NB: Open with an advanced pdf reader (e.g., Acrobat to have animations)

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

---

Aymeric Dieuleveut & Margaux Zaffran

July 22th, 2024

41st International Conference on Machine Learning (ICML)

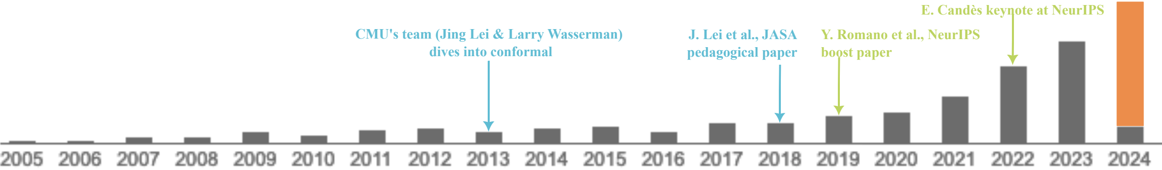
Why are we all here today?

(Slides available on our webpages)

Why are we all here today?

(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)

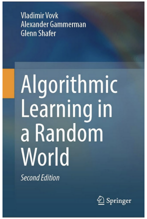
- Because Conformal Prediction has been a **popular** topic recently.
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.

  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.



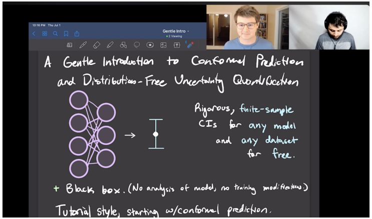
Vladimir Vovk
Alexander Gammerman
Glenn Shafer

Algorithmic Learning in a Random World

Second Edition

Springer

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



A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification

Rigorous, finite-sample CIs for any model and any dataset for free.

+ Black box. (No analysis of model, no training modifications)

Tutorial style, starting w/conformal prediction.

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

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

1 Introduction

Conformal prediction is a relatively new framework for quantifying uncertainty in the predictions made by arbitrary prediction algorithms. Fundamentally, it does so by converting an algorithm's predictions into prediction sets, which have strong finite-sample coverage properties.

The idea behind conformal prediction was born out of conversations between Vladimir Vovk and his colleagues, including Alexander Gammerman and Vladimir Vapnik (who was visiting the university), in the mid 1990s at Royal Holloway, University of College London. The definitive reference is Vovk et al. (2005). This remained a topic of intense interest for Vovk and collaborators for many years, up through current day. It is Larry Wasserman who brought about the interest in the statistics community, and inspired Carnegie Mellon University colleagues including Jing Lei (and the author of these lecture notes), to collaborate on conformal prediction in the mid 2010s. This lecture will draw mostly from the language for conformal prediction developed in Lei et al. (2018); Tibshirani et al. (2019).

The community working on conformal prediction remained fairly small for most of its short history, until fairly recently—just the last few years, really—when it exploded in popularity in the machine learning community. As such, there will be a lot of interesting work about conformal prediction that we will not be covering. A nice recent overview is Angelopoulos and Bates (2023). The last section of that monograph provides some sense of the current trends in the field.

1.1 A lofty goal?

The basic goal of conformal prediction is as follows. Let $(X_i, Y_i) \sim P$, $i = 1, \dots, n$ be i.i.d. feature and response pairs, from a distribution P on $\mathcal{X} \times \mathcal{Y}$. For concreteness, we can think of the feature space as say, $\mathcal{X} = \mathbb{R}^d$, and the response space as $\mathcal{Y} = \mathbb{R}$, though this need not be the case in general. Let $\alpha \in (0,1)$ be and a nominal error level. Then we want to find a prediction band,

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

with the property that for a new i.i.d. pair $(X_{n+1}, Y_{n+1}) \sim P$,

$$\mathbb{P}(Y_{n+1} \in \hat{C}_n(X_{n+1})) \geq 1 - \alpha, \quad (1)$$

where the probability is over all of our data (X_i, Y_i), $i = 1, \dots, n+1$.

On the one hand, without placing any assumptions on P, and without appealing to asymptotics of any kind, this might seem like a really hard goal in general. On the other hand, we can do something trivial to obtain it: for example,

$$\hat{C}_n(X_{n+1}) = \begin{cases} \mathcal{Y} & \text{with probability } 1-\alpha \\ \emptyset & \text{with probability } \alpha \end{cases}$$

will always have exactly $1-\alpha$ coverage, that is, it will achieve (1) with an equality.

So the real question is this (albeit still somewhat vaguely-phrased): can we achieve (1), in finite samples, without any assumptions on P, by doing something "nontrivial"? In particular, we would like our strategy to adapt to the hardness of the problem, in the following sense: the more easily we can predict Y_{n+1} from X_{n+1}, the smaller we would like our set $\hat{C}_n(X_{n+1})$ to be.

1

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.

→ **All material including sources will be accessible soon on**

learn conformal prediction

→ Builds upon earlier material accessible on this webpage

→ 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 insights on how to leverage those techniques in your own fields

Disclaimers

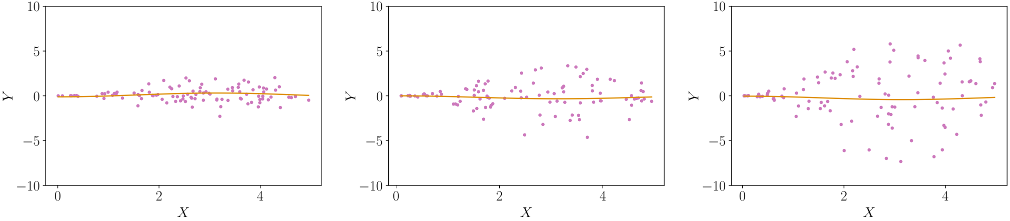
- Many people contributed to the domain - list of references may not be exhaustive
- Multiple other excellent resources

On the importance of quantifying uncertainty

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

On the importance of quantifying uncertainty

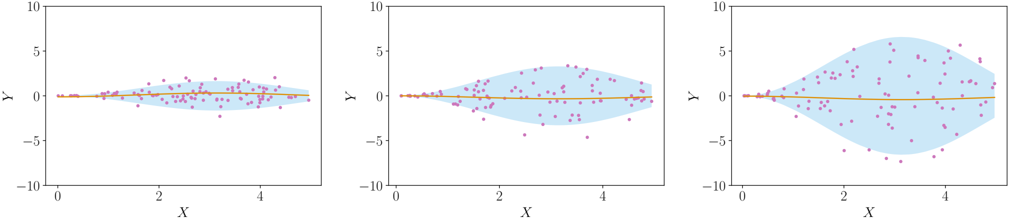
- **Obvious in most applications - weather, medical, markets**
- **Mathematically**



↪ Same “best” predictor, yet 3 distinct underlying phenomena!

On the importance of quantifying uncertainty

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



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

$\Longrightarrow$ Quantifying uncertainty conveys this information.

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

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) (Vovk et al., 2005)

without anim

SCP: implementation details

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

SCP: implementation details

without anim

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*

SCP: implementation details

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

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

▸ without anim

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 $q_{1-\alpha}(\mathcal{S})$[1]

---

[1] Equivalently, let $\mathcal{S}$ be the set of $\#\mathrm{Cal}$ conformity scores (i.e. without adding $\{+\infty\}$). Compute the $(1-\alpha)(1/\#\mathrm{Cal}+1)$ quantile of these scores $\mathcal{S}$.

