)
- $\boxed{\mathrm{s}}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$
- For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \boxed{\mathrm{s}}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, 0.35)$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$
- $\hat{A}(X_{n+1}) = (\mathbf{0.05}, 0.60, 0.35)$

  $\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \textbf{"dog"}) = 0.95$ "dog" $\notin \widehat{C}_\alpha(X_{n+1})$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$
- $\hat{A}(X_{n+1}) = (0.05, \mathbf{0.60}, 0.35)$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "dog"$) = 0.95$ "dog" $\notin \widehat{C}_\alpha(X_{n+1})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ **"tiger"**$) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$ "tiger" $\in \widehat{C}_\alpha(X_{n+1})$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, \mathbf{0.35})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \text{"dog"}) = 0.95$ "dog" $\notin \widehat{C}_\alpha(X_{n+1})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$ "tiger" $\in \widehat{C}_\alpha(X_{n+1})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \textbf{"cat"}) = 0.65 \leq q_{1-\alpha}(\mathcal{S})$ "cat" $\in \widehat{C}_\alpha(X_{n+1})$

SCP: standard classification in practice

Ex: $Y_i \in \{\text{“dog”}, \text{“tiger”}, \text{“cat”}\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	“dog”	“dog”	“dog”	“tiger”	“tiger”	“tiger”	“tiger”	“cat”	“cat”	“cat”
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.50	0.45	0.40	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.30	0.40	0.45	0.40	0.35
S_i	0.05	0.1	0.15	0.40	0.45	0.50	0.55	0.55	0.6	0.65

- $q_{1-\alpha}(\mathcal{S}) = 0.65$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, 0.35)$
  - $\hookrightarrow$ $\mathrm{s}\left(\hat{A}(X_{n+1}), \text{“dog”}\right) = 0.95$ — “dog” $\notin \widehat{C}_\alpha(X_{n+1})$
  - $\hookrightarrow$ $\mathrm{s}\left(\hat{A}(X_{n+1}), \text{“tiger”}\right) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$ — “tiger” $\in \widehat{C}_\alpha(X_{n+1})$
  - $\hookrightarrow$ $\mathrm{s}\left(\hat{A}(X_{n+1}), \text{“cat”}\right) = 0.65 \leq q_{1-\alpha}(\mathcal{S})$ — “cat” $\in \widehat{C}_\alpha(X_{n+1})$
- $\widehat{C}_\alpha(X_{n+1}) = \{\text{“tiger”}, \text{“cat”}\}$

SCP: standard classification in practice, cont'd

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
S_i	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

- $q_{1-\alpha}(\mathcal{S}) = 0.45$

SCP: standard classification in practice, cont'd

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
S_i	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

- $q_{1-\alpha}(\mathcal{S}) = 0.45$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, 0.35)$

$\hookrightarrow$ $\boxed{\mathrm{s}}(\hat{A}(X_{n+1}), \text{“dog”}) = 0.95$

$\hookrightarrow$ $\boxed{\mathrm{s}}(\hat{A}(X_{n+1}), \text{“tiger”}) = 0.40$

$\hookrightarrow$ $\boxed{\mathrm{s}}(\hat{A}(X_{n+1}), \text{“cat”}) = 0.65$

SCP: standard classification in practice, cont'd

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
S_i	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

- $q_{1-\alpha}(\mathcal{S}) = 0.45$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, 0.35)$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "dog"$) = 0.95$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "tiger"$) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "cat"$) = 0.65$

SCP: standard classification in practice, cont'd

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
S_i	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

- $q_{1-\alpha}(\mathcal{S}) = 0.45$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, 0.35)$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \text{"dog"}) = 0.95$ "dog" $\notin \widehat{C}_\alpha(X_{n+1})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$ "tiger" $\in \widehat{C}_\alpha(X_{n+1})$

$\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}), \text{"cat"}) = 0.65$ "cat" $\notin \widehat{C}_\alpha(X_{n+1})$

SCP: standard classification in practice, cont'd

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.05	0.10	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.70	0.25	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.25	0.65	0.60	0.55
S_i	0.05	0.1	0.15	0.15	0.20	0.25	0.30	0.35	0.40	0.45

- $q_{1-\alpha}(\mathcal{S}) = 0.45$
- $\hat{A}(X_{n+1}) = (0.05, 0.60, 0.35)$
  - $\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "dog"$) = 0.95$ — "dog" $\notin \widehat{C}_\alpha(X_{n+1})$
  - $\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "tiger"$) = 0.40 \leq q_{1-\alpha}(\mathcal{S})$ — "tiger" $\in \widehat{C}_\alpha(X_{n+1})$
  - $\hookrightarrow$ $\mathrm{s}(\hat{A}(X_{n+1}),$ "cat"$) = 0.65$ — "cat" $\notin \widehat{C}_\alpha(X_{n+1})$
- $\widehat{C}_\alpha(X_{n+1}) = \{$"tiger"$\}$

SCP: limits of the standard classification case

efficiency yet non-adaptivity of the simplest classification scores

✓ Outputs the most efficient set possible (i.e. achieving the smallest average set size, Sadinle et al., 2018),

✗ Does not allow to discriminate between "easy" and "hard" test point. In practice, it leads to predictive sets that under-cover (resp. over-cover) on "hard" (resp. "easy") subgroups.
This is due to the fact that the same threshold $q_{1-\alpha}(\mathcal{S})$ is applied to any test point.

SCP: classification with Adaptive Prediction Sets[8]

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$

[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS

SCP: classification with Adaptive Prediction Sets[8]

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$
2. $s(x, y; \hat{p}) := \sum_{k=1}^{\sigma_x^{-1}(y)} \hat{p}_{\sigma_x(k)}(x)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y)

[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS

SCP: classification with Adaptive Prediction Sets[8]

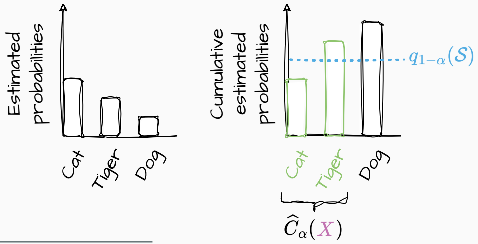
1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$
2. $\mathsf{s}(x, y; \hat{p}) := \sum_{k=1}^{\sigma_x^{-1}(y)} \hat{p}_{\sigma_x(k)}(x)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y)
3. Return the set of classes $\left\{\sigma_{X_{n+1}}(1), \ldots, \sigma_{X_{n+1}}(r^\star)\right\}$, where

$$r^\star = \operatorname*{arg\,max}_{1 \leq r \leq C} \left\{ \sum_{k=1}^{r} \hat{p}_{\sigma_{X_{n+1}}(k)}(X_{n+1}) < q_{1-\alpha}(\mathcal{S}) \right\} + 1$$

[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$
2. $s(x, y; \hat{p}) := \sum_{k=1}^{\sigma_x^{-1}(y)} \hat{p}_{\sigma_x(k)}(x)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y)
3. Return the set of classes $\{\sigma_{X_{n+1}}(1), \ldots, \sigma_{X_{n+1}}(r^\star)\}$, where

$$r^\star = \underset{1 \leq r \leq C}{\arg\max} \left\{ \sum_{k=1}^{r} \hat{p}_{\sigma_{X_{n+1}}(k)}(X_{n+1}) < q_{1-\alpha}(\mathcal{S}) \right\} + 1$$



[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS
Figure highly inspired by Angelopoulos and Bates (2023).

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$“dog”, “tiger”, “cat”$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	“dog”	“dog”	“dog”	“tiger”	“tiger”	“tiger”	“tiger”	“cat”	“cat”	“cat”
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$, $r^\star = 2$

$$\widehat{C}_\alpha(X_{n+1}) = \{\text{"tiger"}, \text{"cat"}\}$$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$, $r^\star = 2$

  $\widehat{C}_\alpha(X_{n+1}) = \{$"tiger", "cat"$\}$

  $\hookrightarrow$ Ex 2: $\hat{A}(X_{n+1}) = (0.03, 0.95, 0.02)$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

$\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$, $r^\star = 2$

$\widehat{C}_\alpha(X_{n+1}) = \{$"tiger", "cat"$\}$

$\hookrightarrow$ Ex 2: $\hat{A}(X_{n+1}) = (0.03, 0.95, 0.02)$, $r^\star = 1$

$\widehat{C}_\alpha(X_{n+1}) = \{$"tiger"$\}$

Split Conformal Prediction: summary

- **Simple** procedure which quantifies the uncertainty of **any** predictive model $\hat{A}$ by returning predictive regions
- **Finite-sample** guarantees
- **Distribution-free** as long as the data are exchangeable (and so are the scores)

Split Conformal Prediction: summary

- **Simple** procedure which quantifies the uncertainty of **any** predictive model $\hat{A}$ by returning predictive regions
- **Finite-sample** guarantees
- **Distribution-free** as long as the data are exchangeable (and so are the scores)
- **Marginal** theoretical guarantee over the joint (X, Y) distribution, and not conditional, i.e., no guarantee that for any x:

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \cancel{| X_{n+1} = x} \right\} \geq 1 - \alpha.$$

Split Conformal Prediction: summary

- **Simple** procedure which quantifies the uncertainty of **any** predictive model $\hat{A}$ by returning predictive regions
- **Finite-sample** guarantees
- **Distribution-free** as long as the data are exchangeable (and so are the scores)
- **Marginal** theoretical guarantee over the joint (X, Y) distribution, and not conditional, i.e., no guarantee that for any x:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha.$$

$\hookrightarrow$ marginal also over the whole calibration set and the test point!

Challenges: open questions (non exhaustive!)

- Conditional coverage
- Computational cost vs statistical power
- Exchangeability

Quantile Regression

Split Conformal Prediction (SCP)

On the design choices of conformity scores and (empirical) conditional guarantees

- On distribution-free X-conditional validity
- Y-conditional validity
- Impact of the calibration set on the coverage

Avoiding data splitting: full conformal and out-of-bags approaches

Beyond exchangeability

Some case studies

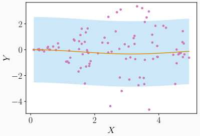
Concluding remarks

SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

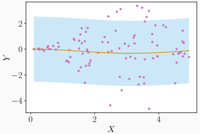
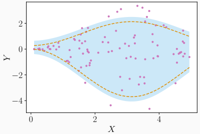
SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \text{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)		
$\text{s}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$		
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$		
Visu.			
✓	black-box around a "usable" prediction		
✗	not adaptive		

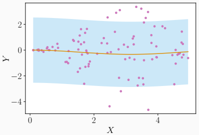
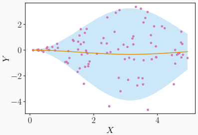
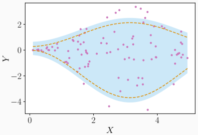
SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \boxed{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)		**CQR** Romano et al. (2019)
$\boxed{s}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$		$\max(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X) - Y,\ Y - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$		$[\widehat{\mathrm{QR}}_{\mathrm{lower}}(x) - q_{1-\alpha}(\mathcal{S});\ \widehat{\mathrm{QR}}_{\mathrm{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
Visu.			
✓	black-box around a "usable" prediction		adaptive
✗	not adaptive		no black-box around a "usable" prediction

SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)	**Locally weighted SCP** Lei et al. (2018)	**CQR** Romano et al. (2019)
$\mathrm{s}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$	$\dfrac{\lvert\hat{\mu}(X) - Y\rvert}{\hat{\rho}(X)}$	$\max(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X) - Y,$ $Y - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})\hat{\rho}(x)]$	$[\widehat{\mathrm{QR}}_{\mathrm{lower}}(x) - q_{1-\alpha}(\mathcal{S});$ $\widehat{\mathrm{QR}}_{\mathrm{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
Visu.			
✓	black-box around a "usable" prediction	black-box around a "usable" prediction	adaptive
✗	not adaptive	limited adaptiveness	no black-box around a "usable" prediction

Definition of distribution-free features conditional validity

$\widehat{C}_\alpha$ = estimated predictive set based on n data points.

