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

Lecture on Conformal Prediction

Margaux Zaffran

December 1-5, 2025

ECAS-SFdS School

Why a lecture on conformal prediction?

- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
  - 1996-1999: emergence of Conformal Prediction (CP) in the historical machine learning community, by Vladimir Vovk, Alexander Gammerman, Vladimir Vapnik, Glenn Shafer
  - 2012+: Jing Lei, Larry Wasserman and colleagues at Carnegie Mellon University (CMU) bring it to the statistics community
  - 2020+: explosion, notably thanks to Rina Barber, Emmanuel Candès, Aaditya Ramdas and Ryan J. Tibshirani

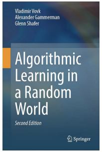
Why a lecture on conformal prediction?

- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
- Because I believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.

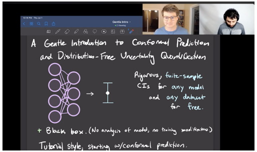
Why a lecture on conformal prediction?

- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
- Because I believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.
  Successfully applied to
  - Medical applications
  - Markets / demand forecasting
  - Computer Vision

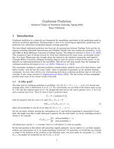
- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
- Because I believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.
- To be part of the **diffusion** effort that many colleagues are making.



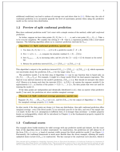
Book reference: Vovk et al. (2005)
(new edition in 2022)



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



R. J. Tibshirani
introductive lecture's notes



Forthcoming book: Angelopoulos, Barber, Bates (2024)
arXiv version

Why a lecture on conformal prediction?

- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
- Because I 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.

$\hookrightarrow$ Based upon earlier material accessible on `this webpage`

Why a lecture on conformal prediction?

- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
- Because I 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.

↪ Based upon earlier material accessible on `this webpage`

↪ Presented as lectures and tutorials in various places (ENBIS conference, UAI and ICML tutorials, etc.)

Why a lecture on conformal prediction?

- Conformal Prediction comes from the 90s, and has emerged as a popular topic in the early 2020's.
- Because I 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.

↪ Based upon earlier material accessible on `this webpage`

↪ Presented as lectures and tutorials in various places (ENBIS conference, UAI and ICML tutorials, etc.)

Aymeric Dieuleveut

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

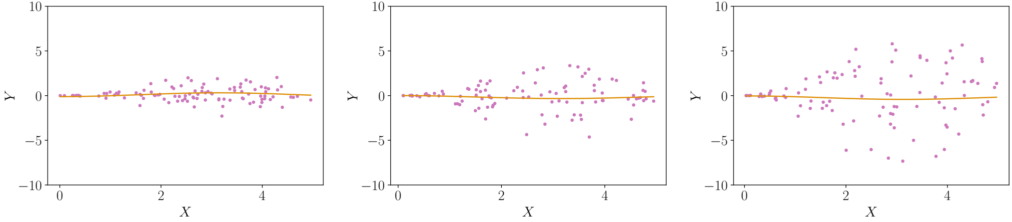
- Many people contributed to the domain - list of references may not be exhaustive
- Multiple other excellent resources

On the importance of quantifying uncertainty

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

On the importance of quantifying uncertainty

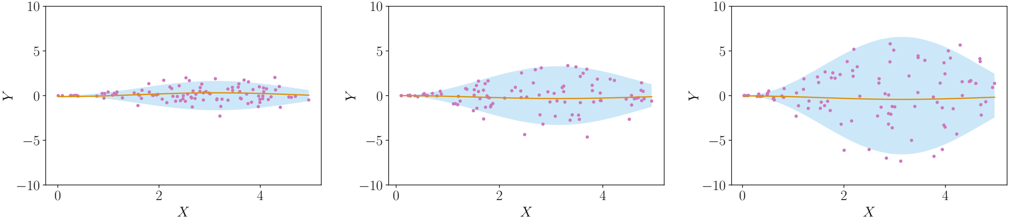
- **Obvious in most applications - weather, medical, markets**
- **Mathematically**



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

On the importance of quantifying uncertainty

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

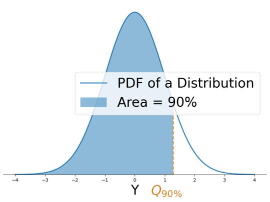


↪ Same "best" predictor, yet 3 distinct underlying phenomena!

$\Longrightarrow$ Quantifying uncertainty conveys this information.

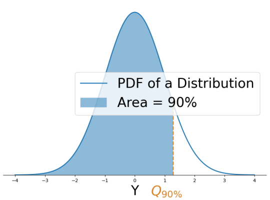
Reminder about quantiles

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

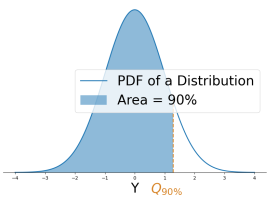


Reminder about quantiles

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



Reminder about quantiles



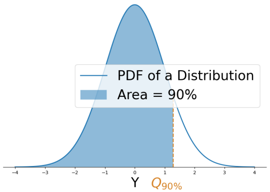
- Quantile level $\beta \in [0, 1]$
- $Q_Y(\beta) := \inf\{t \in \mathbb{R}, \mathbb{P}(Y \leq t) \geq \beta\}$
- Empirical quantile $q_\beta(Y_1, \ldots, Y_n)$
  $:= \lceil \beta \times n \rceil$ smallest value of $(Y_1, \ldots, Y_n)$

$\hookrightarrow$ Defining the order statistics of a random vector $(Y_1, \ldots, Y_n)$ by $(Y_{(1)}, \ldots, Y_{(n)})$, with $Y_{(1)} \leq \ldots \leq Y_{(n)}$, we have:

$$q_\beta(Y_1, \ldots, Y_n) = Y_{(\lceil \beta \times n \rceil)}$$

.

Reminder about quantiles



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

$\hookrightarrow$ Defining the order statistics of a random vector $(Y_1, \ldots, Y_n)$ by $(Y_{(1)}, \ldots, Y_{(n)})$, with $Y_{(1)} \leq \ldots \leq Y_{(n)}$, we have:

$$q_\beta(Y_1, \ldots, Y_n) = Y_{(\lceil \beta \times n \rceil)}$$

Example (a special quantile: the median).