SCP: implementation details

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 $q_{1-\alpha}(\mathcal{S})$[1]
6. For a new point X_{n+1}, return

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

---

[1] Equivalently, let $\mathcal{S}$ be the set of $\#\mathrm{Cal}$ conformity scores (i.e. without adding $\{+\infty\}$). Compute the $(1-\alpha)(1/\#\mathrm{Cal}+1)$ quantile of these scores $\mathcal{S}$.

SCP: theoretical foundation

SCP: theoretical foundation

Definition (Exchangeability)**.**

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

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

SCP: theoretical foundation

Definition (Exchangeability)**.**

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

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

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

SCP: theoretical foundation

Definition (Exchangeability).

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

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

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

Example (exchangeable sequences).

- i.i.d. samples

SCP: theoretical foundation

Definition (Exchangeability).

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

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

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

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.

Proof architecture of SCP guarantees

Lemma (Quantile lemma).

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

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

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

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

Proof architecture of SCP guarantees

Lemma (Quantile lemma).

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

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

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

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

$(X_i, Y_i)_{i=1}^{n+1}$ exchangeable $\Rightarrow \{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ exchangeable.

Proof architecture of SCP guarantees

Lemma (Quantile lemma).

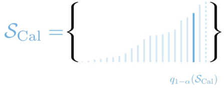
If $(U_1, \dots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0,1[$:

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

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

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

$(X_i, Y_i)_{i=1}^{n+1}$ exchangeable $\Rightarrow \{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ exchangeable.

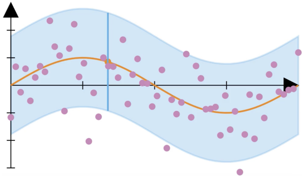


$\hookrightarrow$ quantile lemma to the scores gives the result.

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

$$= \left\{\left|Y_{n+1} - \hat{\mu}\left(X_{n+1}\right)\right| \leq q_{1-\alpha}\left(\mathcal{S}\right)\right\}$$

$$\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right)\right\} = \left\{S_{n+1} \leq q_{1-\alpha}\left(\mathcal{S}\right)\right\}.$$



Proof of the quantile lemma

Def: Empirical quantile $q_{\beta}(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

Proof of the quantile lemma

Def: Empirical quantile $q_{\beta}(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

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

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \dots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \dots, U_k)$

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

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \dots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \dots, U_k)$

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

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

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.

✓ Distribution free, model (regressor) free, finite sample average validity guarantee.

Standard mean-regression SCP – strength: validity – good vs bad estimator

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:

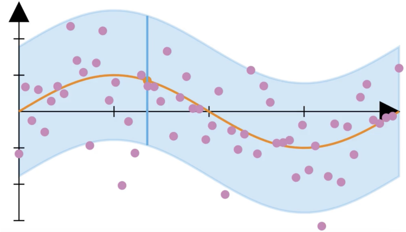
$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

---

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

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

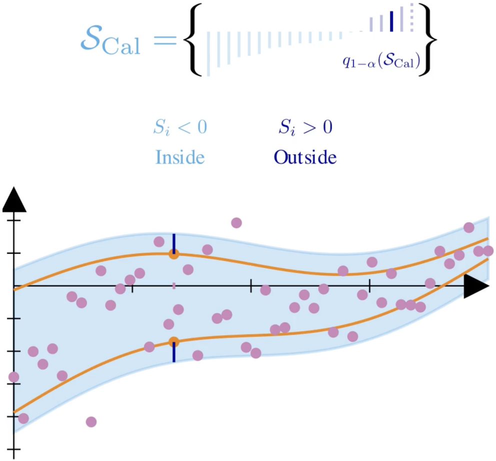
Standard mean-regression SCP – weakness: not adaptive



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

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

Conformalized Quantile Regression (CQR)



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

Thus

$$\left\{ Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \right\} = \left\{ S_{n+1} \leq q_{1-\alpha}\left(\mathcal{S}\right) \right\}.$$

⇝ Marginal validity is ensured, independently of the underlying quantile level or regressor quality. ✓

CQR: under vs over coverage

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $q_{1-\alpha}\left(\mathcal{S}\right)$
5. For a new point X_{n+1}, return

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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)

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:

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

↪ marginal also over the whole calibration set and the test point!

Challenges: open questions (non exhaustive!)

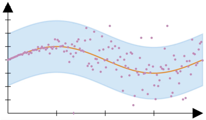
- Conditional coverage
- Computational cost vs statistical power
- Exchangeability

SCP: what choices for the regression scores?

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

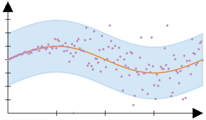
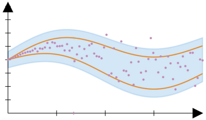
SCP: what choices for the regression scores?

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

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

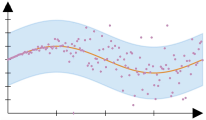
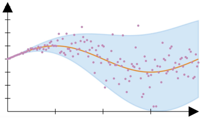
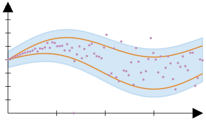
SCP: what choices for the regression scores?

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

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

SCP: what choices for the regression scores?

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

	Standard SCP Vovk et al. (2005)	**Locally weighted SCP** Lei et al. (2018)	**CQR** Romano et al. (2019)
$\mathrm{s}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$	$\frac{\lvert\hat{\mu}(X) - Y\rvert}{\hat{\rho}(X)}$	$\max(\widehat{\mathrm{QR}}_{\text{lower}}(X) - Y,$ $Y - \widehat{\mathrm{QR}}_{\text{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})\hat{\rho}(x)]$	$[\widehat{\mathrm{QR}}_{\text{lower}}(x) - q_{1-\alpha}(\mathcal{S});$ $\widehat{\mathrm{QR}}_{\text{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
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:

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

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:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(x\right)\right) = \infty\right) \geq 1 - \alpha.$$

Analogous statement is also available for the classification framework.

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:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(x\right)\right) = \infty\right) \geq 1 - \alpha.$$

↪ distribution-free X-conditional hardness result applies beyond CP

Analogous statement is also available for the classification framework.

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:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\operatorname{mes}\left(\widehat{C}_\alpha\left(x\right)\right) = \infty\right) \geq 1 - \alpha.$$

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

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

---

Analogous statement is also available for the classification framework.

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:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\text{mes}\left(\widehat{C}_\alpha(x)\right) = \infty\right) \geq 1 - \alpha.$$

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

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

$\hookrightarrow$ the lower bound is tight

Example (Naive estimator).

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

Analogous statement is also available for the classification framework.

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 $\mathscr{X} \subseteq \mathcal{X}$ such that $\mathbb{P}_{\mathcal{D}_X}(X \in \mathscr{X}) \geq \delta$, and for any associated exchangeable joint distribution $\mathcal{D}^{\mathrm{exch}(n+1)}$, we have:

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

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 $\mathscr{X} \subseteq \mathcal{X}$ 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:

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

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 practice, consider small δ → unefficient predictive sets.

Getting closer to X-conditional coverage

Getting closer to X-conditional coverage

- Approximate conditional coverage

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

Getting closer to X-conditional coverage

- Approximate conditional coverage

  $\hookrightarrow$ Romano et al. (2020); 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.

Split Conformal Prediction (SCP)

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

- On distribution-free X-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

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 $0 < \delta \leq 0.5$:

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

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

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

- Full Conformal Prediction
- Jackknife+

Beyond exchangeability

Some case studies

Concluding remarks

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

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

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

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

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

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

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

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

✗ $\hat{A}$ 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 CP (Vovk et al., 2005) does not discard training points!

- Full Conformal Prediction
    - avoids data splitting

Full CP (Vovk et al., 2005) does not discard training points!