Definition (Distribution-free X-conditional validity).

$\widehat{C}_\alpha$ achieves distribution-free X-conditional validity if:

- for any distribution $\mathcal{D}$,
- for any associated exchangeable joint distribution $\mathcal{D}^{\mathrm{exch}(n+1)}$,

we have that:

$$\mathbb{P}_{\mathcal{D}^{\mathrm{exch}(n+1)}}\left(Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}\right) | X^{(n+1)}\right) \overset{a.s.}{\geq} 1 - \alpha.$$

Informative conditional coverage as such is impossible

Theorem (Impossibility results Vovk (2012); Lei and Wasserman (2014)).

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

- *Regression:* $\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left(\text{mes} \left(\widehat{C}_\alpha \left(x \right) \right) = \infty \right) \geq 1 - \alpha$,
- *Classification:* for any $y \in \mathcal{Y}$, $\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left(y \in \widehat{C}_\alpha \left(x \right) \right) \geq 1 - \alpha$.

Informative conditional coverage as such is impossible

Theorem (Impossibility results Vovk (2012); Lei and Wasserman (2014)).

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

- *Regression:* $\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left(\text{mes} \left(\widehat{C}_\alpha \left(x \right) \right) = \infty \right) \geq 1 - \alpha$,
- *Classification:* for any $y \in \mathcal{Y}$, $\mathbb{P}_{\mathcal{D}^{\otimes(n)}} \left(y \in \widehat{C}_\alpha \left(x \right) \right) \geq 1 - \alpha$.

$\hookrightarrow$ distribution-free X-conditional hardness result apply beyond CP

Informative conditional coverage as such is impossible

Theorem (Impossibility results Vovk (2012); Lei and Wasserman (2014)).

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, **for any** $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

- *Regression:* $\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\text{mes}\left(\widehat{C}_\alpha(x)\right) = \infty\right) \geq 1 - \alpha$,
- *Classification:* for any $y \in \mathcal{Y}$, $\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(y \in \widehat{C}_\alpha(x)\right) \geq 1 - \alpha$.

$\hookrightarrow$ distribution-free X-conditional hardness result apply beyond CP

$\hookrightarrow$ X-conditional estimators are overly large even on easy cases

Informative conditional coverage as such is impossible

Theorem (Impossibility results Vovk (2012); Lei and Wasserman (2014)).

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

- *Regression:* $\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\operatorname{mes}\left(\widehat{C}_\alpha(x)\right) = \infty\right) \geq 1 - \alpha$,
- *Classification:* for any $y \in \mathcal{Y}$, $\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(y \in \widehat{C}_\alpha(x)\right) \geq 1 - \alpha$.

$\hookrightarrow$ distribution-free X-conditional hardness result apply beyond CP

$\hookrightarrow$ X-conditional estimators are overly large even on easy cases

$\hookrightarrow$ the lower bounds are tight

Example (Naive estimator).

$\mathcal{C}_\alpha(\cdot;\xi) \equiv \mathcal{Y}\mathbb{1}\{\xi \leq 1-\alpha\} + \emptyset\mathbb{1}\{\xi > \alpha\}$, where $\xi \sim \mathcal{U}([0,1])$.

Informative conditional coverage as such is impossible

Theorem (Impossibility results Vovk (2012); Lei and Wasserman (2014)).

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

- *Regression:* $\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\text{mes}\left(\widehat{C}_\alpha(x)\right) = \infty\right) \geq 1 - \alpha$,
- **Classification:** for any $y \in \mathcal{Y}$, $\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(y \in \widehat{C}_\alpha(x)\right) \geq 1 - \alpha$.

$\hookrightarrow$ distribution-free X-conditional hardness result apply beyond CP

$\hookrightarrow$ X-conditional estimators are overly large even on easy cases

$\hookrightarrow$ the lower bounds are tight

$\hookrightarrow$ <u>Classification</u>: *every label* is likely to be included in $\widehat{C}_\alpha$.
$\widehat{C}_\alpha$ is likely to be large: for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$,
$\mathbb{E}_{\mathcal{D}^{\otimes(n)}}\left[\#\widehat{C}_\alpha(x)\right] \geq (1-\alpha)\#\mathcal{Y}$.

Definition (distribution-free $(1-\alpha, \delta)$–X-conditional validity).

Let $\delta > 0$ be a tolerance level.
An estimator $\widehat{C}_\alpha$ achieves distribution-free $(1-\alpha, \delta)$–X-conditional validity if for any distribution $\mathcal{D}$, for any $\mathscr{X} \subseteq \mathcal{X}$ such that $\mathbb{P}_{\mathcal{D}_X}(X \in \mathscr{X}) \geq \delta$, and for any associated exchangeable joint distribution $\mathcal{D}^{\text{exch}(n+1)}$, we have:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left(Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \middle| X_{n+1} \in \mathscr{X}\right) \geq 1-\alpha.$$

Weaker notion of X-conditional validity (Barber et al., 2021a)

Definition (distribution-free $(1-\alpha,\delta)$–X-conditional validity).

Let $\delta > 0$ be a tolerance level.
An estimator $\widehat{C}_\alpha$ achieves distribution-free $(1-\alpha,\delta)$–X-conditional validity if for any distribution $\mathcal{D}$, for any $\mathcal{X} \subseteq \mathcal{X}$ such that $\mathbb{P}_{\mathcal{D}_X}(X \in \mathcal{X}) \geq \delta$, and for any associated exchangeable joint distribution $\mathcal{D}^{\text{exch}(n+1)}$, we have:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,|X_{n+1} \in \mathcal{X}\right) \geq 1-\alpha.$$

Informal theorem (lower bound on $(1-\alpha,\delta)$–X-cond. valid efficiency)

An estimator achieving $(1-\alpha,\delta)$–X-conditional validity can not be more efficient than an estimator achieving **distribution-free marginal validity at the level** $1-\alpha\delta$.

$\hookrightarrow$ In practive, consider small $\delta \rightarrow$ unefficient predictive sets.

Definition (distribution-free group-features-conditional validity)**.**

Let $G := \left(G^{(k)}\right)_{k=1}^{K}$ represents groups on the features space (possibly overlapping).

An estimator $\widehat{C}_\alpha$ achieves distribution-free G-conditional validity if for any distribution $\mathcal{D}$, and for any associated exchangeable joint distribution $\mathcal{D}^{\mathrm{exch}(n+1)}$, we have:

$$\mathbb{P}_{\mathcal{D}^{\mathrm{exch}(n+1)}}\left(Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}, G_{n+1}\right) \middle| G_{n+1}\right) \overset{a.s.}{\geq} 1 - \alpha.$$

Hardness results on G-conditional-validity (Zaffran et al., 2024)

Theorem (General MCV hardness result)**.**

If $\widehat{C}_\alpha$ is distribution-free group-features-conditionally valid then for any distribution $\mathcal{D}$, for any group g such that $\mathcal{D}_G(g) := \mathbb{P}_{\mathcal{D}}(G = g) > 0$, it holds:

- *Regression*

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\mathrm{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1},$$

- *Classification*

for any $y \in \mathcal{Y}$, $\mathbb{P}_{P^{\otimes(n+1)}}\left(\mathit{alert} < 2|\mathit{handout} : 0 > y \in \widehat{C}_\alpha\left(X_{n+1}, g\right)\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$

Theorem (General MCV hardness result)**.**

If $\widehat{C}_\alpha$ is distribution-free group-features-conditionally valid then for any distribution $\mathcal{D}$, for any group g such that $\mathcal{D}_G(g) := \mathbb{P}_{\mathcal{D}}(G = g) > 0$, it holds:

- *Regression*

  $\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}$,

- *Classification*

  for any $y \in \mathcal{Y}$, $\mathbb{P}_{P^{\otimes(n+1)}}\left(alert < 2|handout : 0 > y \in \widehat{C}_\alpha\left(X_{n+1}, g\right)\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}$.

Theorem (General MCV hardness result).

If $\widehat{C}_\alpha$ is distribution-free group-features-conditionally valid then for any distribution $\mathcal{D}$, for any group g such that $\mathcal{D}_G(g) := \mathbb{P}_{\mathcal{D}}(G = g) > 0$, it holds:

- *Regression*

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\mathrm{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1},$$

- *Classification*

$$\text{for any } y \in \mathcal{Y},\ \mathbb{P}_{P^{\otimes(n+1)}}\left(alert < 2|handout : 0 > y \in \widehat{C}_\alpha\left(X_{n+1}, g\right)\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$$

Hardness results on G-conditional-validity (Zaffran et al., 2024)

Theorem (General MCV hardness result)**.**

If $\widehat{C}_\alpha$ is distribution-free group-features-conditionally valid then for any distribution $\mathcal{D}$, for any group g such that $\mathcal{D}_G(g) := \mathbb{P}_\mathcal{D}(G = g) > 0$, it holds:

- *Regression*
$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1},$$
- *Classification*
$$\text{for any } y \in \mathcal{Y},\ \mathbb{P}_{P^{\otimes(n+1)}}\left(alert < 2|handout : 0 > y \in \widehat{C}_\alpha\left(X_{n+1}, g\right)\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$$

Irreducible term: consider $\widehat{C}_\alpha$ outputting $\mathcal{Y}$ with probability $1 - \alpha$ and $\emptyset$ otherwise.

Theorem (General MCV hardness result)**.**

If $\widehat{C}_\alpha$ is distribution-free group-features-conditionally valid then for any distribution $\mathcal{D}$, for any group g such that $\mathcal{D}_G(g) := \mathbb{P}_{\mathcal{D}}(G = g) > 0$, it holds:

- *Regression*

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1},$$

- *Classification*

for any $y \in \mathcal{Y}$, $\mathbb{P}_{P^{\otimes(n+1)}}\left(\mathit{alert} < 2|\mathit{handout} : 0 > y \in \widehat{C}_\alpha\left(X_{n+1}, g\right)\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}$.

Irreducible term: consider $\widehat{C}_\alpha$ outputting $\mathcal{Y}$ with probability $1-\alpha$ and $\emptyset$ otherwise.

$\Delta_{g,n}$ term: smaller than $\mathcal{D}_G(g)\sqrt{n+1}$

Hardness results on G-conditional-validity (Zaffran et al., 2024)

Theorem (General MCV hardness result).

If $\widehat{C}_\alpha$ is distribution-free group-features-conditionally valid then for any distribution $\mathcal{D}$, for any group g such that $\mathcal{D}_G(g) := \mathbb{P}_{\mathcal{D}}(G = g) > 0$, it holds:

- *Regression*

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1},$$

- *Classification*

$$\text{for any } y \in \mathcal{Y},\ \mathbb{P}_{P^{\otimes(n+1)}}\left(\mathit{alert} < 2|\mathit{handout} : 0 > y \in \widehat{C}_\alpha\left(X_{n+1}, g\right)\right) \geq 1 - \alpha - \Delta_{g,n} \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$$

Irreducible term: consider $\widehat{C}_\alpha$ outputting $\mathcal{Y}$ with probability $1-\alpha$ and $\emptyset$ otherwise.