$\beta = 0.5$

$\hookrightarrow$ $q_{0.5}(Y_1, \ldots, Y_n)$ is the empirical median of $(Y_1, \ldots, Y_n)$;

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

$\hookrightarrow$ $Q_Y(0.5) = \arg\min_c \mathbb{E}[|Y - c|]$.

Median regression

- The Bayes predictor depends on the chosen loss function.

$\hookrightarrow$ Bayes predictor $f^\star \in \underset{f}{\operatorname{argmin}} \mathsf{Risk}_\ell(f)$

$$:= \underset{f}{\operatorname{argmin}} \mathbb{E}\left[\ell(Y, f(X))\right]$$

Median regression

- The Bayes predictor depends on the chosen loss function.
  $\hookrightarrow$ Bayes predictor $f^\star \in \underset{f}{\operatorname{argmin}} \operatorname{Risk}_\ell(f)$

$$:= \underset{f}{\operatorname{argmin}} \mathbb{E}\left[\ell(Y, f(X))\right]$$

- Mean Absolute Error (MAE): $\ell(Y, Y') = |Y - Y'|$
  Associated risk: $\operatorname{Risk}_\ell(f) = \mathbb{E}\left[|Y - f(X)|\right]$

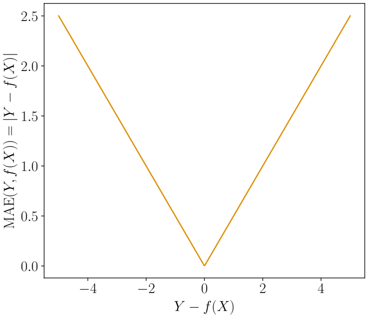
Median regression

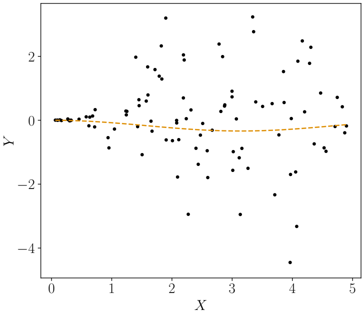
- The Bayes predictor depends on the chosen loss function.
  $\hookrightarrow$ Bayes predictor $f^\star \in \underset{f}{\operatorname{argmin}} \mathrm{Risk}_\ell(f)$

$$:= \underset{f}{\operatorname{argmin}} \mathbb{E}\left[\ell(Y, f(X))\right]$$

- Mean Absolute Error (MAE): $\ell(Y, Y') = |Y - Y'|$
  Associated risk: $\mathrm{Risk}_\ell(f) = \mathbb{E}\left[|Y - f(X)|\right]$

$$\Rightarrow f^\star(X) = \mathrm{median}\left[Y|X\right] = Q_{Y|X}(0.5)$$





Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta |Y - Y'| \mathbb{1}_{\{Y - Y' \geq 0\}} + (1 - \beta) |Y - Y'| \mathbb{1}_{\{Y - Y' \leq 0\}}$
  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y - Y' \geq 0\}} + (1 - \beta)|Y - Y'|\mathbb{1}_{\{Y - Y' \leq 0\}}$
  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

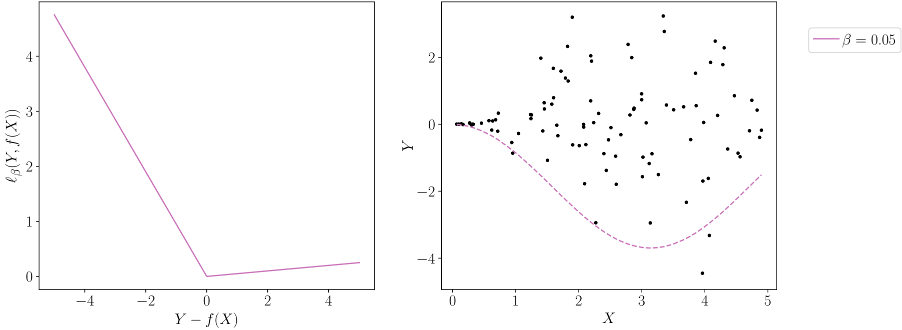
$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$

Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta |Y - Y'| \mathbb{1}_{\{Y - Y' \geq 0\}} + (1 - \beta) |Y - Y'| \mathbb{1}_{\{Y - Y' \leq 0\}}$

  Bayes predictor: $f^\star \in \underset{f}{\operatorname{argmin}} \, \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

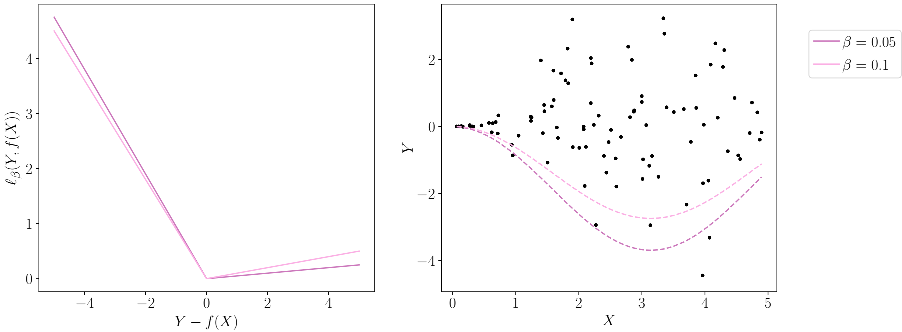
$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$



- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y-Y' \geq 0\}} + (1-\beta)|Y - Y'|\mathbb{1}_{\{Y-Y' \leq 0\}}$

  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

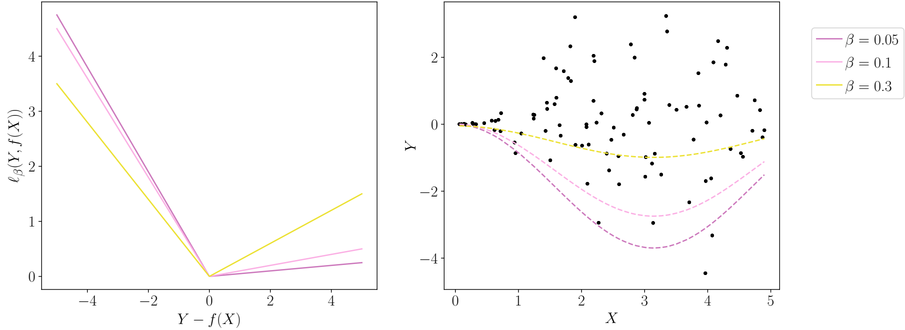
$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$



- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y - Y' \geq 0\}} + (1 - \beta)|Y - Y'|\mathbb{1}_{\{Y - Y' \leq 0\}}$

  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$

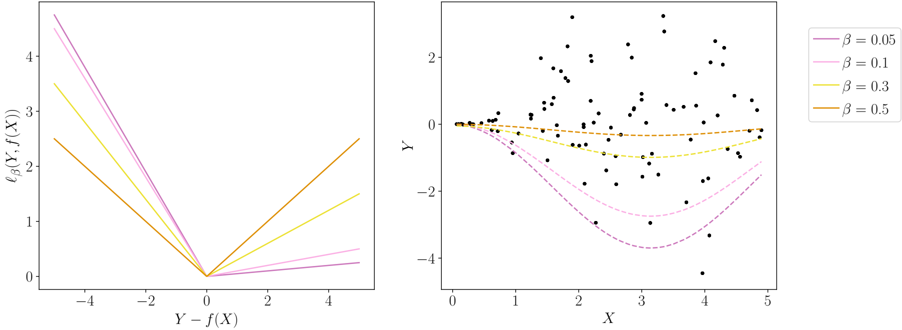


Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta |Y - Y'| \mathbb{1}_{\{Y - Y' \geq 0\}} + (1 - \beta) |Y - Y'| \mathbb{1}_{\{Y - Y' \leq 0\}}$

  Bayes predictor: $f^\star \in \underset{f}{\operatorname{argmin}}\, \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

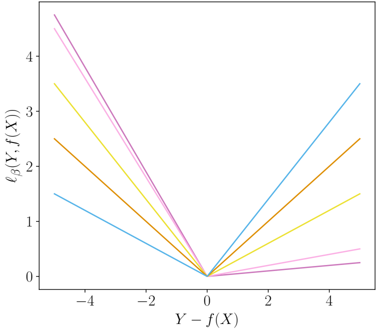
$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$

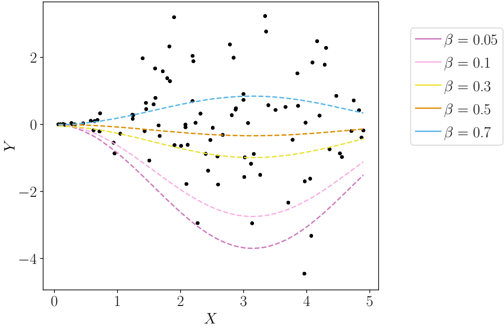


- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y-Y'\geq 0\}} + (1-\beta)|Y - Y'|\mathbb{1}_{\{Y-Y'\leq 0\}}$

  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$



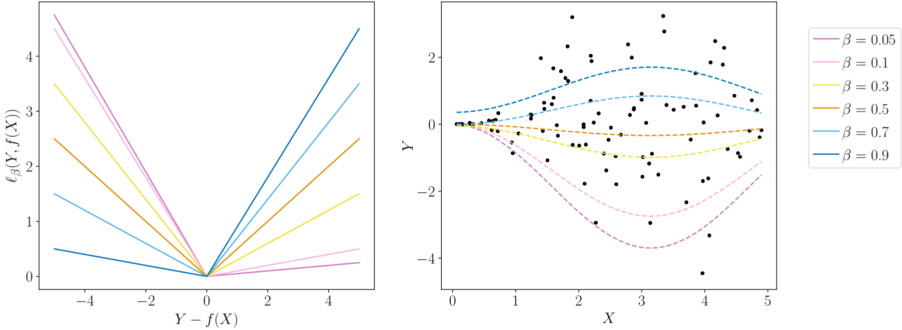


Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y-Y'\geq 0\}} + (1-\beta)|Y - Y'|\mathbb{1}_{\{Y-Y'\leq 0\}}$

  Bayes predictor: $f^\star \in \underset{f}{\mathrm{argmin}}\, \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$

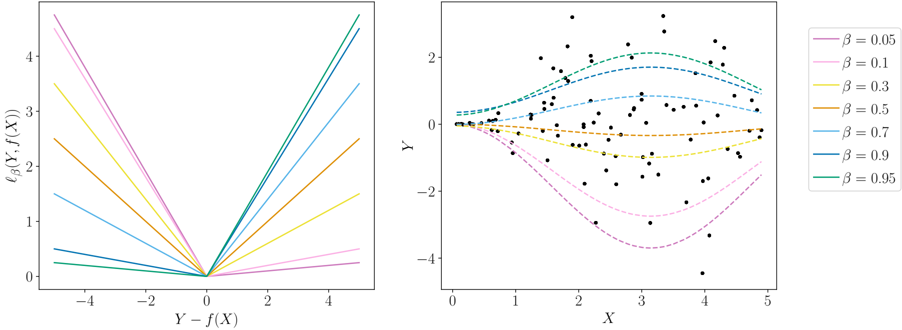


Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y - Y' \geq 0\}} + (1 - \beta)|Y - Y'|\mathbb{1}_{\{Y - Y' \leq 0\}}$

  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$

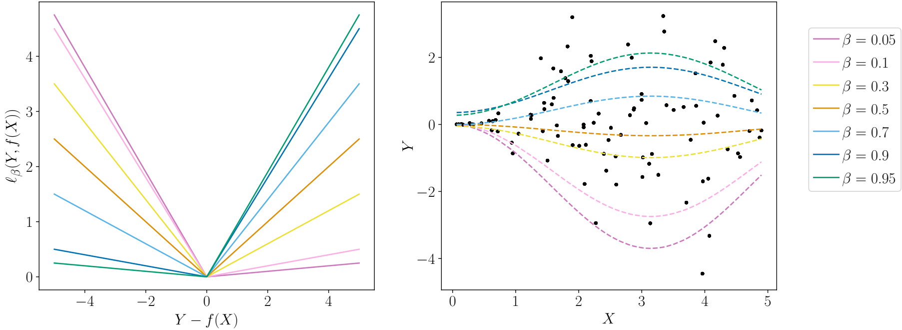


Generalization: Quantile regression

- Quantile level $\beta \in [0, 1]$
- Pinball loss: $\ell_\beta(Y, Y') = \beta|Y - Y'|\mathbb{1}_{\{Y-Y'\geq 0\}} + (1-\beta)|Y - Y'|\mathbb{1}_{\{Y-Y'\leq 0\}}$

  Bayes predictor: $f^\star \in \operatorname*{argmin}_f \mathbb{E}\left[\ell_\beta(Y, f(X))\right]$

$$\Rightarrow f^\star(X) = Q_{Y|X}(\beta)$$



Quantifying predictive uncertainty

- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ 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

- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ 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

→ we do not want to blindly trust a model's assumption;

→ we do not want to test if the data's assumptions are correct;

→ yet, we do want to have theoretically valid tools.