- Full Conformal Prediction
    - avoids data splitting
    - at the cost of many more model fits

Full CP (Vovk et al., 2005) does not discard training points!

- Full 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.

Full 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 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
$$\mathcal{S}_y^{\text{(train)}} = \left\{ \mathsf{s}\left(X_i, Y_i; \hat{A}_y\right)\right\}_{i=1}^{n} \cup \left\{ \mathsf{s}\left(X_{n+1}, y; \hat{A}_y\right)\right\}$$
and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{\text{(train)}}\right)$

Full 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
$$\mathcal{S}_y^{(\text{train})} = \left\{ \text{s} \left(X_i, Y_i; \hat{A}_y \right) \right\}_{i=1}^{n} \cup \{ \text{s} \left(X_{n+1}, y; \hat{A}_y \right) \}$$
and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha} \left(\mathcal{S}_y^{(\text{train})} \right)$

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

Full 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
$$\mathcal{S}_y^{(\text{train})} = \left\{ \mathrm{s} \left(X_i, Y_i; \hat{A}_y \right) \right\}_{i=1}^{n} \cup \left\{ \mathrm{s} \left(X_{n+1}, y; \hat{A}_y \right) \right\}$$
and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha} \left(\mathcal{S}_y^{(\text{train})} \right)$

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

✓ Test point treated in the same way than train points

Full 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

$$\mathcal{S}_y^{(\text{train})} = \left\{ \text{s}\left(X_i, Y_i; \hat{A}_y\right)\right\}_{i=1}^{n} \cup \{ \text{s}\left(X_{n+1}, y; \hat{A}_y\right)\}$$

   and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

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

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

Full 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

$$\mathcal{S}_y^{\text{(train)}} = \left\{ \mathrm{s} \left(X_i, Y_i; \hat{A}_y \right) \right\}_{i=1}^{n} \cup \{ \mathrm{s} \left(X_{n+1}, y; \hat{A}_y \right) \}$$

   and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha} \left(\mathcal{S}_y^{\text{(train)}} \right)$

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

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

Full CP: theoretical foundation

Definition (Symmetrical algorithm)**.**

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

Full CP: theoretical foundation

Definition (Symmetrical algorithm).

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

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

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

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 $[\![1, n]\!]$: $\mathcal{A}\left(U_1, \dots, U_n\right) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \dots, U_{\sigma(n)}\right)$.

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

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

Moreover

$$Y_{n+1} \in \widehat{C}_\alpha^{\text{Full}}(X_{n+1}) := \left\{ y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}_y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)\right\}$$

$$\Leftrightarrow \mathrm{s}\left(X_{n+1}, Y_{n+1}; \hat{A}_{Y_{n+1}}\right) \leq q_{1-\alpha}\left(\mathcal{S}_{Y_{n+1}}^{(\text{train})}\right)$$

$$\Leftrightarrow S_{n+1} \leq q_{1-\alpha}(S_1, \dots, S_n, S_{n+1}) \text{ !}$$

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:

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

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

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 $k \in [\![1, n+1]\!]$

Interpolation regime

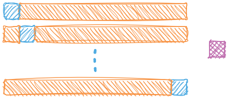
Example (FCP sets with an interpolating algorithm)**.**

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 $k \in [\![1, n+1]\!]$

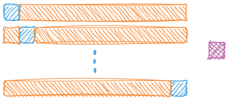
$\Rightarrow$ Full Conformal Prediction *(with standard score functions)* outputs $\mathcal{Y}$ (the whole label space) for any new test point!

Jackknife: the naive idea does not enjoy valid coverage



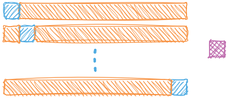
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$

Jackknife: the naive idea does not enjoy valid coverage



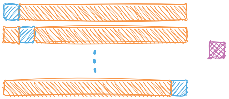
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean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

Jackknife: the naive idea does not enjoy valid coverage



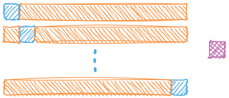
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard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

Jackknife: the naive idea does not enjoy valid coverage



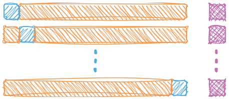
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean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

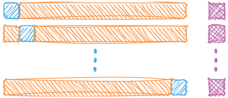
Jackknife+ (Barber et al., 2021b)

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 $\mathcal{D}_n \setminus (X_i, Y_i)$

Jackknife+ (Barber et al., 2021b)

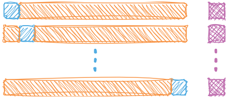


- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals

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

(in standard mean regression)

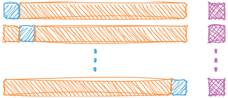
Jackknife+ (Barber et al., 2021b)



- 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 $q_{\beta,\text{inf}}(X_1, \ldots, X_k) := \lfloor \beta \times k \rfloor$ smallest value of $(X_1, \ldots, X_k)$

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)

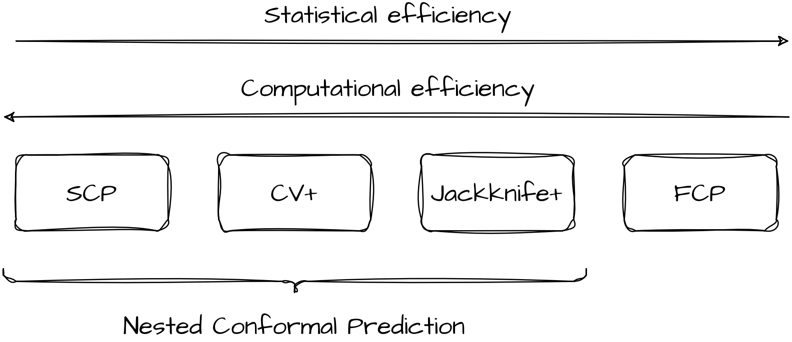
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

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

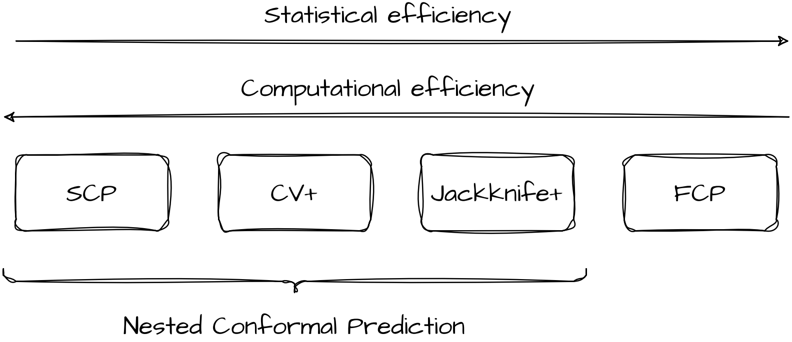
If $\mathcal{D}_n \cup (X_{n+1}, Y_{n+1})$ are exchangeable and $\mathcal{A}$ is symmetric: $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - 2\alpha$.

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

General overview

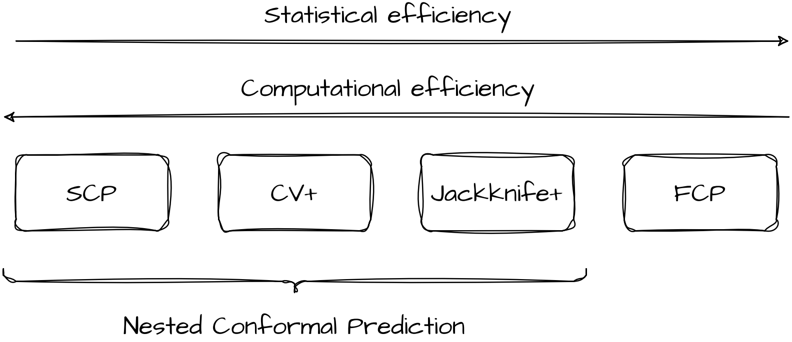


General overview



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

General overview



- Generalized framework encapsulating out-of-sample methods: Nested CP (Gupta et al., 2022) → 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.