$\Delta_{g,n}$ term: smaller than $\mathcal{D}_G(g)\sqrt{n+1}$

$\hookrightarrow$ gets negligible (making the lower bound nearly $1-\alpha$) **only** for low probability groups compared to n.

Restricting the link between G and (X or Y) does not allow informative G-conditional-coverage (Zaffran et al., 2024)

Analogous statements are also available for the classification framework.

Restricting the link between G and (X or Y) does not allow informative G-conditional-coverage (Zaffran et al., 2024)

Theorem ($G \perp\!\!\!\perp X$ hardness result).

If any $\widehat{C}_\alpha$ is G-conditionally-valid under $G \perp\!\!\!\perp X$, then for any distribution $\mathcal{D}$ such that $G \perp\!\!\!\perp X$, for any group g such that $\mathcal{D}_G(g) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$$

Analogous statements are also available for the classification framework.

Restricting the link between G and (X or Y) does not allow informative G-conditional-coverage (Zaffran et al., 2024)

Theorem ($G \perp\!\!\!\perp X$ hardness result)**.**

If any $\widehat{C}_\alpha$ is G-conditionally-valid under $G \perp\!\!\!\perp X$, then for any distribution $\mathcal{D}$ such that $G \perp\!\!\!\perp X$, for any group g such that $\mathcal{D}_G(g) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$$

Theorem ($Y \perp\!\!\!\perp G \mid X$ hardness result)**.**

If any $\widehat{C}_\alpha$ is G-conditionally-valid under $Y \perp\!\!\!\perp G \mid X$, then for any distribution $\mathcal{D}$ such that $Y \perp\!\!\!\perp G \mid X$, for any mask g such that $\frac{1}{\sqrt{2}} \geq \mathcal{D}_G(g) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - 2\mathcal{D}_G(g)\sqrt{n+1}.$$

Analogous statements are also available for the classification framework.

Restricting the link between G and (X or Y) does not allow informative G-conditional-coverage (Zaffran et al., 2024)

Theorem ($G \perp\!\!\!\perp X$ hardness result)**.**

If any $\widehat{C}_\alpha$ is G-conditionally-valid under $G \perp\!\!\!\perp X$, then for any distribution $\mathcal{D}$ such that $G \perp\!\!\!\perp X$, for any group g such that $\mathcal{D}_G(g) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - \mathcal{D}_G(g)\sqrt{n+1}.$$

Theorem ($Y \perp\!\!\!\perp G \mid X$ hardness result)**.**

If any $\widehat{C}_\alpha$ is G-conditionally-valid under $Y \perp\!\!\!\perp G \mid X$, then for any distribution $\mathcal{D}$ such that $Y \perp\!\!\!\perp G \mid X$, for any mask g such that $\frac{1}{\sqrt{2}} \geq \mathcal{D}_G(g) > 0$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, g\right)\right) = \infty\right) \geq 1 - \alpha - 2\mathcal{D}_G(g)\sqrt{n+1}.$$

$\Rightarrow$ Need to restrict **both** the link between G and X, **as well as** between G and Y.

Analogous statements are also available for the classification framework.

Getting closer to X-conditional coverage

Getting closer to X-conditional coverage

- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020a); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,|X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$

Non exhaustive references.

Getting closer to X-conditional coverage

- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020a); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) | X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$
- Asymptotic (with the sample size) conditional coverage
  $\hookrightarrow$ Romano et al. (2019); Kivaranovic et al. (2020); Chernozhukov et al. (2021); Sesia and Romano (2021); Izbicki et al. (2022)

Non exhaustive references.

- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020a); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) | X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$
- Asymptotic (with the sample size) conditional coverage
  $\hookrightarrow$ **Romano et al. (2019)**; Kivaranovic et al. (2020); Chernozhukov et al. (2021); Sesia and Romano (2021); Izbicki et al. (2022)

Non exhaustive references.

Achieving Y-conditional validity in classification

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ (*by training $\mathcal{A}$ on the proper training set* $(X_i, Y_i)_{i \in \mathrm{Tr}}$)
3. **For** any candidate $y \in \mathcal{Y}$:
   On the calibration set, obtain a set of $\#\mathrm{Cal}_y + 1$ conformity scores:
   $$\mathcal{S}_y = \{S_i = \mathrm{s}\left(X_i, y; \hat{A}\right), i \in \mathrm{Cal} \text{ such that } Y_i = y\} \cup \{+\infty\}$$
4. For a new point X_{n+1}, return $\widehat{C}_{n,\alpha}(X_{n+1}) \left\{y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S}_y)\right\}$

Achieving Y-conditional validity in classification

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{A}$ (*by training $\mathcal{A}$ on the proper training set* $(X_i, Y_i)_{i\in\mathrm{Tr}}$)
3. **For** any candidate $y \in \mathcal{Y}$:
   On the calibration set, obtain a set of $\#\mathrm{Cal}_y + 1$ conformity scores:
   $$\mathcal{S}_y = \{S_i = \mathrm{s}\left(X_i, y; \hat{A}\right), i \in \mathrm{Cal} \text{ such that } Y_i = y\} \cup \{+\infty\}$$
4. For a new point X_{n+1}, return $\widehat{C}_{n,\alpha}(X_{n+1}) \left\{y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S}_y)\right\}$

$\hookrightarrow$ What if there is a high class imbalance?

Ding et al. (2023) proposed to instead obtain **cluster**-conditional coverage.

Probably Approximately Correct bounds on calibration-conditional coverage (Vovk, 2012; Bian and Barber, 2023)

Theorem (calibration conditional validity of SCP).

SCP outputs $\widehat{C}_\alpha$ such that for any distribution $\mathcal{D}$ and any $0 < \delta \leq 0.5$:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\mathbb{P}_{\mathcal{D}}\left(Y_{n+1} \notin \widehat{C}_{n,\alpha}\left(X_{n+1}\right) \middle| \left(X_i, Y_i\right)_{i=1}^n\right) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2\#\mathrm{Cal}}}\right) \geq 1-\delta.$$

$\hookrightarrow$ controls the deviation of miscoverage with respect to the nominal level of a predictive set built on a given calibration set.

Splitting the data might not be desired

SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability

Splitting the data might not be desired

SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability

Can we avoid splitting the data set?

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - compute the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \left\{ \mathsf{s}\left(\hat{A}(X_i), Y_i\right)\right\}_{i=1}^{n} \cup \{\infty\}.$$

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - compute the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \left\{ \text{s} \left(\hat{A}\left(X_i\right), Y_i \right) \right\}_{i=1}^{n} \cup \{\infty\}.$$

  - output the set $\left\{ y \text{ such that } \text{s} \left(\hat{A}\left(X_{n+1}\right), y \right) \leq q_{1-\alpha}(\mathcal{S}) \right\}$.

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - compute the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \left\{ \mathrm{s}\left(\hat{A}(X_i), Y_i\right)\right\}_{i=1}^{n} \cup \{\infty\}.$$

  - output the set $\left\{y \text{ such that } \mathrm{s}\left(\hat{A}(X_{n+1}), y\right) \leq q_{1-\alpha}(\mathcal{S})\right\}$.

✗ $\hat{A}$ obtained w. the training set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ but not X_{n+1}.

Example ("Naive Idea" sets with an interpolating algorithm)**.**

Assume $\mathcal{A}$ interpolates:

- $\hat{A} = \mathcal{A}((x_1, y_1), \ldots, (x_n, y_n))$
- $\hat{A}(x_k) - y_k = 0$ for any $k \in [\![1, n]\!]$

$\Rightarrow$ Naive method above *(with MAE score functions)* outputs $\{\hat{A}(X_{n+1})\}$ (a single point) for any new test point!

Full Conformal Prediction[9] does not discard training points!

- Full *(or transductive)* Conformal Prediction
  - avoids data splitting

[9] Vovk et al. (2005), *Algorithmic Learning in a Random World*

Full Conformal Prediction[9] does not discard training points!

- Full *(or transductive)* Conformal Prediction
  - avoids data splitting
  - at the cost of many more model fits

[9] Vovk et al. (2005), *Algorithmic Learning in a Random World*

Full Conformal Prediction[9] does not discard training points!

- Full *(or transductive)* Conformal Prediction
  - avoids data splitting
  - at the cost of many more model fits
- **Idea**: the most probable labels Y_{n+1} live in $\mathcal{Y}$, and have a low enough conformity score. By looping over all possible $y \in \mathcal{Y}$, the ones leading to the smallest conformity scores will be found.

[9] Vovk et al. (2005), *Algorithmic Learning in a Random World*

Full Conformal Prediction (CP): recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$

Full Conformal Prediction (CP): recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores
$$\mathcal{S}_y^{(\text{train})} = \left\{ \mathrm{s}(\hat{A}_y(X_i), Y_i) \right\}_{i=1}^{n} \cup \{ \mathrm{s}(\hat{A}_y(X_{n+1}), y)\}$$
and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Full Conformal Prediction (CP): recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores
$$\mathcal{S}_y^{(\text{train})} = \left\{ \text{s}(\hat{A}_y(X_i), Y_i) \right\}_{i=1}^{n} \cup \{ \text{s}(\hat{A}_y(X_{n+1}), y)\}$$
and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Output the set $\left\{ y \text{ such that } \text{s}\left(\hat{A}_y\left(X_{n+1}\right), y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)\right\}$.

Full Conformal Prediction (CP): recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores
$$\mathcal{S}_y^{(\text{train})} = \left\{ \boxed{s} (\hat{A}_y(X_i), Y_i) \right\}_{i=1}^{n} \cup \{ \boxed{s} (\hat{A}_y(X_{n+1}), y)\}$$
and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Output the set $\left\{ y \text{ such that } \boxed{s} \left(\hat{A}_y\left(X_{n+1}\right), y \right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right) \right\}$.

✓ Test point treated in the same way than train points

Full Conformal Prediction (CP): recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores
$$\mathcal{S}_y^{\text{(train)}} = \left\{ \boxed{\mathsf{s}}(\hat{A}_y(X_i), Y_i) \right\}_{i=1}^{n} \cup \{ \boxed{\mathsf{s}}(\hat{A}_y(X_{n+1}), y)\}$$
and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{\text{(train)}}\right)$

Output the set $\left\{ y \text{ such that } \boxed{\mathsf{s}}\left(\hat{A}_y\left(X_{n+1}\right), y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{\text{(train)}}\right) \right\}$.

- ✓ Test point treated in the same way than train points
- ✓ Any score works

Full Conformal Prediction (CP): recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores
$$\mathcal{S}_y^{(\text{train})} = \left\{ \boxed{\mathrm{s}}(\hat{A}_y(X_i), Y_i) \right\}_{i=1}^{n} \cup \{ \boxed{\mathrm{s}}(\hat{A}_y(X_{n+1}), y)\}$$
and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Output the set $\left\{ y \text{ such that } \boxed{\mathrm{s}}\left(\hat{A}_y\left(X_{n+1}\right), y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)\right\}$.

- ✓ Test point treated in the same way than train points
- ✓ Any score works
- ✗ Computationally costly

Definition (Symmetrical algorithm)**.**

A deterministic algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric if for any permutation σ of $[\![1, n]\!]$: $\mathcal{A}\,(U_1, \ldots, U_n) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \ldots, U_{\sigma(n)}\right)$.