Quantifying predictive uncertainty

- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ 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

→ we do not want to blindly trust a model's assumption;

→ we do not want to test if the data's assumptions are correct;

→ yet, we do want to have theoretically valid tools.

Post-hoc design (black-box view)

We aim for an approach that works in a post-hoc fashion on top of *any* $\mathcal{A}$, with no impact of the choice of $\mathcal{A}$ on the validity of the method.

- $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ 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

Attempt to define distribution-free validity

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

Definition ((lofty) Distribution-free validity).

$\widehat{C}_\alpha$ achieves (lofty) distribution-free validity if for any joint distribution $\mathcal{D}_{n+1}$ on $(\mathcal{X} \times \mathcal{Y})^{n+1}$, we have that:

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

Attempt to define distribution-free validity

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

Definition ((lofty) Distribution-free validity).

$\widehat{C}_\alpha$ achieves (lofty) distribution-free validity if for any joint distribution $\mathcal{D}_{n+1}$ on $(\mathcal{X} \times \mathcal{Y})^{n+1}$, we have that:

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

✗ Wishing to satisfy this definition without any constraint on $\mathcal{D}_{n+1}$ is way too ambitious!

Attempt to define distribution-free validity

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

Definition ((lofty) Distribution-free validity).

$\widehat{C}_\alpha$ achieves (lofty) distribution-free validity if for any joint distribution $\mathcal{D}_{n+1}$ on $(\mathcal{X} \times \mathcal{Y})^{n+1}$, we have that:

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

- ✗ Wishing to satisfy this definition without any constraint on $\mathcal{D}_{n+1}$ is way too ambitious!
- ✓ What is achievable is to restrict $\mathcal{D}_{n+1}$ to any product distribution $\mathcal{D}^{\otimes(n+1)}$ (i.i.d.) without restriction on $\mathcal{D}$ (distribution on $(\mathcal{X} \times \mathcal{Y})$).

Attempt to define distribution-free validity

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

Definition ((lofty) Distribution-free validity).

$\widehat{C}_\alpha$ achieves (lofty) distribution-free validity if for any joint distribution $\mathcal{D}_{n+1}$ on $(\mathcal{X} \times \mathcal{Y})^{n+1}$, we have that:

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

- ✗ Wishing to satisfy this definition without any constraint on $\mathcal{D}_{n+1}$ is way too ambitious!
- ✓ What is achievable is to restrict $\mathcal{D}_{n+1}$ to any product distribution $\mathcal{D}^{\otimes(n+1)}$ (i.i.d.) without restriction on $\mathcal{D}$ (distribution on $(\mathcal{X} \times \mathcal{Y})$).
- ❣ In fact, we can even weaken this to exchangeable distributions.

This is what we are jumping into!

1. On exchangeability

This is what we are jumping into!

1. On exchangeability
2. Split conformal prediction

This is what we are jumping into!

1. On exchangeability
2. Split conformal prediction
3. Towards conditional coverage?

This is what we are jumping into!

1. On exchangeability
2. Split conformal prediction
3. Towards conditional coverage?
4. Beyond exchangeability

This is what we are jumping into!

1. On exchangeability
2. Split conformal prediction
3. Towards conditional coverage?
4. Beyond exchangeability
5. Computational and statistical trade-offs

This is what we are jumping into!

1. On exchangeability
2. Split conformal prediction
3. Towards conditional coverage?
4. Beyond exchangeability
5. Computational and statistical trade-offs
6. Handling missing data

This is what we are jumping into!

1. On exchangeability (theory)
2. Split conformal prediction (theory)
3. Towards conditional coverage? (theory)
4. Beyond exchangeability
5. Computational and statistical trade-offs (theory)
6. Handling missing data

This is what we are jumping into!

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory)
3. Towards conditional coverage? (theory)
4. Beyond exchangeability (methods)
5. Computational and statistical trade-offs (methods) (theory)
6. Handling missing data (methods)

This is what we are jumping into!

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory) (practical session)
3. Towards conditional coverage? (practical session) (theory)
4. Beyond exchangeability (methods)
5. Computational and statistical trade-offs (methods) (theory)
6. Handling missing data (methods)

This is what we are jumping into!

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory) (practical session)
3. Towards conditional coverage? (practical session) (theory) (case studies)
4. Beyond exchangeability (methods) (case studies)
5. Computational and statistical trade-offs (methods) (theory)
6. Handling missing data (methods)

This is what we are jumping into!

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory)

Introductory remarks

Exchangeability

Split Conformal Prediction (SCP)

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

Exchangeable random vectors

Definition (Exchangeability)**.**

$(Z_i)_{i=1}^n$, a random vector taking values in $\mathcal{Z}^n$, is exchangeable if, for any permutation σ of $[\![1, n]\!]$:

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

Definition (Exchangeability).

$(Z_i)_{i=1}^n$, a random vector taking values in $\mathcal{Z}^n$, is exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$(Z_1, \ldots, Z_n) \stackrel{d}{=} \left(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}\right) .$$

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

Exchangeable random vectors

Definition (Exchangeability).

$(Z_i)_{i=1}^n$, a random vector taking values in $\mathcal{Z}^n$, is exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$(Z_1, \dots, Z_n) \overset{d}{=} \left(Z_{\sigma(1)}, \dots, Z_{\sigma(n)}\right) .$$

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

In particular, exchangeability implies that all the entries of $(Z_i)_{i=1}^n$ are identically distributed.

Some examples of exchangeable random vectors

Example (exchangeable sequences).

- $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{D}$

Some examples of exchangeable random vectors

Example (exchangeable sequences).

- $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{D}$
- The components of $\mathcal{N}\left(\begin{pmatrix} m \\ \vdots \\ \vdots \\ m \end{pmatrix}, \begin{pmatrix} \sigma^2 & & & \\ & \ddots & \gamma^2 & \\ & \gamma^2 & \ddots & \\ & & & \sigma^2 \end{pmatrix}\right)$

Some examples of exchangeable random vectors

Example (exchangeable sequences).