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)

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

Covariate shift (Tibshirani et al., 2019)

- 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

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

Covariate shift (Tibshirani et al., 2019)

- 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 $w(X_i) = \dfrac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

Covariate shift (Tibshirani et al., 2019)

- 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. $\omega_i = \omega(X_i) = \dfrac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

Covariate shift (Tibshirani et al., 2019)

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

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

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

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)
      ↪ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓
    - Barber et al. (2022)
      ↪ Quantifies the coverage loss depending on the strength of exchangeability violation
      $$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha - \text{average violation of exchangeability by each calibration point}$$
      ↪ 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:

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

often relaxed in:

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

Online setting

- Data: T_0 random variables $(X_1, Y_1), \dots, (X_{T_0}, Y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$
- Aim: predict the response values as well as predictive intervals for T_1 subsequent observations $X_{T_0+1}, \dots, X_{T_0+T_1}$ sequentially: at any prediction step $t \in [\![T_0 + 1, T_0 + T_1]\!]$, $Y_{t-T_0}, \dots, Y_{t-1}$ 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:
$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{Y_t \in \widehat{C}_{\alpha}^{t}(X_t)\right\} \approx 1 - \alpha.$$

$\rightsquigarrow$ More during the case study!

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

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

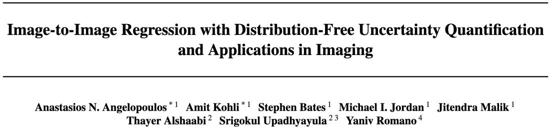


Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

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 ⇝ 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.

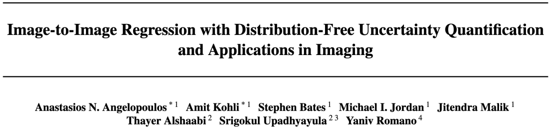


Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

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]

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

- Medical application
- Image based task
- Pixel by pixel analysis ⇝ 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.

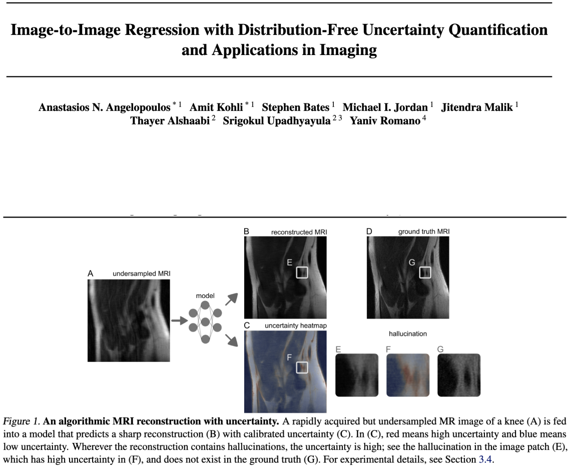


Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

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 1: 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 - $\hat{\mu} : \mathbb{R}^{NM} \rightarrow [0; 1]$

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 - $\hat{\mu} : \mathbb{R}^{NM} \rightarrow [0; 1]$
    - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \rightarrow [0; 1]$, such that

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

$\rightarrow$ 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

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

      $\rightarrow$ 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

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

   → 3 regressors are used

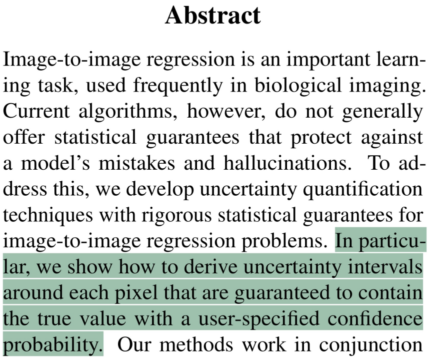
   4 techniques are experimented for these regressors, including QR.
3. Calibration step: leverage the calibration set.
   - In spirit, almost equivalent to CQR but with a multiplicative form.
   - Precisely, relies on RCPS (Bates et al., 2021a)

Guarantee:

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

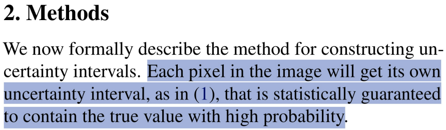
→ Marginal validity on the test, with high probability w.r.t. the calibration set.

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

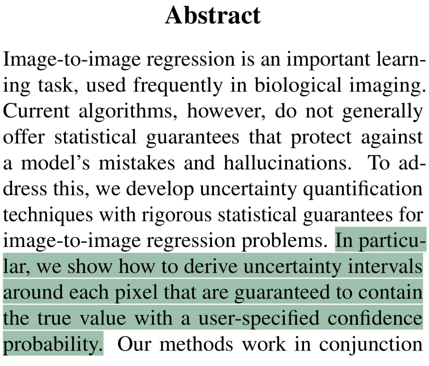


2. Methods

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.

How do you understand that?

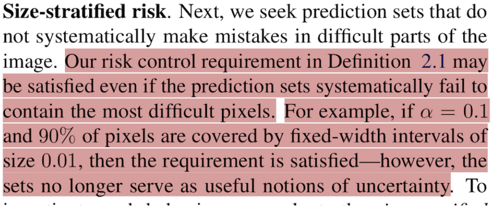
Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)



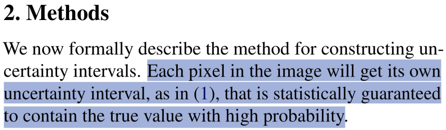
Abstract

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

- 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



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.

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

How do you understand that?

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 ⟶

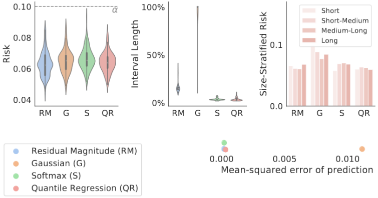
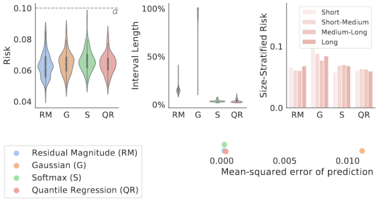


Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

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

- Link between the size of the PI and the coverage level ⟶
- Localization of the errors ↓



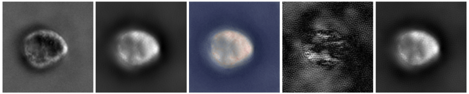


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth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

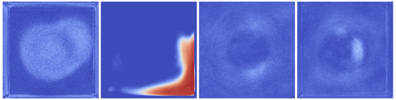


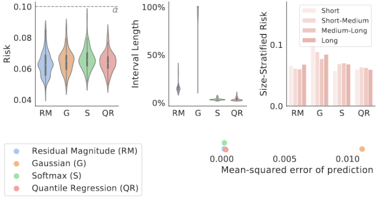
Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

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

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 ⟶
- Localization of the errors ↓



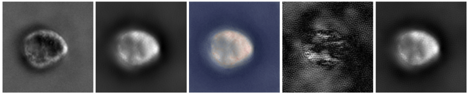


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth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

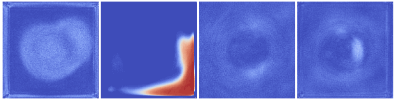


Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

Figure 2: 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 → look at metrics to evaluate which methods performs best!