Definition (Symmetrical algorithm).

A deterministic algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric if for any permutation σ of $[\![1, n]\!]$: $\mathcal{A}\left(U_1, \ldots, U_n\right) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \ldots, U_{\sigma(n)}\right)$.

Lemma (Exchangeable scores).

If the algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric, and $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, then $S_1, \ldots, S_{n+1}$ are exchangeable, with

$$S_i := \mathrm{s}(\hat{A}_{Y_{n+1}}(X_i), Y_i).$$

Full CP: theoretical foundation

Definition (Symmetrical algorithm).

A deterministic algorithm $\mathcal{A} : (U_1, \dots, U_n) \mapsto \hat{A}$ is symmetric if for any permutation σ of $[\![1, n]\!]$: $\mathcal{A}(U_1, \dots, U_n) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \dots, U_{\sigma(n)}\right)$.

Lemma (Exchangeable scores).

If the algorithm $\mathcal{A} : (U_1, \dots, U_n) \mapsto \hat{A}$ is symmetric, and $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, then $S_1, \dots, S_{n+1}$ are exchangeable, with

$$S_i := \mathrm{s}(\hat{A}_{Y_{n+1}}(X_i), Y_i).$$

Moreover

$$\begin{aligned} Y_{n+1} \in \widehat{C_\alpha^{\text{Full}}}(X_{n+1}) &:= \left\{ y \text{ such that } \mathrm{s}\left(\hat{A}_y\left(X_{n+1}\right), y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)\right\} \\ &\Leftrightarrow \mathrm{s}\left(\hat{A}_{Y_{n+1}}\left(X_{n+1}\right), Y_{n+1}\right) \leq q_{1-\alpha}\left(\mathcal{S}_{Y_{n+1}}^{(\text{train})}\right) \\ &\Leftrightarrow S_{n+1} \leq q_{1-\alpha}(S_1, \dots, S_n, S_{n+1}) \ ! \end{aligned}$$

Full CP enjoys finite sample guarantees proved in Vovk et al. (2005).

Theorem (Marginal validity of Full CP Vovk et al. (2005))**.**

Suppose that

(i) $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable,

(ii) the algorithm $\mathcal{A}$ is symmetric.

Full CP applied on $(X_i, Y_i)_{i=1}^{n} \cup \{X_{n+1}\}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores are a.s. distinct:

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

Full CP: theoretical guarantees

Full CP enjoys finite sample guarantees proved in Vovk et al. (2005).

Theorem (Marginal validity of Full CP Vovk et al. (2005)).

Suppose that

(i) $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable,

(ii) the algorithm $\mathcal{A}$ is symmetric.

Full CP applied on $(X_i, Y_i)_{i=1}^{n} \cup \{X_{n+1}\}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores are a.s. distinct:

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

✗ Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha$

Interpolation regime

Example (FCP sets with an interpolating algorithm)**.**

Assume $\mathcal{A}$ interpolates:

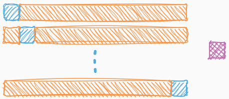
- $\hat{A} = \mathcal{A}\left((x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\right)$
- $\hat{A}(x_k) - y_k = 0$ for any $k \in [\![1, n+1]\!]$

Example (FCP sets with an interpolating algorithm).

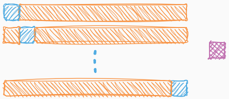
Assume $\mathcal{A}$ interpolates:

- $\hat{A} = \mathcal{A}\left((x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\right)$
- $\hat{A}(x_k) - y_k = 0$ for any $k \in [\![1, n+1]\!]$

$\Rightarrow$ Full Conformal Prediction *(with standard score functions)* outputs $\mathcal{Y}$ (the whole label space) for any new test point!

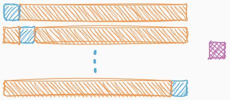


- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$



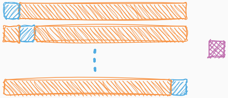
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)

Jackknife: the naive idea does not enjoy valid coverage

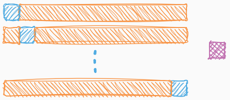


- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$

Jackknife: the naive idea does not enjoy valid coverage



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ \left| \hat{A}_{-i}(X_i) - Y_i \right| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$
- Build the predictive interval: $\left[\hat{A}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S}) \right]$

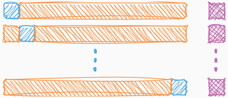


- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$
- Build the predictive interval: $\left[\hat{A}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S}) \right]$

Warning

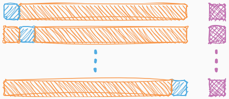
No guarantee on the prediction of $\hat{A}$ with scores based on $(\hat{A}_{-i})_i$, without assuming a form of **stability** on $\mathcal{A}$.

Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$

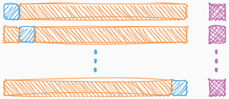
Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$

(in standard mean regression)

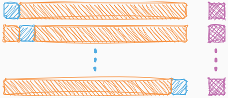
Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$
  (in standard mean regression)
- Build the predictive interval: $[q_{\alpha,\text{inf}}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Recall $q_{\beta,\text{inf}}(X_1, \ldots, X_n) := \lfloor \beta \times n \rfloor$ smallest value of $(X_1, \ldots, X_n)$

Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals

  $$\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$$

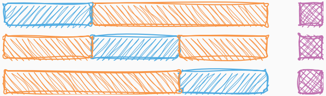
  (in standard mean regression)
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Theorem (Marginal validity of Jackknife+ Barber et al. (2021b)).

If $\mathcal{D}_n \cup (X_{n+1}, Y_{n+1})$ are exchangeable and $\mathcal{A}$ is symmetric: $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - 2\alpha$.

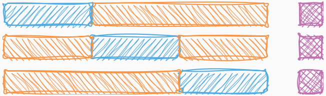
Recall $q_{\beta,\inf}(X_1, \ldots, X_n) := \lfloor \beta \times n \rfloor$ smallest value of $(X_1, \ldots, X_n)$

CV+ Barber et al. (2021b) (see also cross-conformal predictors: Vovk, 2015)



- Based on cross-validation residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Split $\mathcal{D}_n$ into K folds $F_1, \ldots, F_K$
- Get $\hat{A}_{-F_k}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus F_k$

CV+ Barber et al. (2021b) (see also cross-conformal predictors: Vovk, 2015)

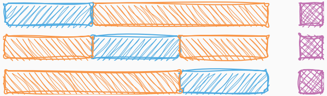


- Based on cross-validation residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Split $\mathcal{D}_n$ into K folds $F_1, \ldots, F_K$
- Get $\hat{A}_{-F_k}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus F_k$
- Cross-val predictions / predictive intervals

$$\mathcal{S}_{\text{up/down}} = \left\{ \left\{ \hat{A}_{-F_k}(X_{n+1}) \pm |\hat{A}_{-F_k}(X_i) - Y_i| \right\}_{i \in F_k} \right\}_k \cup \{\pm\infty\}$$

(in standard mean regression)

CV+ Barber et al. (2021b) (see also cross-conformal predictors: Vovk, 2015)



- Based on cross-validation residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Split $\mathcal{D}_n$ into K folds $F_1, \ldots, F_K$
- Get $\hat{A}_{-F_k}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus F_k$
- Cross-val predictions / predictive intervals

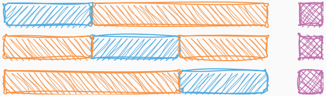
$$\mathcal{S}_{\text{up/down}} = \left\{ \left\{ \hat{A}_{-F_k}(X_{n+1}) \pm \left| \hat{A}_{-F_k}(X_i) - Y_i \right| \right\}_{i \in F_k} \right\}_k \cup \{\pm\infty\}$$

(in standard mean regression)

- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Recall $q_{\beta,\inf}(X_1, \ldots, X_n) := \lfloor \beta \times n \rfloor$ smallest value of $(X_1, \ldots, X_n)$

CV+ Barber et al. (2021b) (see also cross-conformal predictors: Vovk, 2015)



- Based on cross-validation residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Split $\mathcal{D}_n$ into K folds $F_1, \ldots, F_K$
- Get $\hat{A}_{-F_k}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus F_k$
- Cross-val predictions / predictive intervals

$$\mathcal{S}_{\text{up/down}} = \left\{ \left\{ \hat{A}_{-F_k}(X_{n+1}) \pm |\hat{A}_{-F_k}(X_i) - Y_i| \right\}_{i \in F_k} \right\}_k \cup \{\pm\infty\}$$

(in standard mean regression)

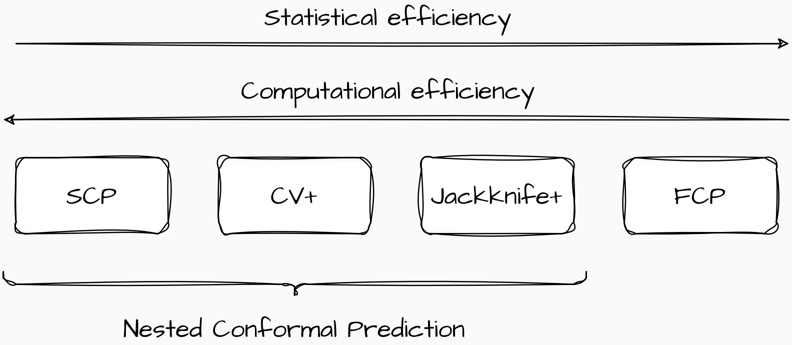
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Theorem (Marginal validity of CV+ Barber et al. (2021b)).

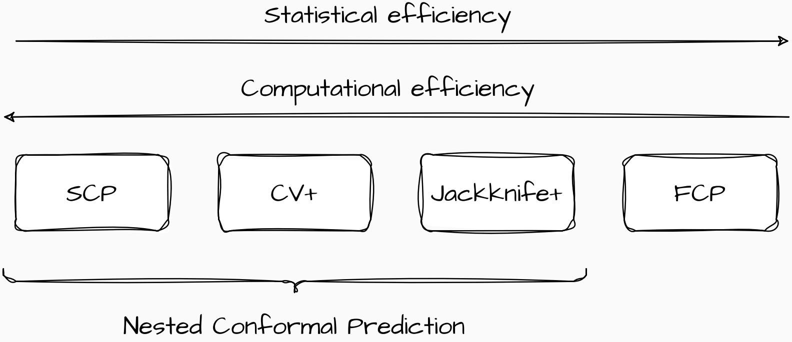
If $\mathcal{D}_n \cup (X_{n+1}, Y_{n+1})$ are exchangeable and $\mathcal{A}$ is symmetric: $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - 2\alpha - \min\left(\frac{2(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right) \geq 1 - 2\alpha - \sqrt{2/n}$.