- $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \mathcal{D}$
- The components of $\mathcal{N}\left(\begin{pmatrix} m \\ \vdots \\ \vdots \\ m \end{pmatrix}, \begin{pmatrix} \sigma^2 & & & \\ & \ddots & \gamma^2 & \\ & \gamma^2 & \ddots & \\ & & & \sigma^2 \end{pmatrix} \right)$
- Sampling without replacement from $\{z_1, \ldots, z_n\}$

Other characterizations of exchangeability

Characterization in a countable space

Denote p the probability mass function of the joint distribution of $(Z_1, \ldots, Z_n)$.
$(Z_1, \ldots, Z_n)$ are exchangeable if and only if, for any permutation σ of $[\![1, n]\!]$, for any $(z_1, \ldots, z_n) \in \mathcal{Z}^n$ we have:

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

Other characterizations of exchangeability

Characterization in a countable space

Denote p the probability mass function of the joint distribution of $(Z_1, \dots, Z_n)$.
$(Z_1, \dots, Z_n)$ are exchangeable if and only if, for any permutation σ of $[\![1, n]\!]$, for any $(z_1, \dots, z_n) \in \mathcal{Z}^n$ we have:

$$p\left(z_1, \dots, z_n\right) = p\left(z_{\sigma(1)}, \dots, z_{\sigma(n)}\right).$$

Characterization in a continuous space

Assume $(Z_1, \dots, Z_n)$ admits a probability density function, denoted f.
$(Z_1, \dots, Z_n)$ are exchangeable if and only if, for any permutation σ of $[\![1, n]\!]$, for almost any $(z_1, \dots, z_n) \in \mathcal{Z}^n$ we have:

$$f\left(z_1, \dots, z_n\right) = f\left(z_{\sigma(1)}, \dots, z_{\sigma(n)}\right).$$

Interesting properties of exchangeability

Blackboard time!

Achievable definition of distribution-free validity

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

Definition (Distribution-free validity)**.**

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

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

we have that:

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

Introductory remarks

Exchangeability

Split Conformal Prediction (SCP)

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

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

Split Conformal Prediction (SCP) (Vovk et al., 2005)

▸ without anim

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

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*

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

without anim V0 with $+\infty$

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

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

(+ worst-case scenario)

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

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

($+$ worst-case scenario)

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

SCP: implementation details

without anim V0 with $+\infty$

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

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

(+ worst-case scenario)

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

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

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

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

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

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

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

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

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

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

SCP: theoretical 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:

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

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

Blackboard time!

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[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.

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

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

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

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

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

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

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

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

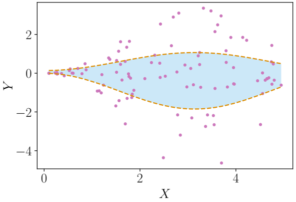
Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

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

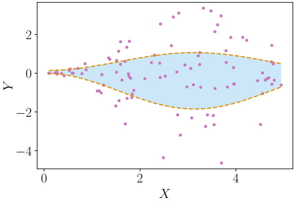
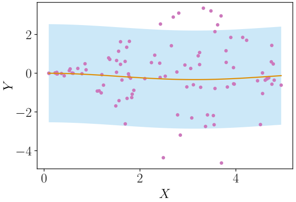
[a] Only the calibration and test data need to be exchangeable.

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

Conditional coverage implies adaptiveness

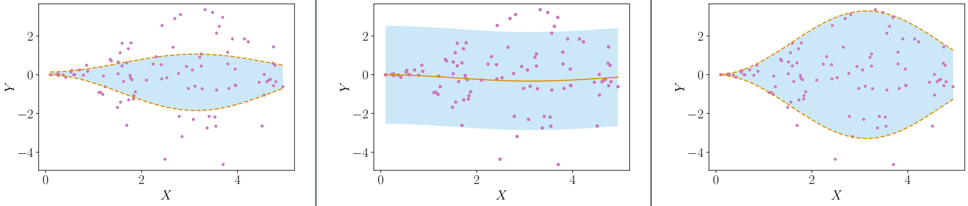
no coverage		
		

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

no coverage	marginal	
		

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

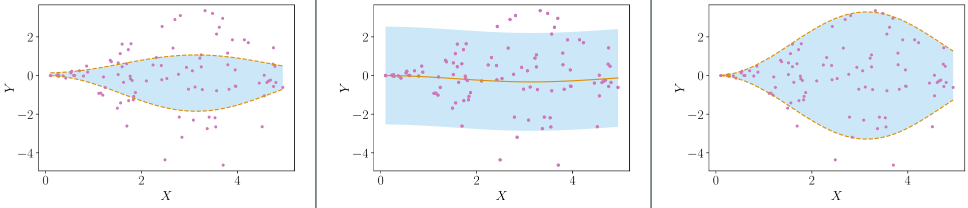
no coverage	marginal	conditional



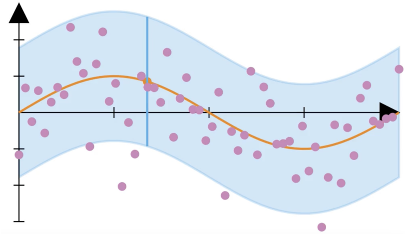
Conditional coverage implies adaptiveness

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

no coverage	marginal	conditional



Standard mean-regression SCP – 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)

CQR: under vs over coverage

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

CQR: implementation details

without anim V0 with $+\infty$

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

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

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

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

CQR: implementation details

without anim   V0 with $+\infty$

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

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

CQR: theoretical guarantees

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

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

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

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

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

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

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

Proof: blackboard!

CQR: theoretical guarantees

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

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

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

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

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

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

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

Proof: blackboard!

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

SCP 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

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

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

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

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

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

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

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

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

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

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

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

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

SCP is defined by the conformity score function

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

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

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

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

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

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

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

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

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

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

$\hookrightarrow$ 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.

SCP: theoretical guarantees

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

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

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

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

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

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

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

Proof: application of the quantile lemma.

SCP: theoretical guarantees

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

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

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

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

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

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

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

Proof: application of the quantile lemma.