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

- Healthcare
- Electricity

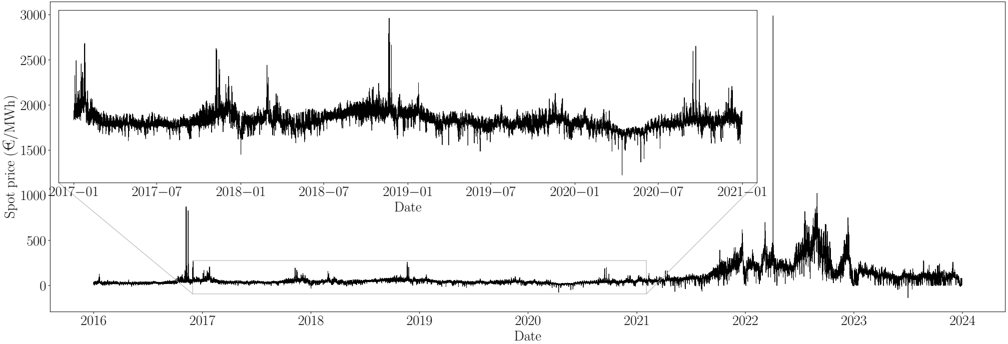
Concluding remarks

Forecasting French spot electricity prices

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

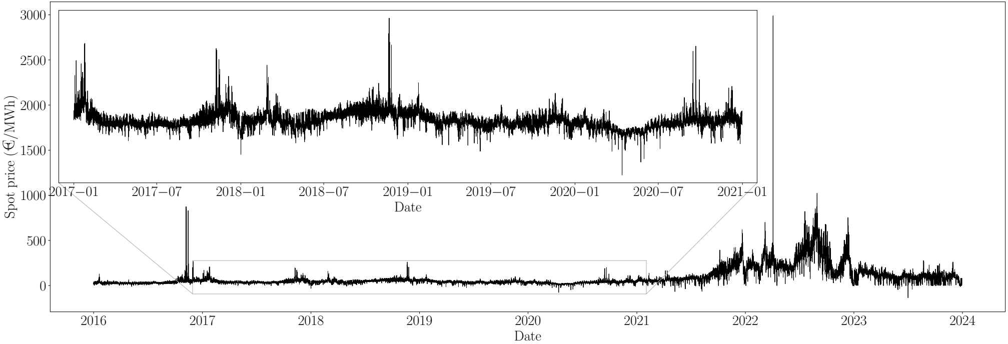
Forecasting French spot electricity prices

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



Forecasting French spot electricity prices

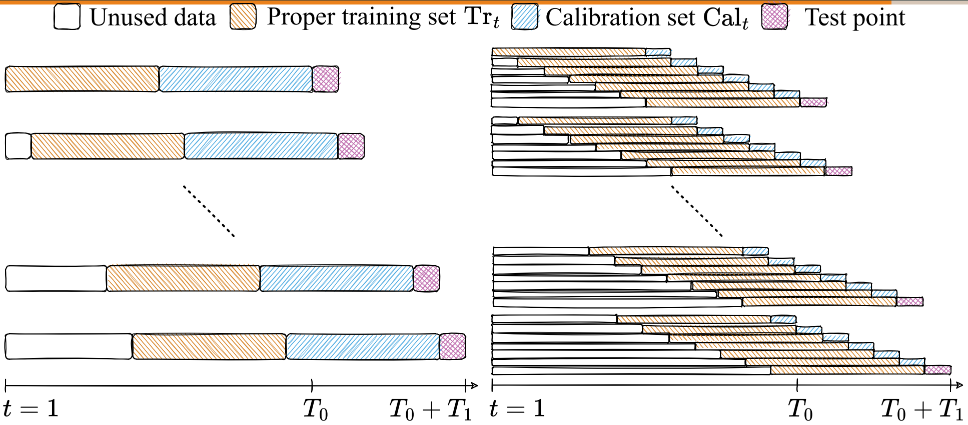
Hourly day-ahead market prices (between producers and suppliers)



To which extent are they forecastable?

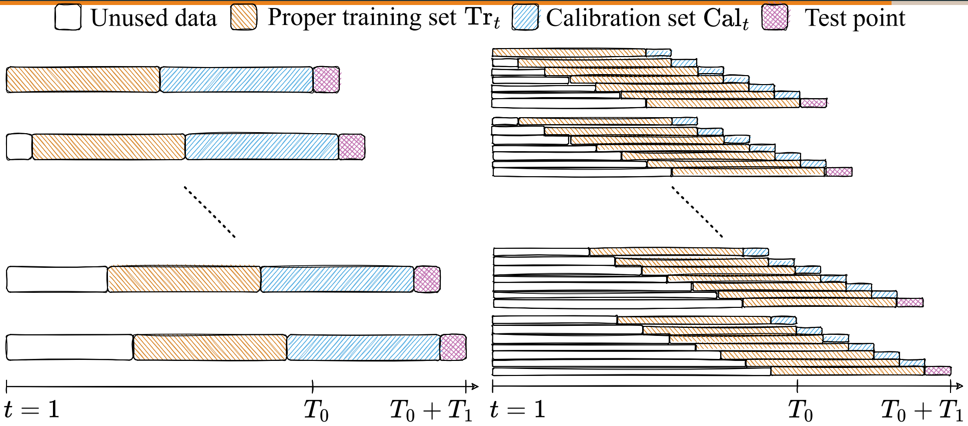
↪ forecasts errors **no lower than 10%** of the realized price!

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



(a) OSSCP (b) OSSCP-horizon

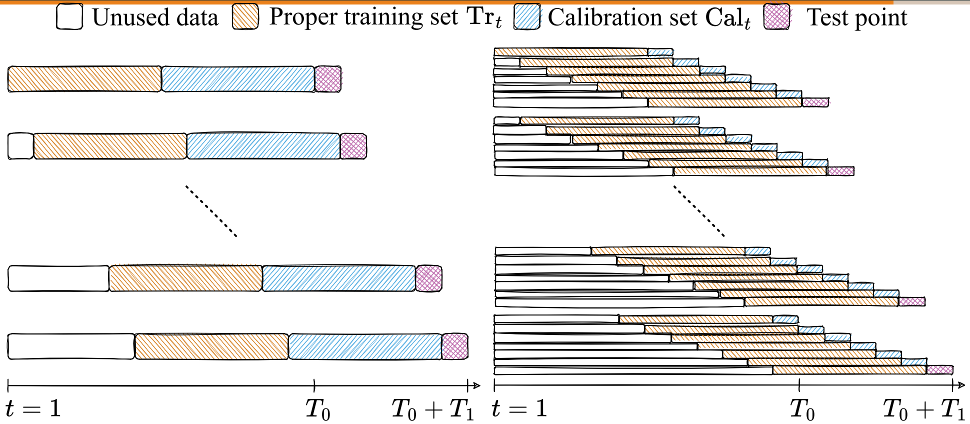
Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

↪ `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

↪ `OSSCP` improves robustness in temporal settings;

↪ `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 $\alpha_1 = \alpha$, $\gamma \geq 0$.