Recall $q_{\beta,\inf}(X_1, \ldots, X_n) := \lfloor \beta \times n \rfloor$ smallest value of $(X_1, \ldots, X_n)$

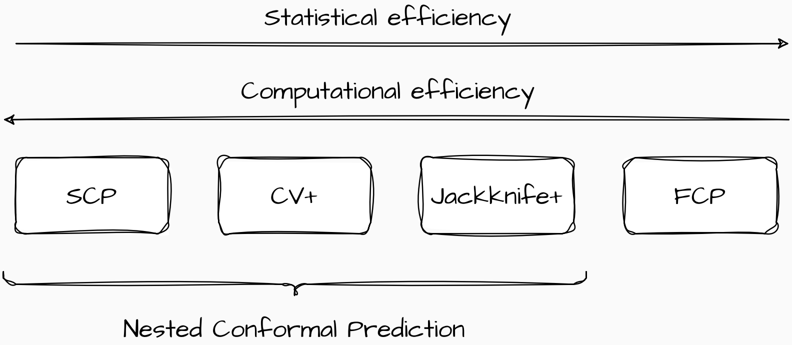
General overview



General overview



- Generalized framework encapsulating out-of-sample methods: Nested CP (Gupta et al., 2022) $\rightarrow$ extends *JK+/CV+ for any score.*



- Generalized framework encapsulating out-of-sample methods: Nested CP (Gupta et al., 2022) $\rightarrow$ extends *JK+/CV+ for any score.*
- Accelerating FCP: Nouretdinov et al. (2001); Lei (2019); Ndiaye and Takeuchi (2019); Cherubin et al. (2021); Ndiaye and Takeuchi (2022); Ndiaye (2022)

Non exhaustive references.

Quantile Regression

Split Conformal Prediction (SCP)

On the design choices of conformity scores and (empirical) conditional guarantees

Avoiding data splitting: full conformal and out-of-bags approaches

Beyond exchangeability

Some case studies

Concluding remarks

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant

✗ Arbitrary distribution shift

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant

✗ Arbitrary distribution shift

✗ Possibly many shifts, not only one

Covariate shift (Tibshirani et al., 2019)[12]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$

[12] Tibshirani et al. (2019), *Conformal Prediction Under Covariate Shift*, NeurIPS

Covariate shift (Tibshirani et al., 2019)[12]

- Setting:
  - $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point

[12] Tibshirani et al. (2019), *Conformal Prediction Under Covariate Shift*, NeurIPS

Covariate shift (Tibshirani et al., 2019)[12]

- Setting:
  - $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point
- In practice:
  1. estimate the likelihood ratio $w(X_i) = \dfrac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$

[12] Tibshirani et al. (2019), *Conformal Prediction Under Covariate Shift*, NeurIPS

Covariate shift (Tibshirani et al., 2019)[12]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point
- In practice:
  1. estimate the likelihood ratio $w(X_i) = \dfrac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$
  2. normalize the weights, i.e. $\omega_i = \omega(X_i) = \dfrac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$

[12] Tibshirani et al. (2019), *Conformal Prediction Under Covariate Shift*, NeurIPS

Covariate shift (Tibshirani et al., 2019)[12]

- Setting:
  - $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point
- In practice:
  1. estimate the likelihood ratio $w(X_i) = \dfrac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$
  2. normalize the weights, i.e. $\omega_i = \omega(X_i) = \dfrac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$
  3. outputs $\widehat{C}_\alpha(X_{n+1}) =$

$$\left\{ y : \mathrm{s}(\hat{A}(X_{n+1}), y) \leq Q_{1-\alpha}\left(\sum_{i \in \mathrm{Cal}} \omega_i \delta_{S_i} + \omega_{n+1}\delta_\infty \right) \right\}$$

[12] Tibshirani et al. (2019), *Conformal Prediction Under Covariate Shift*, NeurIPS

Label shift (Podkopaev and Ramdas, 2021)[13]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification

[13] Podkopaev and Ramdas (2021), *Distribution-free uncertainty quantification for classification under label*

Label shift (Podkopaev and Ramdas, 2021)[13]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point

[13] Podkopaev and Ramdas (2021), *Distribution-free uncertainty quantification for classification under label*

Label shift (Podkopaev and Ramdas, 2021)[13]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown

[13] Podkopaev and Ramdas (2021), *Distribution-free uncertainty quantification for classification under label*

Label shift (Podkopaev and Ramdas, 2021)[13]

- Setting:
  - $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown
- In practice:
  1. estimate the likelihood ratio $w(Y_i) = \dfrac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ using algorithms from the existing label shift literature

[13] Podkopaev and Ramdas (2021), *Distribution-free uncertainty quantification for classification under label*

Label shift (Podkopaev and Ramdas, 2021)[13]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown
- In practice:
  1. estimate the likelihood ratio $w(Y_i) = \dfrac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ using algorithms from the existing label shift literature
  2. normalize the weights, i.e. $\omega_i^y = \omega^y(X_i) = \dfrac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}$

[13] Podkopaev and Ramdas (2021), *Distribution-free uncertainty quantification for classification under label*

Label shift (Podkopaev and Ramdas, 2021)[13]

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown
- In practice:
  1. estimate the likelihood ratio $w(Y_i) = \frac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ using algorithms from the existing label shift literature
  2. normalize the weights, i.e. $\omega_i^y = \omega^y(X_i) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}$
  3. outputs $\widehat{C}_\alpha(X_{n+1}) =$

$$\left\{ y : s(\hat{A}(X_{n+1}), y) \leq Q_{1-\alpha}\left(\sum_{i \in \mathrm{Cal}} \omega_i^y \delta_{S_i} + \omega_{n+1}^y \delta_\infty \right) \right\}$$

[13] Podkopaev and Ramdas (2021), *Distribution-free uncertainty quantification for classification under label*

Generalizations

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:

Generalizations

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:
  - Chernozhukov et al. (2018)
    $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓

Generalizations

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:
  - Chernozhukov et al. (2018)
    $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓
  - Barber et al. (2022)
    $\hookrightarrow$ Quantifies the coverage loss depending on the strength of exchangeability violation

    $$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha - \substack{\text{average violation of exchangeability} \\ \text{by each calibration point}}$$

    $\hookrightarrow$ proposed algorithm: reweighting again!
    e.g., in a temporal setting, give higher weights to more recent points.

Online setting

- Data: T_0 random variables $(X_1, Y_1), \ldots, (X_{T_0}, Y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$
- Aim: predict the response values as well as predictive intervals for T_1 subsequent observations $X_{T_0+1}, \ldots, X_{T_0+T_1}$ sequentially: at any prediction step $t \in [\![T_0 + 1, T_0 + T_1]\!]$, $Y_{t-T_0}, \ldots, Y_{t-1}$ have been revealed
- Build the smallest interval $\widehat{C}_\alpha^t$ such that:

$$\mathbb{P}\left\{ Y_t \in \widehat{C}_\alpha^t(X_t) \right\} \geq 1 - \alpha, \quad \text{for } t \in [\![T_0 + 1, T_0 + T_1]\!],$$

often relaxed in:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y_t \in \widehat{C}_\alpha^t(X_t) \right\} \approx 1 - \alpha.$$

Recent developments

- Consider splitting strategies that respect the temporal structure

Non exhaustive references.

Recent developments

- Consider splitting strategies that respect the temporal structure
- Gibbs and Candès (2021) propose a method which reacts faster to temporal evolution
  - Idea: track the previous coverages of the predictive intervals $(\mathbb{1}\{Y_t \in \widehat{C}_\alpha(X_t)\})$
  - Tool: update the empirical quantile level with a learning rate γ
  - Asymptotic guarantee (on average) for any distribution (even adversarial)

Non exhaustive references.

Recent developments

- Consider splitting strategies that respect the temporal structure
- Gibbs and Candès (2021) propose a method which reacts faster to temporal evolution
  - Idea: track the previous coverages of the predictive intervals ($\mathbb{1}\{Y_t \in \widehat{C}_\alpha(X_t)\}$)
  - Tool: update the empirical quantile level with a learning rate γ
  - Asymptotic guarantee (on average) for any distribution (even adversarial)
- Zaffran et al. (2022) studies the influence of this learning rate γ and proposes, along with Gibbs and Candès (2022), a method not requiring to choose γ

Non exhaustive references.

Recent developments

- Consider splitting strategies that respect the temporal structure
- Gibbs and Candès (2021) propose a method which reacts faster to temporal evolution
    - Idea: track the previous coverages of the predictive intervals ($\mathbb{1}\{Y_t \in \widehat{C}_\alpha(X_t)\}$)
    - Tool: update the empirical quantile level with a learning rate γ
    - Asymptotic guarantee (on average) for any distribution (even adversarial)
- Zaffran et al. (2022) studies the influence of this learning rate γ and proposes, along with Gibbs and Candès (2022), a method not requiring to choose γ
- Bhatnagar et al. (2023) enjoys **anytime** regret bound, by leveraging tools from the strongly adaptive regret minimization literature

Non exhaustive references.

Recent developments

- Consider splitting strategies that respect the temporal structure
- Gibbs and Candès (2021) propose a method which reacts faster to temporal evolution
  - Idea: track the previous coverages of the predictive intervals $(\mathbb{1}\{Y_t \in \widehat{C}_\alpha(X_t)\})$
  - Tool: update the empirical quantile level with a learning rate γ
  - Asymptotic guarantee (on average) for any distribution (even adversarial)
- Zaffran et al. (2022) studies the influence of this learning rate γ and proposes, along with Gibbs and Candès (2022), a method not requiring to choose γ
- Bhatnagar et al. (2023) enjoys anytime regret bound, by leveraging tools from the strongly adaptive regret minimization literature
- Bastani et al. (2022) proposes an algorithm achieving stronger coverage guarantees (conditional on specified overlapping subsets, and threshold calibrated) without hold-out set

Non exhaustive references.

- Medical application
- Image based task
- Pixel by pixel analysis $\rightsquigarrow$ applications to segmentation for self-driving cars

Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos [*1] Amit Kohli [*1] Stephen Bates [1] Michael I. Jordan [1] Jitendra Malik [1]
Thayer Alshaabi [2] Srigokul Upadhyayula [2 3] Yaniv Romano [4]

Image to Image regression with DF-UQ - Angelopoulos et al. (2022b)

- Medical application
- Image based task
- Pixel by pixel analysis $\rightsquigarrow$ applications to segmentation for self-driving cars

1. **Task**: *Image to Image regression* - for each pixel of an image, predict a real valued output from the entire image.
2. **UQ Goal:** provide a predictive interval for each pixel, such that the output is in the interval at least 90% of the time.

Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos [*1] **Amit Kohli** [*1] **Stephen Bates** [1] **Michael I. Jordan** [1] **Jitendra Malik** [1]
Thayer Alshaabi [2] **Srigokul Upadhyayula** [2 3] **Yaniv Romano** [4]

- Medical application
- Image based task
- Pixel by pixel analysis $\rightsquigarrow$ applications to segmentation for self-driving cars

1. **Task**: *Image to Image regression* - for each pixel of an image, predict a real valued output from the entire image.
2. **UQ Goal:** provide a predictive interval for each pixel, such that the output is in the interval at least 90% of the time.

Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos [*1] Amit Kohli [*1] Stephen Bates [1] Michael I. Jordan [1] Jitendra Malik [1] Thayer Alshaabi [2] Srigokul Upadhyayula [2 3] Yaniv Romano [4]

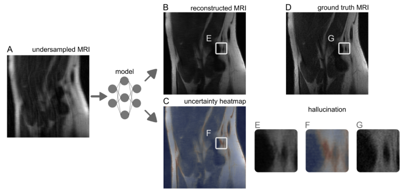


Figure 1. **An algorithmic MRI reconstruction with uncertainty.** A rapidly acquired but undersampled MR image of a knee (A) is fed into a model that predicts a sharp reconstruction (B) with calibrated uncertainty (C). In (C), red means high uncertainty and blue means low uncertainty. Wherever the reconstruction contains hallucinations, the uncertainty is high; see the hallucination in the image patch (E), which has high uncertainty in (F), and does not exist in the ground truth (G). For experimental details, see Section 3.4.