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

SCP: standard classification

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

SCP: standard classification

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

SCP: standard classification

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

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

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

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

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

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

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

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

SCP: standard classification in practice

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

- Scores on the calibration set

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

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

SCP: standard classification in practice, cont'd

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

- Scores on the calibration set

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

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

SCP: standard classification in practice, cont'd

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

- Scores on the calibration set

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

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

SCP: standard classification in practice, cont'd

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

- Scores on the calibration set

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

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

SCP: standard classification in practice, cont'd

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

- Scores on the calibration set

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

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

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

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

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

SCP: standard classification in practice, cont'd

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

- Scores on the calibration set

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

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

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

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

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

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

SCP: limits of the standard classification case

Efficiency yet non-adaptivity of the simplest classification scores

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

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

SCP: classification with Adaptive Prediction Sets[8]

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

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

SCP: classification with Adaptive Prediction Sets[8]

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

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

SCP: classification with Adaptive Prediction Sets[8]

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

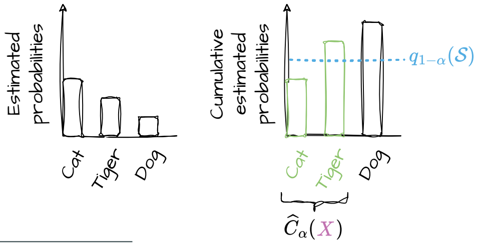
$$r^\star = \operatorname*{arg\,max}_{1 \leq r \leq C} \left\{ \sum_{k=1}^{r} \hat{p}_{\sigma_{X_{n+1}}(k)}(X_{n+1}) < q_{1-\alpha}(\mathcal{S}) \right\} + 1$$

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

SCP: classification with Adaptive Prediction Sets[8]

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

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



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

SCP: classification with Adaptive Prediction Sets in practice

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

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.05	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.50	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.45	0.60	0.55
S_i	0	0	0	0	0	0	0	0.5	0	0

SCP: classification with Adaptive Prediction Sets in practice

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

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.05	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.50	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.45	0.60	0.55
S_i	0	0	0	0	0	0	0	0.5	0	0

- $q_{1-\alpha}(\mathcal{S}) = 0.5$

SCP: classification with Adaptive Prediction Sets in practice

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

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.05	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.50	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.45	0.60	0.55
S_i	0	0	0	0	0	0	0	0.5	0	0

- $q_{1-\alpha}(\mathcal{S}) = 0.5$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.15, 0.4, 0.45)$

SCP: classification with Adaptive Prediction Sets in practice

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

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.05	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.50	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.45	0.60	0.55
S_i	0	0	0	0	0	0	0	0.5	0	0

- $q_{1-\alpha}(\mathcal{S}) = 0.5$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.15, 0.4, 0.45)$, $r^\star = 2$

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

SCP: classification with Adaptive Prediction Sets in practice

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

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.05	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.50	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.45	0.60	0.55
S_i	0	0	0	0	0	0	0	0.5	0	0

- $q_{1-\alpha}(\mathcal{S}) = 0.5$

$\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.15, 0.4, 0.45)$, $r^\star = 2$

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

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

SCP: classification with Adaptive Prediction Sets in practice

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

- Scores on the calibration set

Cal_i	“dog”	“dog”	“dog”	“tiger”	“tiger”	“tiger”	“tiger”	“cat”	“cat”	“cat”
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.05	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.50	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.45	0.60	0.55
S_i	0	0	0	0	0	0	0	0.5	0	0

- $q_{1-\alpha}(\mathcal{S}) = 0.5$

$\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.15, 0.4, 0.45)$, $r^\star = 2$

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

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

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

Split Conformal Prediction: summary

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

Split Conformal Prediction: summary

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

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

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)
- **Marginal** theoretical guarantee over the joint (X, Y) distribution, and not conditional, i.e., no guarantee that for any x:

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

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

Side comment: it is possible to achieve exactly $1-\alpha$ coverage!

Note this other view on conformal sets, through conformal *p*-values:

$$
\begin{aligned}
\widehat{C}_\alpha(x) &= \left\{ y \in \mathcal{Y}, s\left(x, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S}) \right\} \\
&= \left\{ y \in \mathcal{Y}, \sum_{i \in \mathrm{Cal}} \mathbb{1}\left\{ s\left(x, y; \hat{A}\right) > S_i \right\} < (1-\alpha)(1+\#\mathrm{Cal}) \right\} \\
&= \left\{ y \in \mathcal{Y}, \frac{1 + \sum_{i \in \mathrm{Cal}} \mathbb{1}\left\{ s\left(x, y; \hat{A}\right) \leq S_i \right\}}{1 + \#\mathrm{Cal}} > \alpha \right\}.
\end{aligned}
$$

Side comment: it is possible to achieve exactly $1-\alpha$ coverage!

Note this other view on conformal sets, through conformal p-values:

$$
\begin{aligned}
\widehat{C}_\alpha(x) &= \left\{ y \in \mathcal{Y}, s\left(x, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S}) \right\} \\
&= \left\{ y \in \mathcal{Y}, \sum_{i \in \mathrm{Cal}} \mathbb{1}\left\{ s\left(x, y; \hat{A}\right) > S_i \right\} < (1-\alpha)(1+\#\mathrm{Cal}) \right\} \\
&= \left\{ y \in \mathcal{Y}, \frac{1 + \sum_{i \in \mathrm{Cal}} \mathbb{1}\left\{ s\left(x, y; \hat{A}\right) \leq S_i \right\}}{1+\#\mathrm{Cal}} > \alpha \right\}.
\end{aligned}
$$

The randomization procedures drawing $U \sim \mathcal{U}([0,1])$ and outputting: $\widehat{C}^{\mathrm{r}}_\alpha(x) :=$

$$
\left\{ y \in \mathcal{Y}, \frac{\sum_{i \in \mathrm{Cal}} \mathbb{1}\left\{ s\left(x, y; \hat{A}\right) < S_i \right\} + U\left(1 + \sum_{i \in \mathrm{Cal}} \mathbb{1}\left\{ s\left(x, y; \hat{A}\right) = S_i \right\}\right)}{\#\mathrm{Cal}+1} > \alpha \right\},
$$

achieves exactly $1-\alpha$ coverage under exchangeability (smoothed conformal prediction, Vovk et al., 2005).

Introductory remarks

Exchangeability

Split Conformal Prediction (SCP)

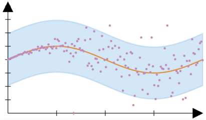
On the design choices of conformity scores and (empirical) conditional guarantees

SCP: what choices for the regression scores?