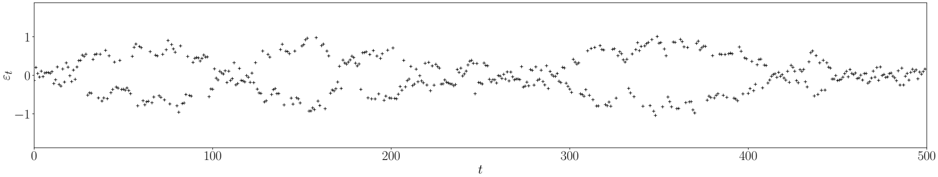
Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

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

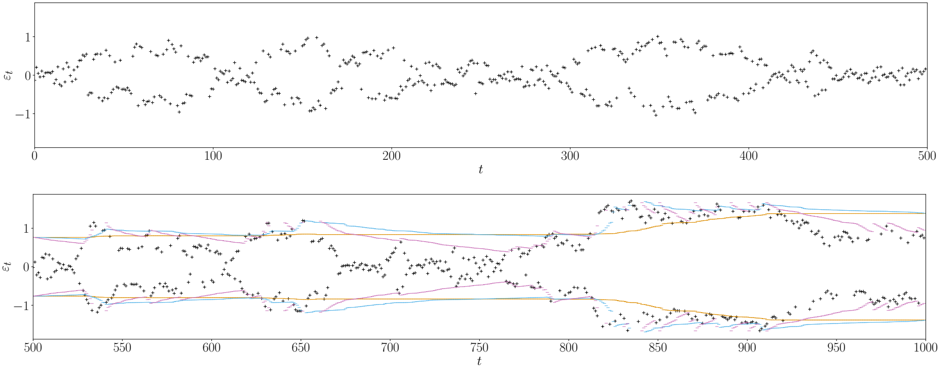
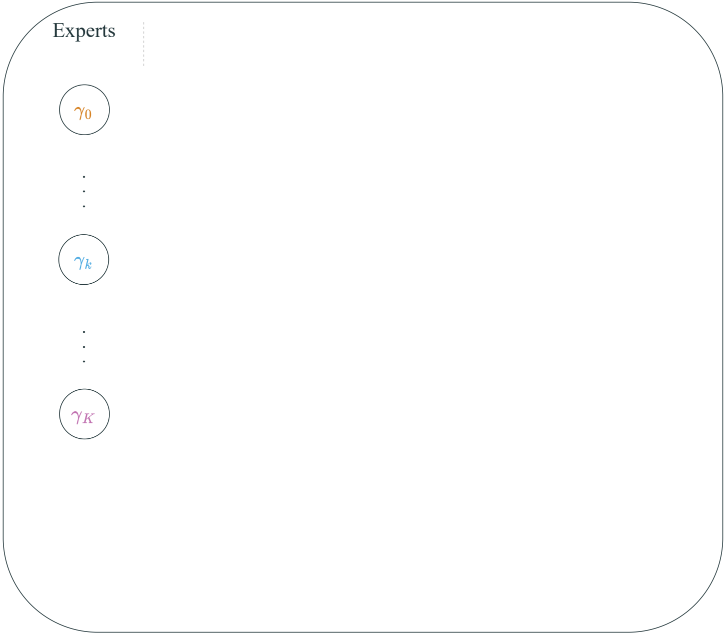


Figure 3: Visualisation of ACI with different values of γ ($\gamma = 0$, $\gamma = 0.01$, $\gamma = 0.05$)

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



Experts

γ_0

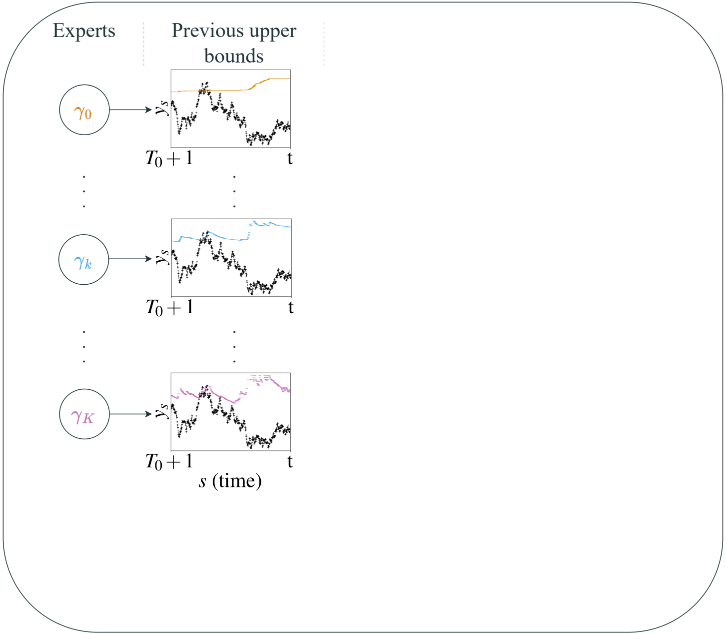
$\vdots$

γ_k

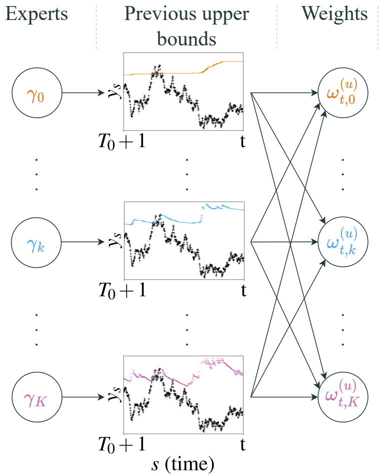
$\vdots$

γ_K

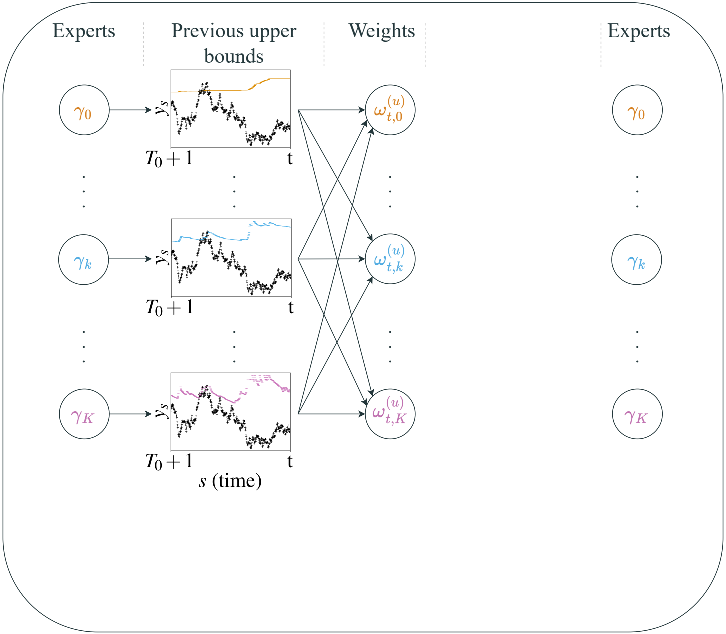
`AgACI`: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



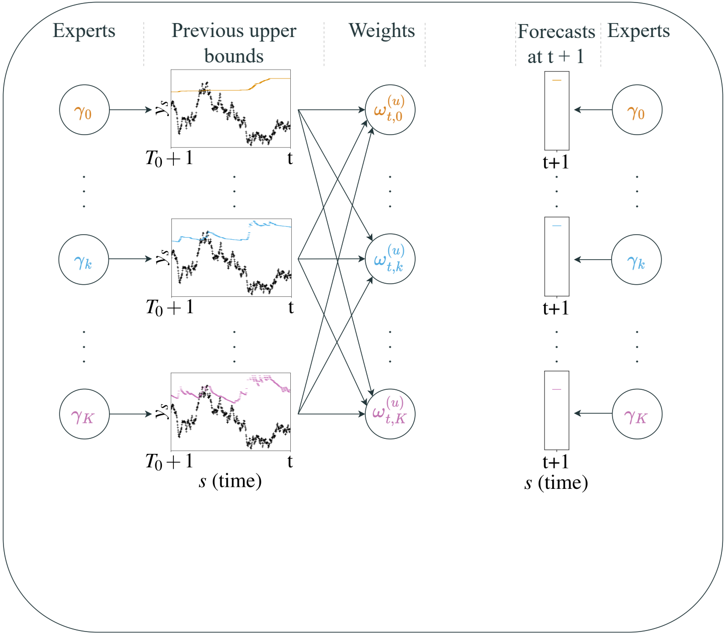
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