Figure 2: Image from Angelopoulos et al. (2022b)

Image to Image regression with DF-UQ - Angelopoulos et al. (2022b)

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
    - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0; 1]$

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0;1]$
   - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \to [0;1]$, such that

$$[\hat{\mu}(X) - \tilde{\ell}(X); \hat{\mu}(X) + \tilde{u}(X)]$$

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.

Image to Image regression with DF-UQ - Angelopoulos et al. (2022b)

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0; 1]$
   - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \to [0; 1]$, such that

   $$[\hat{\mu}(X) - \tilde{\ell}(X); \hat{\mu}(X) + \tilde{u}(X)]$$

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.
3. Calibration step: leverage the calibration set.
   - In spirit, almost equivalent to CQR but with a multiplicative form.
   - Precisely, relies on RCPS (Bates et al., 2021a)

Image to Image regression with DF-UQ - Angelopoulos et al. (2022b)

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0; 1]$
   - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \to [0; 1]$, such that

   $$[\hat{\mu}(X) - \tilde{\ell}(X); \hat{\mu}(X) + \tilde{u}(X)]$$

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.
3. Calibration step: leverage the calibration set.
   - In spirit, almost equivalent to CQR but with a multiplicative form.
   - Precisely, relies on RCPS (Bates et al., 2021a)

Guarantee:

$$\mathbb{P}\left[\mathbb{E}\left[\text{Average miscoverage on all pixels of a test image} \geq \alpha | \text{Cal}\right]\right] \leq \delta$$

$\to$ Marginal validity on the test, with high probability w.r.t. the calibration set.

Abstract

Image-to-image regression is an important learning task, used frequently in biological imaging. Current algorithms, however, do not generally offer statistical guarantees that protect against a model's mistakes and hallucinations. To address this, we develop uncertainty quantification techniques with rigorous statistical guarantees for image-to-image regression problems. In particular, we show how to derive uncertainty intervals around each pixel that are guaranteed to contain the true value with a user-specified confidence probability. Our methods work in conjunction

2. Methods

We now formally describe the method for constructing uncertainty intervals. Each pixel in the image will get its own uncertainty interval, as in (1), that is statistically guaranteed to contain the true value with high probability.

How do you understand that?

Image to Image regression with DF-UQ - Angelopoulos et al. (2022b)

Abstract

Image-to-image regression is an important learning task, used frequently in biological imaging. Current algorithms, however, do not generally offer statistical guarantees that protect against a model's mistakes and hallucinations. To address this, we develop uncertainty quantification techniques with rigorous statistical guarantees for image-to-image regression problems. In particular, we show how to derive uncertainty intervals around each pixel that are guaranteed to contain the true value with a user-specified confidence probability. Our methods work in conjunction

2. Methods

We now formally describe the method for constructing uncertainty intervals. Each pixel in the image will get its own uncertainty interval, as in (1), that is statistically guaranteed to contain the true value with high probability.

- Not a conditional coverage claim!
- The statement is on-average on the test point - easy or hard.

Size-stratified risk. Next, we seek prediction sets that do not systematically make mistakes in difficult parts of the image. Our risk control requirement in Definition 2.1 may be satisfied even if the prediction sets systematically fail to contain the most difficult pixels. For example, if $\alpha = 0.1$ and 90% of pixels are covered by fixed-width intervals of size 0.01, then the requirement is satisfied—however, the sets no longer serve as useful notions of uncertainty. To

- Hard problem (impossibility results!)
- Introduce metrics to see *if* and *on which underlying regressors* such problem happens.

How do you understand that?

Example of such metrics (see also Feldman et al., 2021) **:**

- Link between the size of the PI and the coverage level $\longrightarrow$

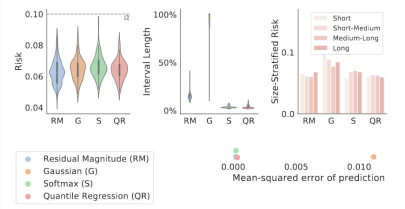
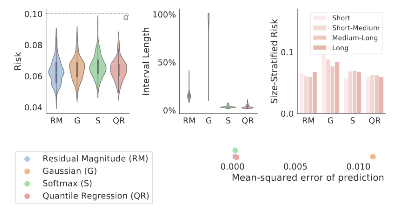


Image to Image regression with DF-UQ - Angelopoulos et al. (2022b)

Example of such metrics (see also Feldman et al., 2021) **:**

- Link between the size of the PI and the coverage level $\longrightarrow$
- Localization of the errors $\downarrow$



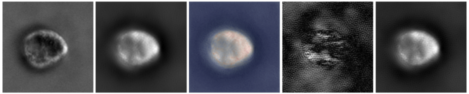


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth (renormalized for visualization), ground truth.

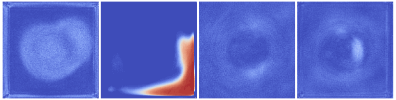
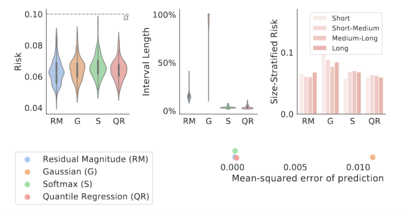


Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

Figure 3: All images from Angelopoulos et al. (2022b)

Example of such metrics (see also Feldman et al., 2021) **:**

- Link between the size of the PI and the coverage level $\longrightarrow$
- Localization of the errors $\downarrow$



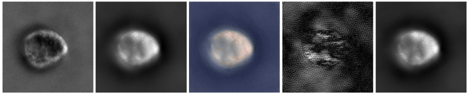


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth (renormalized for visualization), ground truth.

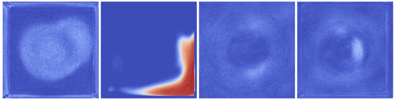


Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

Figure 3: All images from Angelopoulos et al. (2022b)

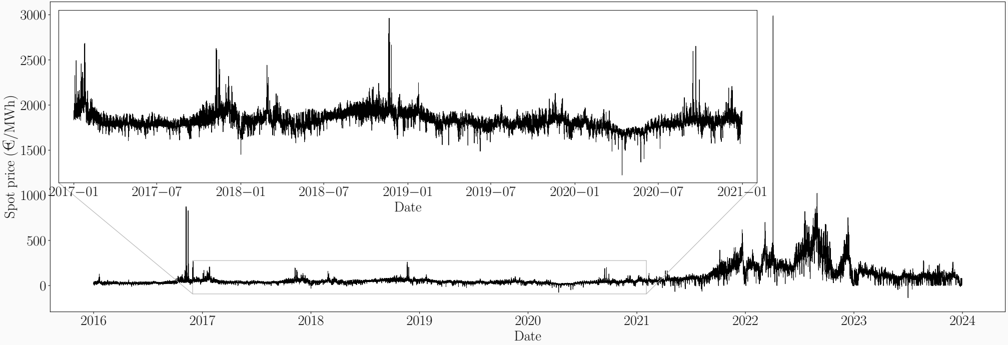
Take aways:

- Elegant application of SCP with CQR type score
- Test marginal and calibration + train conditional validity guarantees with HP
- Main problem is Test conditionality $\rightarrow$ look at metrics to evaluate which methods performs best!

Hourly day-ahead market prices (between producers and suppliers)

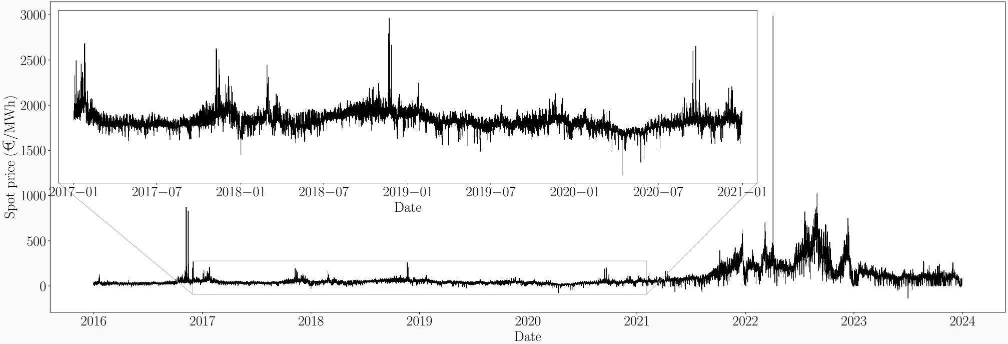
Forecasting French spot electricity prices

Hourly day-ahead market prices (between producers and suppliers)



Forecasting French spot electricity prices

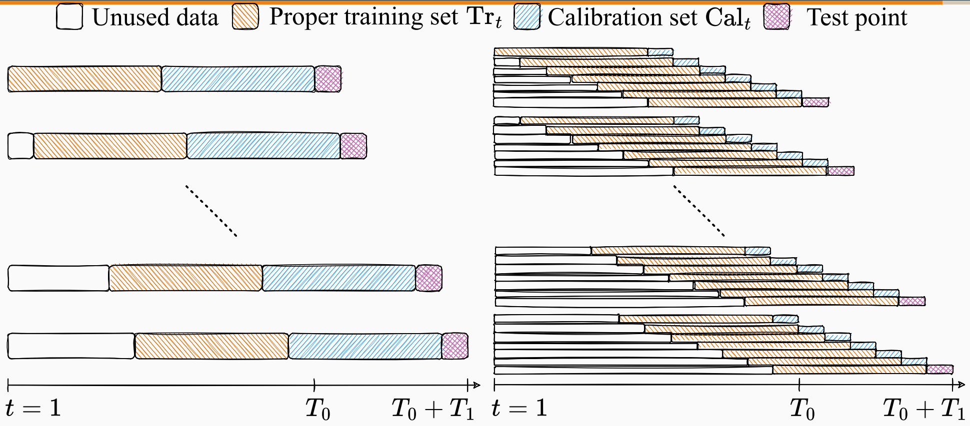
Hourly day-ahead market prices (between producers and suppliers)



To which extent are they forecastable?

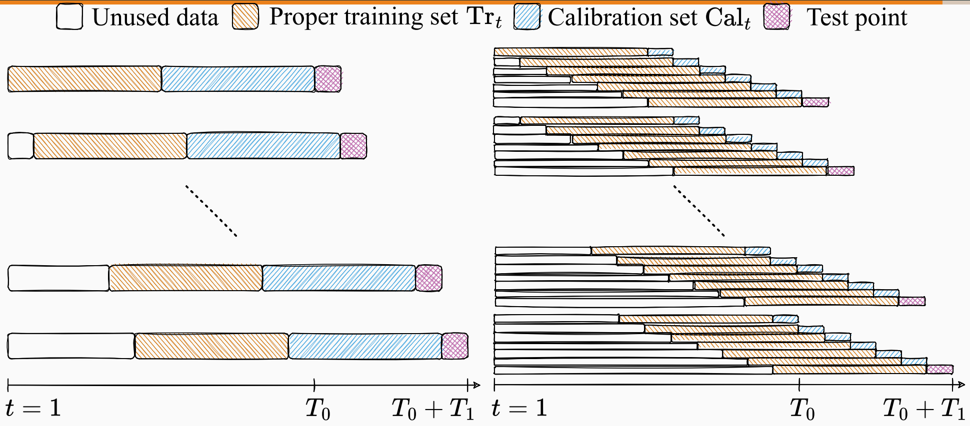
$\hookrightarrow$ forecasts errors **no lower than 10%** of the realized price!

Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

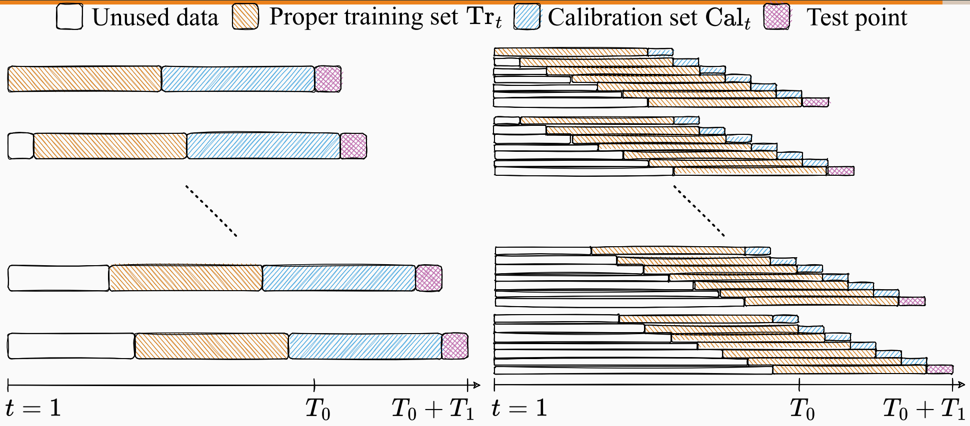
Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

$\hookrightarrow$ `OSSCP` improves robustness in temporal settings;

Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

$\hookrightarrow$ `OSSCP` improves robustness in temporal settings;

$\hookrightarrow$ `OSSCP-horizon` drastically improves robustness in non-stationary temporal settings.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}\left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t}\left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}\left\{Y^{(t)} \notin \widehat{C}_{\alpha_t}\left(X^{(t)}\right)\right\}\right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Guarantee: Asymptotic validity result for any sequence of observations.

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y^{(t)} \in \widehat{C}_{\alpha_t}\left(X^{(t)}\right)\right\} \underset{T_1 \to +\infty}{\longrightarrow} 1 - \alpha$$

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Guarantee: Asymptotic validity result for any sequence of observations.

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1} \left\{ Y^{(t)} \in \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} - (1-\alpha) \right| \leq \frac{2}{\gamma T_1}$$

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}\left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t}\left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

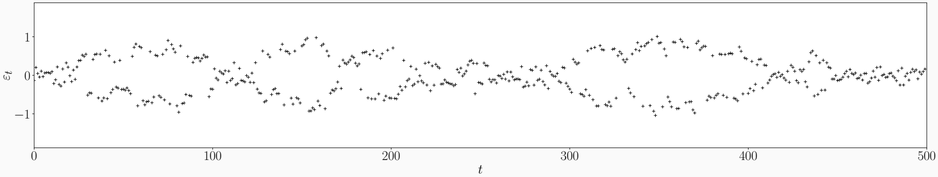
Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Guarantee: Asymptotic validity result for any sequence of observations.

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y^{(t)} \in \widehat{C}_{\alpha_t}\left(X^{(t)} \right) \right\} - (1-\alpha) \right| \leq \frac{2}{\gamma T_1}$$

$\Rightarrow$ favors large γ.

Visualisation of ACI procedure



Visualisation of ACI procedure

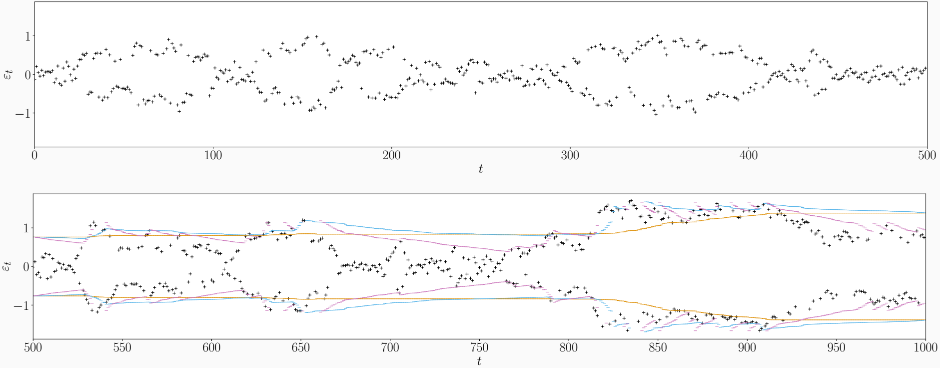
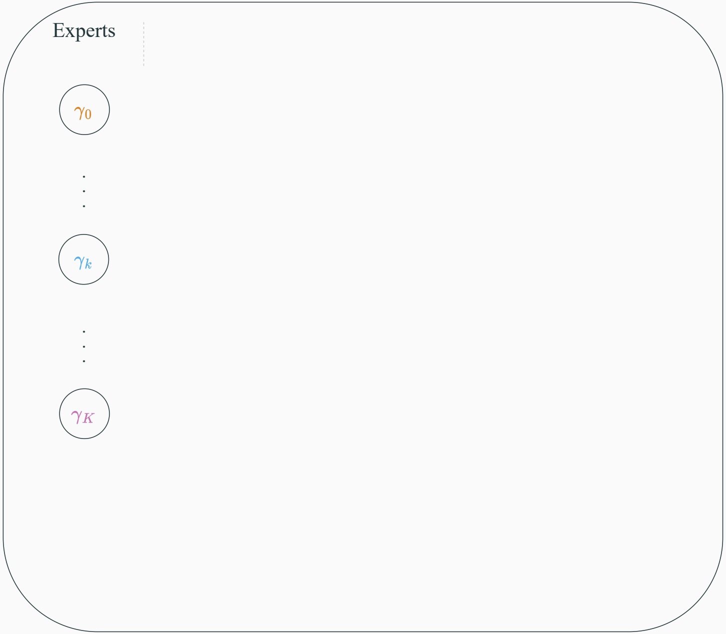
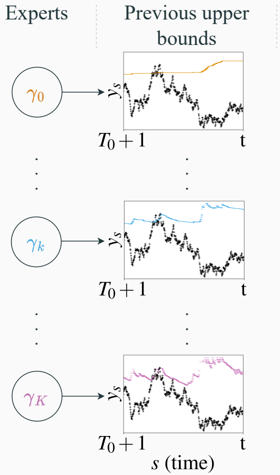


Figure 4: Visualisation of ACI with different values of γ ($\gamma = 0$, $\gamma = 0.01$, $\gamma = 0.05$)

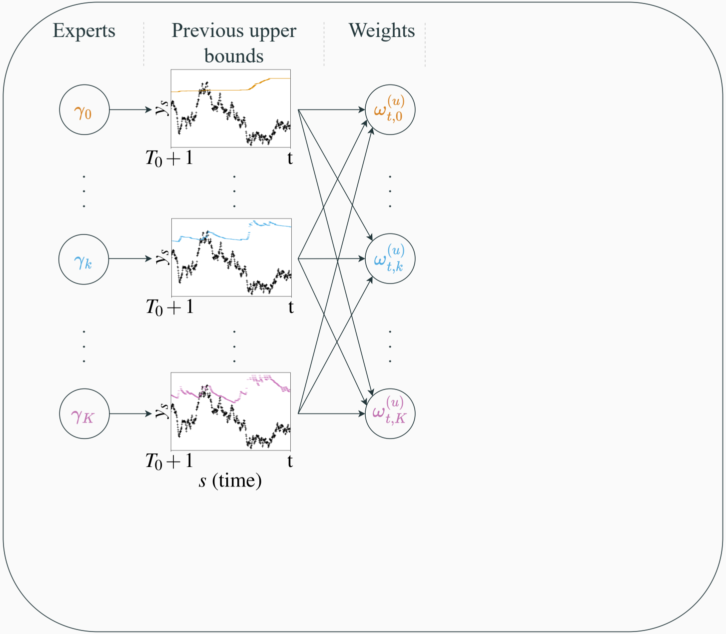
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



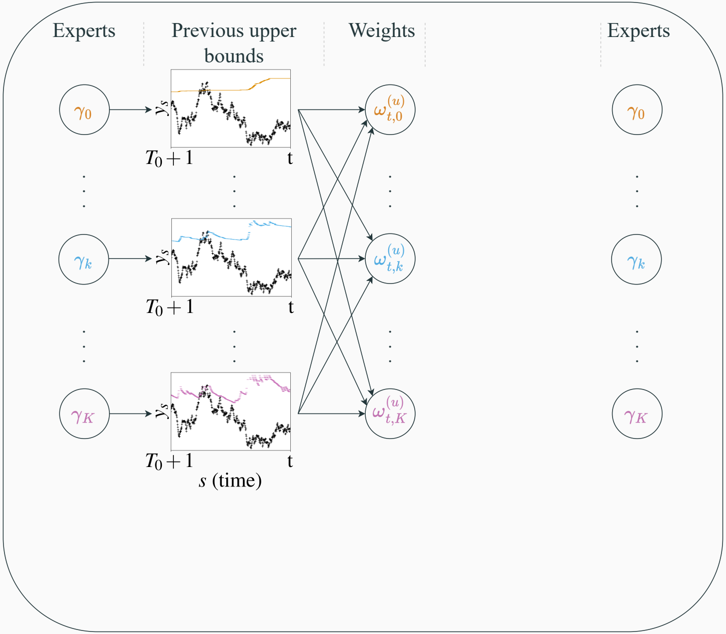
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



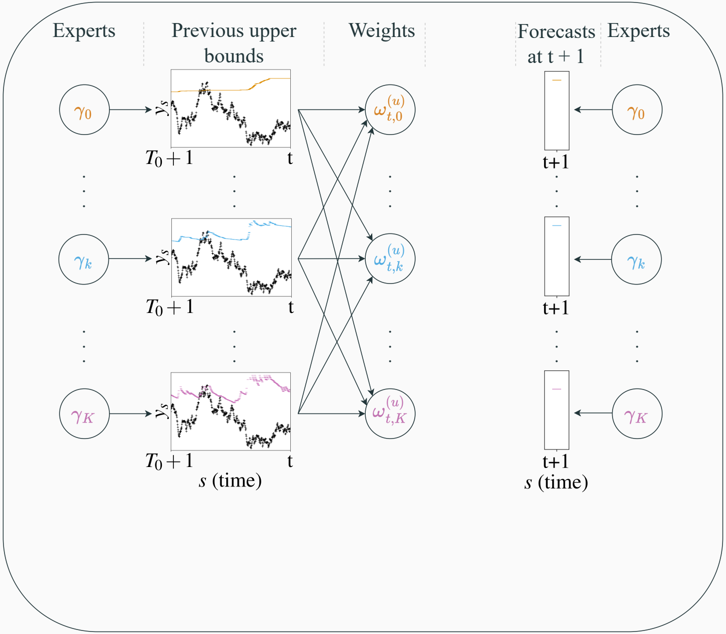
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