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

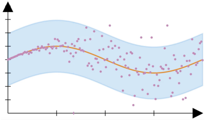
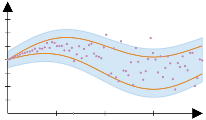
SCP: what choices for the regression scores?

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

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

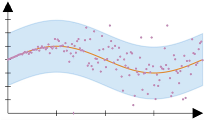
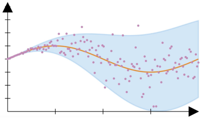
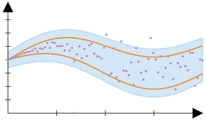
SCP: what choices for the regression scores?

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

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

SCP: what choices for the regression scores?

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

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

Challenges that we are going to explore in the next lectures

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory)

Challenges that we are going to explore in the next lectures

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory) (practical session)

Challenges that we are going to explore in the next lectures

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory) (practical session)
3. Towards conditional coverage? (practical session) (theory) (case studies)
4. Beyond exchangeability (methods) (case studies)

Challenges that we are going to explore in the next lectures

1. On exchangeability (theory)
2. Split conformal prediction (methods) (theory) (practical session)
3. Towards conditional coverage? (practical session) (theory) (case studies)
4. Beyond exchangeability (methods) (case studies)
5. Computational and statistical trade-offs (methods) (theory)
6. Handling missing data (methods)

Some (other, non-exhaustives!) current open directions

- On conformal prediction efficiency (Bars and Humbert, 2025)
- Selective inference (Marandon et al., 2024a)
- Transductive conformal prediction, outlier detection (Vovk et al., 2003; Bates et al., 2023; Marandon et al., 2024b; Gazin et al., 2024)
- Aggregating predictive sets (Gasparin and Ramdas, 2024b,a; Gasparin et al., 2024)
- Multivariate outputs (Klein et al., 2025; Thurin et al., 2025; Kondratyev et al., 2025; Ndiaye, 2025)
- Beyond the indicator loss (Angelopoulos et al., 2022a; Bates et al., 2021; Angelopoulos et al., 2023; Lekeufack et al., 2024)
- Decision-theoretic foundations (Kiyani et al., 2025)

Some (other, non-exhaustives!) current open directions, cont'd

- CP is a very active field of research. Many developments focus on **adapting CP to specific frameworks**, such as:
  - Survival Analysis (Candès et al., 2023),
  - Causal Inference (Lei and Candès, 2021; Jin et al., 2023),
  - NLP (Schuster et al., 2022),
  - RL (Taufiq et al., 2022),
  - federated learning (Humbert et al., 2023; Plassier et al., 2023; Lu et al., 2023),
  - applications (medical (Angelopoulos et al., 2022b; Lu et al., 2022), energy (Kath and Ziel, 2021), etc.) and more.

Angelopoulos, A. N., Barber, R. F., and Bates, S. (2024). Theoretical foundations of conformal prediction.

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 *ICML*.

Bars, B. L. and Humbert, P. (2025). On volume minimization in conformal regression. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J., editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 3030–3053. PMLR.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. (2021). Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6).

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178.

Candès, E., Lei, L., and Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1).

Gasparin, M. and Ramdas, A. (2024a). Conformal online model aggregation.

Gasparin, M. and Ramdas, A. (2024b). Merging uncertainty sets via majority vote.

Gasparin, M., Wang, R., and Ramdas, A. (2024). Combining exchangeable p-values.

Gazin, U., Blanchard, G., and Roquain, E. (2024). Transductive conformal inference with adaptive scores.

Humbert, P., Le Bars, B., Bellet, A., and Arlot, S. (2023). One-shot federated conformal prediction. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14153–14177. PMLR.

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences of the United States of America*, 120(6).

Kath, C. and Ziel, F. (2021). Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2).

Kiyani, S., Pappas, G. J., Roth, A., and Hassani, H. (2025). Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *Forty-second International Conference on Machine Learning*.

Klein, M., Bethune, L., Ndiaye, E., and Cuturi, M. (2025). Multivariate conformal prediction using optimal transport.

Kondratyev, V., Fishkov, A., Kotelevskii, N., Hegazy, M., Flamary, R., Panov, M., and Moulines, E. (2025). Neural optimal transport meets multivariate conformal prediction.

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, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5).

Lekeufack, J., Angelopoulos, A. N., Bajcsy, A., Jordan, M. I., and Malik, J. (2024). Conformal decision theory: Safe autonomous decisions from imperfect predictions.

Lu, C., Angelopoulos, A. N., and Pomerantz, S. (2022). Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Springer Nature Switzerland.

Lu, C., Yu, Y., Karimireddy, S. P., Jordan, M., and Raskar, R. (2023). Federated conformal predictors for distributed uncertainty quantification. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22942–22964. PMLR.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2024a). Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157 – 183.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2024b). Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1).

Ndiaye, E. (2025). Beyond uncertainty sets: Leveraging optimal transport to extend conformal predictive distribution to multivariate settings.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In *Machine Learning: ECML*. Springer.

Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., and Panov, M. (2023). Conformal prediction for federated uncertainty quantification under label shift. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27907–27947. PMLR.

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Romano, Y., Sesia, M., and Candes, E. (2020). Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Sadinle, M., Lei, J., and Wasserman, L. (2018). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V. Q., Tay, Y., and Metzler, D. (2022). Confident adaptive language modeling. In *NeurIPS*.

Taufiq, M. F., Ton, J.-F., Cornish, R., Teh, Y. W., and Doucet, A. (2022). Conformal off-policy prediction in contextual bandits. In *NeurIPS*.

Thurin, G., Nadjahi, K., and Boyer, C. (2025). Optimal transport-based conformal prediction. In *Forty-second International Conference on Machine Learning*.

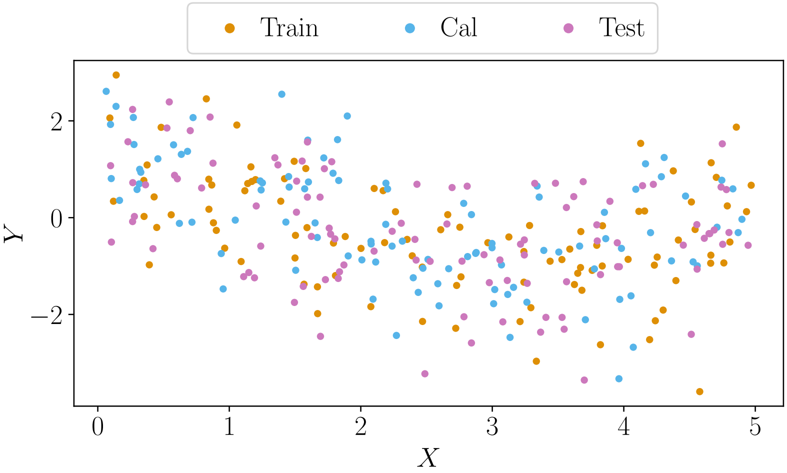
Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

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.