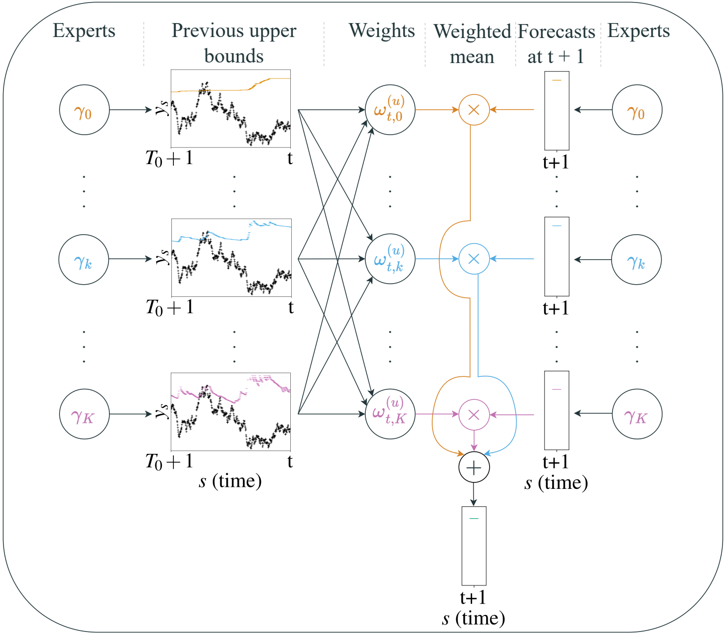
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



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

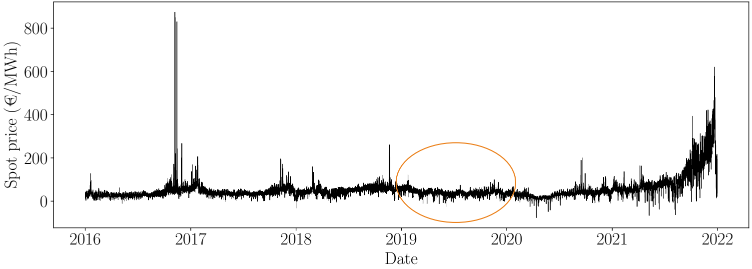


`AgACI`: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



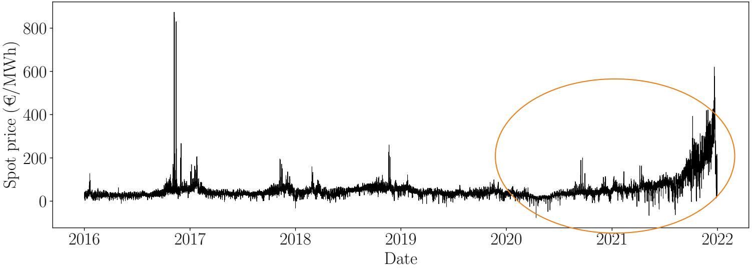
Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- 2019: `AgACI` provides validity with a reasonable efficiency;



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

- 2019: `AgACI` provides validity with a reasonable efficiency;
- 2020 and 2021: `AgACI` fails to ensure validity, and the various forecasting models considered[2] behave differently.



---

[2] Quantile Random Forests, Quantile Generalized Additive Models, Quantile Gradient Boosting, etc.

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)

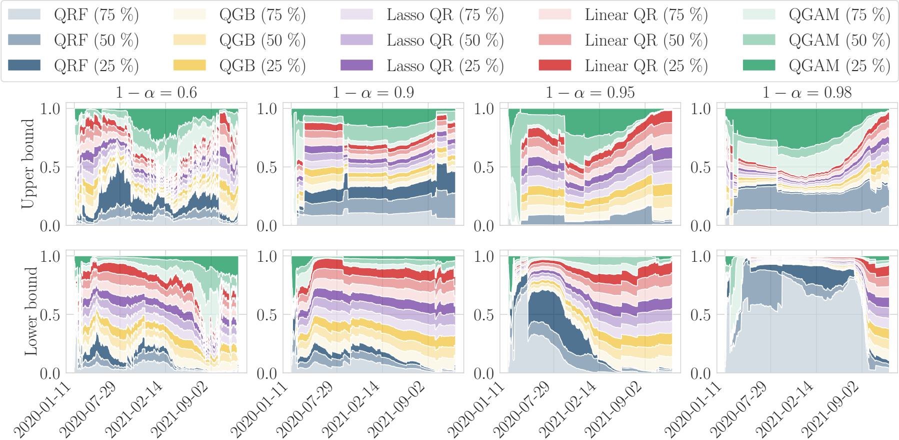
Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ 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

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)

Highlights and perspectives

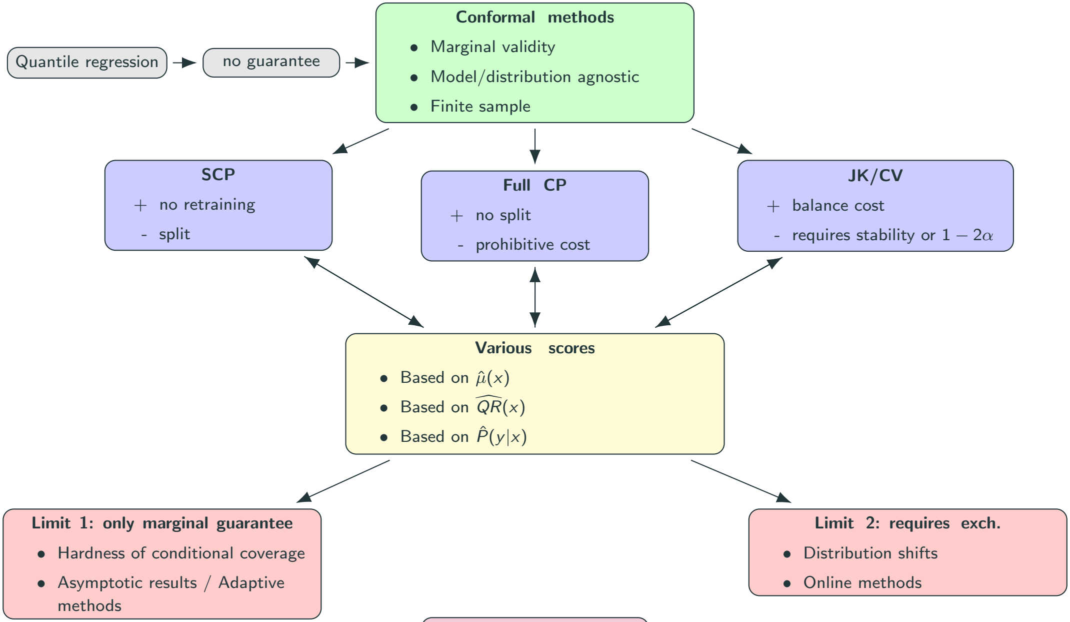
Aggregating the two bounds independently (as in `AgACI` and beyond):

✓ 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)

↪ Weaken the objective and consider a more practical theoretical aim?

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?

References i

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).

References ii

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).

References iii

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.

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.

References v

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.

References vi

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. (2020). 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.

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.