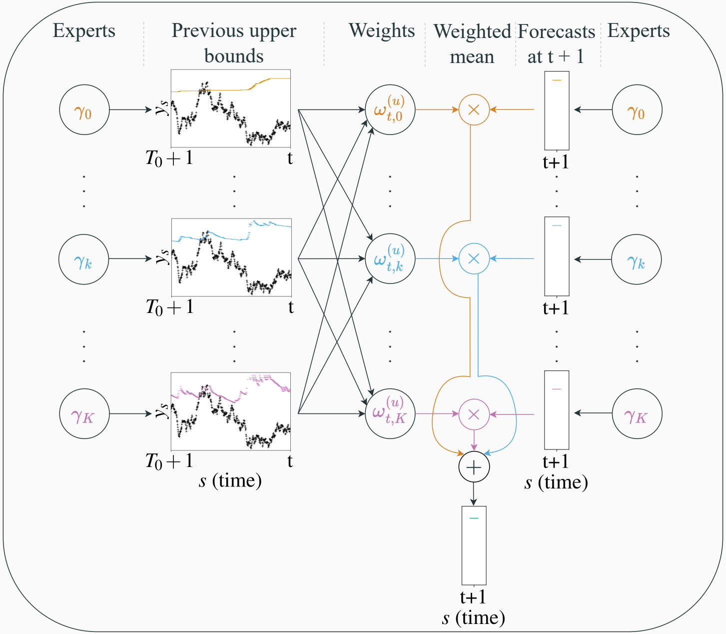
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- **Synthetic data with ARMA noise**

Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- **Synthetic data with ARMA noise**
  - Benchmarks are not robust to the increase in the temporal dependence;

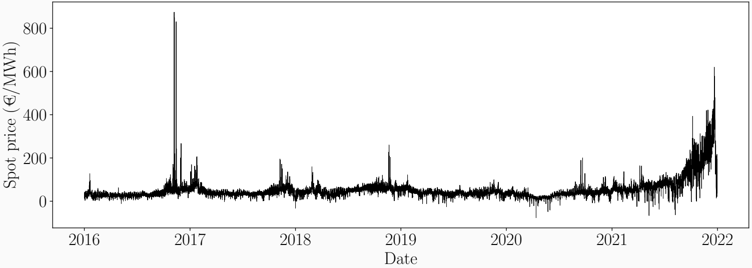
- **Synthetic data with ARMA noise**
  - Benchmarks are not robust to the increase in the temporal dependence;
  - ACI is robust, maintaining validity, with an appropriate γ;

Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

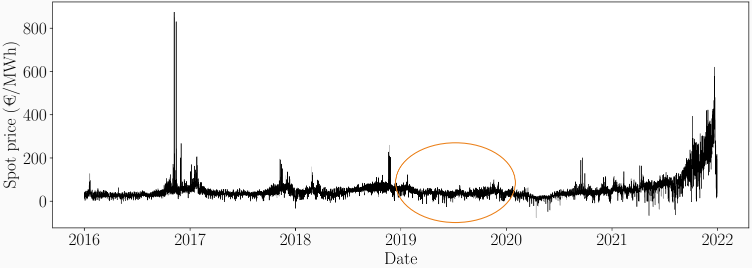
- **Synthetic data with ARMA noise**
  - Benchmarks are not robust to the increase in the temporal dependence;
  - ACI is robust, maintaining validity, with an appropriate γ;
  - `AgACI` is robust, maintaining validity, not the smallest.

Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- **Synthetic data with ARMA noise**
  - Benchmarks are not robust to the increase in the temporal dependence;
  - ACI is robust, maintaining validity, with an appropriate γ;
  - `AgACI` is robust, maintaining validity, not the smallest.
- **French electricity spot prices**

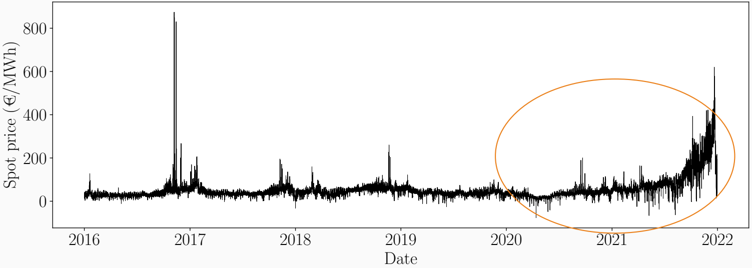


- **Synthetic data with ARMA noise**
  - Benchmarks are not robust to the increase in the temporal dependence;
  - ACI is robust, maintaining validity, with an appropriate γ;
  - `AgACI` is robust, maintaining validity, not the smallest.
- **French electricity spot prices**
  - 2019: `AgACI` provides validity with a reasonable efficiency;



Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- **Synthetic data with ARMA noise**
  - Benchmarks are not robust to the increase in the temporal dependence;
  - ACI is robust, maintaining validity, with an appropriate γ;
  - `AgACI` is robust, maintaining validity, not the smallest.
- **French electricity spot prices**
  - 2019: `AgACI` provides validity with a reasonable efficiency;
  - 2020 and 2021: `AgACI` fails to ensure validity, and the various forecasting models considered behave differently.



Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

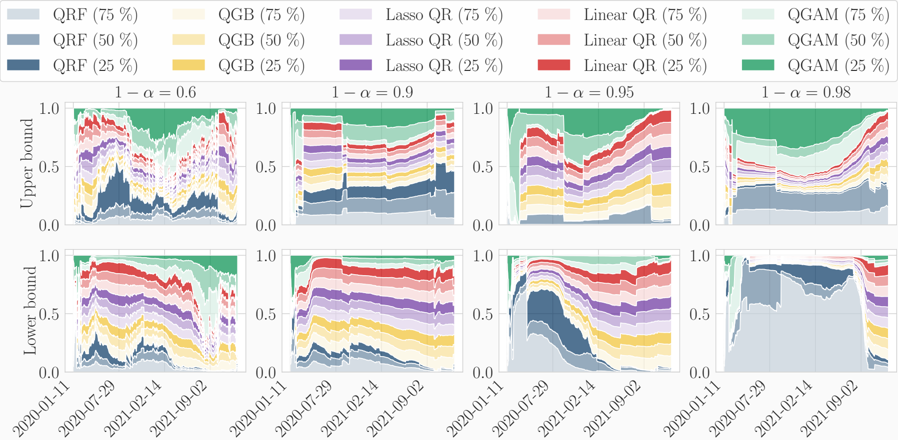
Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.
- ✓ Analyzing its weights provides interpretability.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.
- ✓ Analyzing its weights provides interpretability.



Highlights and perspectives

Aggregating the two bounds independently (as in `AgACI` and beyond):

Highlights and perspectives

Aggregating the two bounds independently (as in `AgACI` and beyond):

✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails

Aggregating the two bounds independently (as in `AgACI` and beyond):

- ✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails
- ✗ Prevents from obtaining theoretical guarantees (by opposition to Gibbs and Candès, 2022)

Highlights and perspectives

Aggregating the two bounds independently (as in `AgACI` and beyond):

- ✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails
- ✗ Prevents from obtaining theoretical guarantees (by opposition to Gibbs and Candès, 2022)

$\hookrightarrow$ Weaken the objective and consider a more practical theoretical aim?

Quantile Regression

Split Conformal Prediction (SCP)

On the design choices of conformity scores and (empirical) conditional guarantees

Avoiding data splitting: full conformal and out-of-bags approaches

Beyond exchangeability

Some case studies

Concluding remarks

Summary: Uncertainty quantification through conformal methods

Quantile regression → no guarantee →

Conformal methods

- Marginal validity
- Model/distribution agnostic
- Finite sample

SCP

+ no retraining
- split

Full CP

+ no split
- prohibitive cost

JK/CV

+ balance cost
- requires stability or $1 - 2\alpha$

Various scores

- Based on $\hat{\mu}(x)$
- Based on $\widehat{QR}(x)$
- Based on $\hat{P}(y|x)$

Limit 1: only marginal guarantee

- Hardness of conditional coverage
- Asymptotic results / Adaptive methods

Limit 2: requires exch.

- Distribution shifts
- Online methods

Some (other, non-exhaustives) current open directions

- Outlier detection (Vovk et al., 2003; Bates et al., 2023)
- Selective inference, false discovery rate guarantees (Marandon et al., 2024; Gazin et al., 2024)
- Beyond the indicator loss (Angelopoulos et al., 2022a; Bates et al., 2021b; Angelopoulos et al., 2023; Lekeufack et al., 2024)
- Aggregating predictive sets (Gasparin and Ramdas, 2024b,a; Gasparin et al., 2024)

Acknowledgments

For discussion and feedback, thanks to

- Julie Josse
- Claire Boyer
- Étienne Roquain

Questions?

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4).

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. (2022a). Learn then test: Calibrating predictive algorithms to achieve risk control.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2023). Conformal risk control.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. (2022b). Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2).

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

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. To appear in *Annals of Statistics (2023)*.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2022). Practical adversarial multivalid conformal prediction. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. (2021a). Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. (2021b). Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6).

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178.

Bhatnagar, A., Wang, H., Xiong, C., and Bai, Y. (2023). Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Bian, M. and Barber, R. F. (2023). Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044 – 2066.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust Validation: Confident Predictions Even When Distributions Shift. arXiv: 2008.04267.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. In *Conference On Learning Theory*. PMLR.

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

Cherubin, G., Chatzikokolakis, K., and Jaggi, M. (2021). Exact optimization of conformal predictors via incremental and decremental learning. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR.

Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. (2023). Class-conditional conformal prediction with many classes. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64555–64576. Curran Associates, Inc.

Dutot, G., Zaffran, M., Féron, O., and Goude, Y. (2024). Adaptive probabilistic forecasting of french electricity spot prices.

Feldman, S., Bates, S., and Romano, Y. (2021). Improving Conditional Coverage via Orthogonal Quantile Regression. *arXiv:2106.00394 [cs]*. arXiv: 2106.00394.

Gasparin, M. and Ramdas, A. (2024a). Conformal online model aggregation.

Gasparin, M. and Ramdas, A. (2024b). Merging uncertainty sets via majority vote.

Gasparin, M., Wang, R., and Ramdas, A. (2024). Combining exchangeable p-values.

Gazin, U., Blanchard, G., and Roquain, E. (2024). Transductive conformal inference with adaptive scores.

Gibbs, I. and Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. arXiv: 2208.08401.

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

Guan, L. (2022). Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1).

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127.

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).

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2023). Batch multivalid conformal prediction. In *International Conference on Learning Representations*.

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2020). Adaptive, Distribution-Free Prediction Intervals for Deep Networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4).

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*.

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).

Lekeufack, J., Angelopoulos, A. N., Bajcsy, A., Jordan, M. I., and Malik, J. (2024). Conformal decision theory: Safe autonomous decisions from imperfect predictions.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2024). Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157 – 183.

Ndiaye, E. (2022). Stable conformal prediction sets. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

Ndiaye, E. and Takeuchi, I. (2019). Computing full conformal prediction set with approximate homotopy. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Ndiaye, E. and Takeuchi, I. (2022). Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1).

Nouretdinov, I., Melluish, T., and Vovk, V. (2001). Ridge regression confidence machine. In *Proceedings of the 18th International Conference on Machine Learning*.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In *Machine Learning: ECML*. Springer.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. PMLR.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. (2020a). With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2).

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Romano, Y., Sesia, M., and Candes, E. (2020b). Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Sadinle, M., Lei, J., and Wasserman, L. (2018). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. (2019). Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Vovk, V. (2012). Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*. PMLR.

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2).

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2003). Testing exchangeability on-line. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 768–775. AAAI Press.

Zaffran, M., Dieuleveut, A., Josse, J., and Romano, Y. (2024). Predictive uncertainty quantification with missing values. Preprint submitted to *Journal of Machine Learning Research*, arXiv arXiv:2405.15641.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

cc BY NC SA