SCP

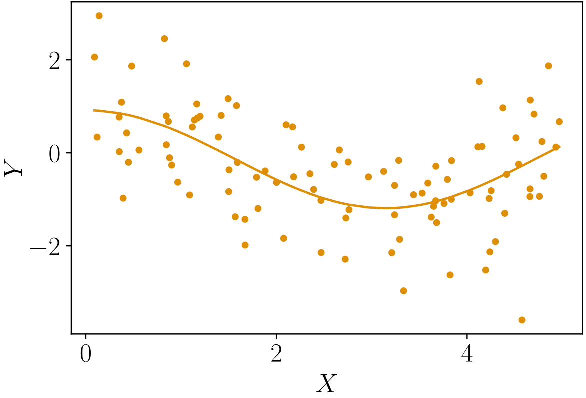
CQR

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 — training step



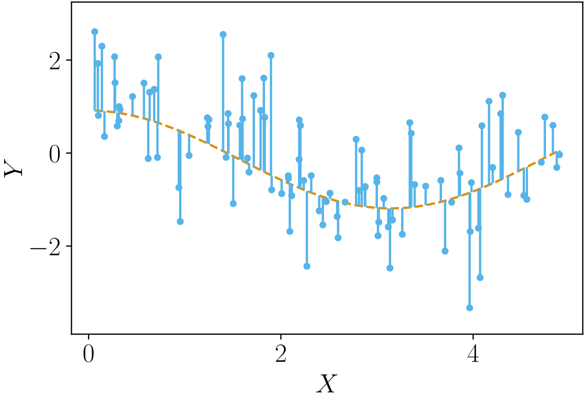
- Learn (or get) $\hat{\mu}$

[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



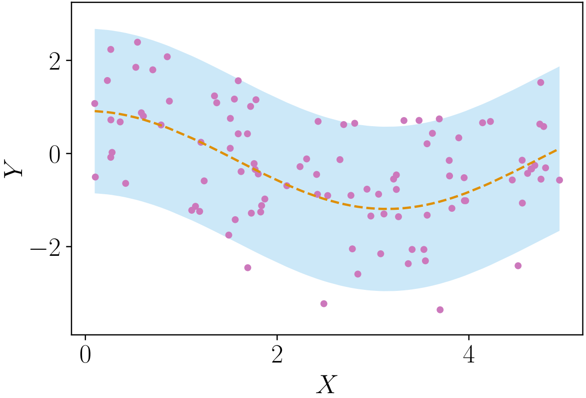
- Predict with $\hat{\mu}$
- Get the $|$residuals$|$, a.k.a. conformity scores
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S} = \{|\text{residuals}|\}_{\text{Cal}} \cup \{+\infty\}$, noted $q_{1-\alpha}(\mathcal{S})$

[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 — prediction step



- 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*

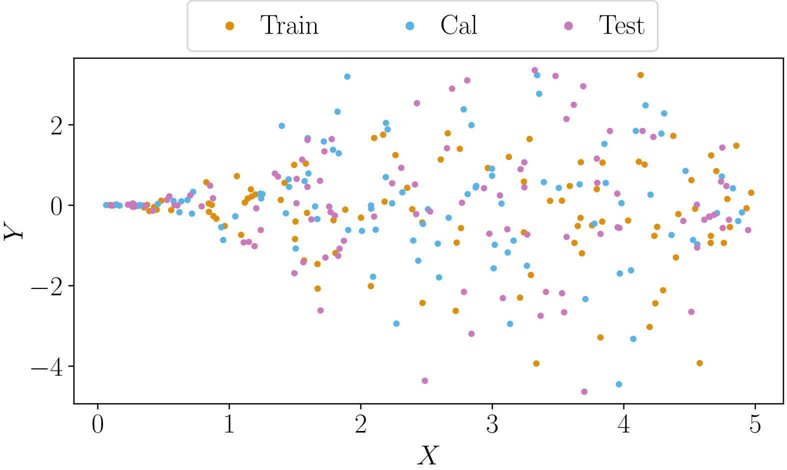
[2]Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML

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

SCP

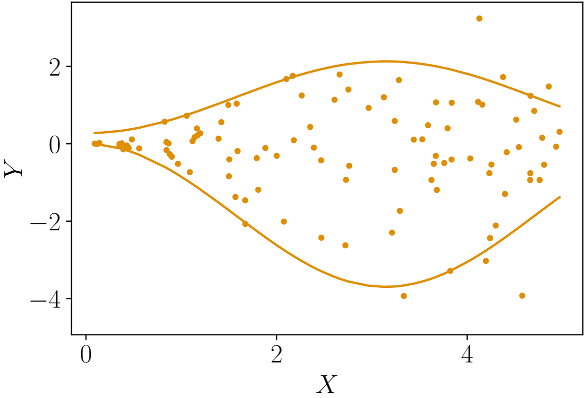
CQR

Conformalized Quantile Regression (CQR)[5]



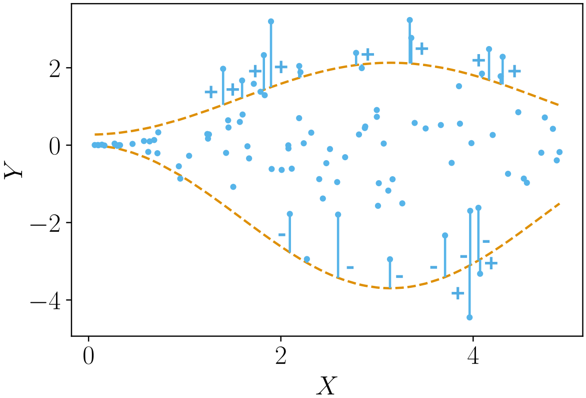
[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS

Conformalized Quantile Regression (CQR)[5]



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

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



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

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

Back to CQR

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