References x

Vovk, V. (2012). Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

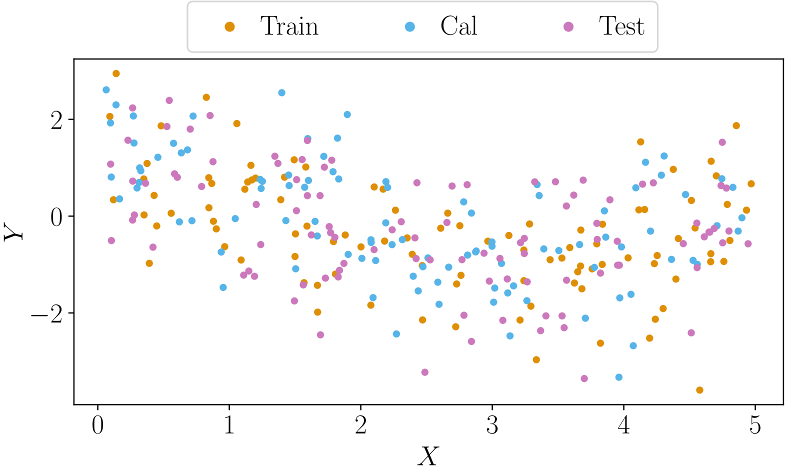
Vovk, V., Nouretdinov, I., and Gammerman, A. (2003). Testing exchangeability on-line. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 768–775. AAAI Press.

Zaffran, M., 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.

SCP

CQR

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



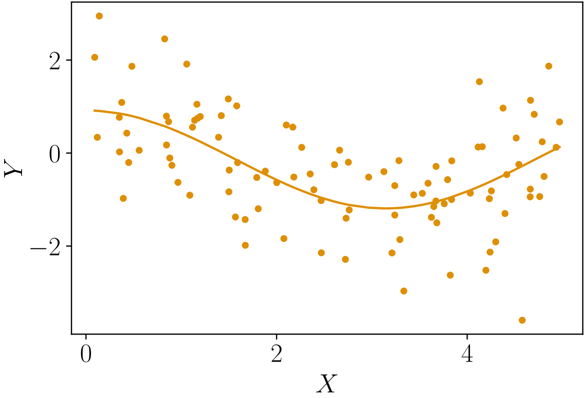
---

[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) $\hat{\mu}$

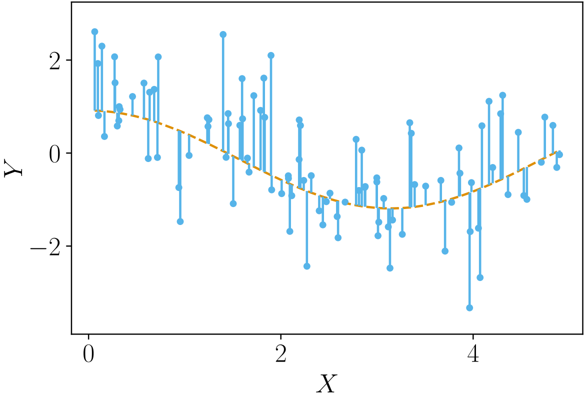
---

[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

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



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

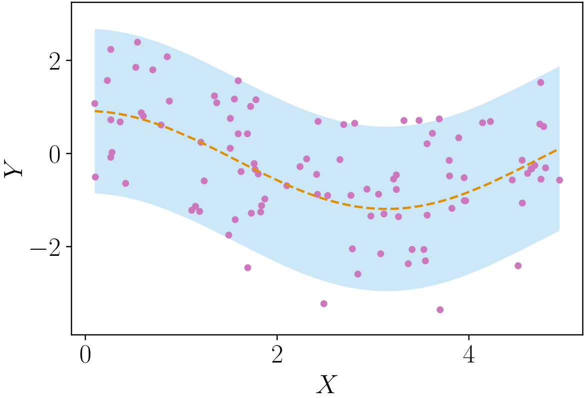
---

[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

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



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

▸ Back to SCP

---

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

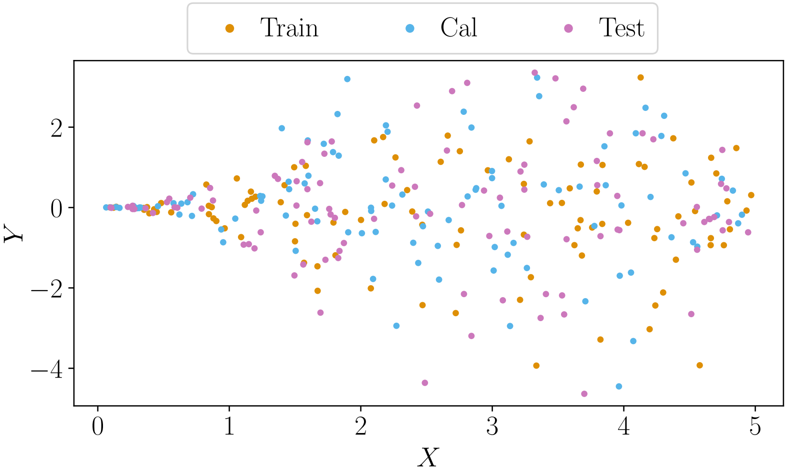
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML

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

SCP

CQR

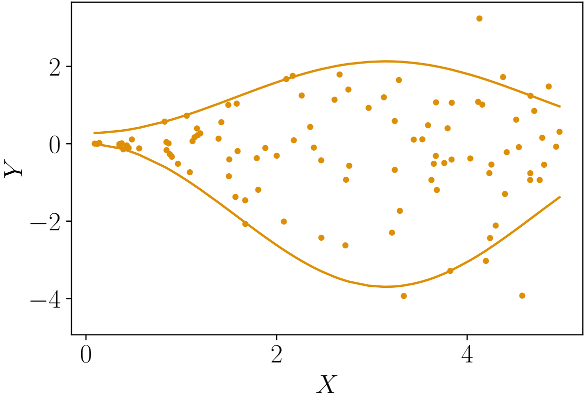
Conformalized Quantile Regression (CQR)[5]



---

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

Conformalized Quantile Regression (CQR)[5] training step

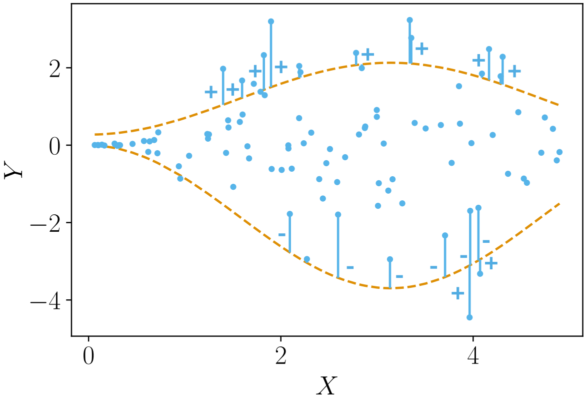


- Learn (or get) $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$

---

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

Conformalized Quantile Regression (CQR)[5] calibration step



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

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

‣ Back to Generalization SCP

---

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

Label shift (Podkopaev and Ramdas, 2021)

- 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

Label shift (Podkopaev and Ramdas, 2021)

- 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

Label shift (Podkopaev and Ramdas, 2021)

- 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

Label shift (Podkopaev and Ramdas, 2021)

- **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

Label shift (Podkopaev and Ramdas, 2021)

- 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
  2. normalize the weights, i.e. $\omega_i^y = \omega^y(X_i) = \dfrac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}$

Label shift (Podkopaev and Ramdas, 2021)

- 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 $w(Y_i) = \frac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ 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 $\widehat{C}_\alpha(X_{n+1}) =$

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