Lecture on Conformal Prediction

Margaux Zaffran

December 1-5, 2025

ECAS-SFdS School

Yesterday's lecture

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

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 $\mathcal{D}^{\text{exch}(n+1)}$,

we have that:

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

SCP is defined by the conformity score function

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

SCP is defined by the conformity score function

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

SCP is defined by the conformity score function

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

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

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

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

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

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

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

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

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

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

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: $\mathrm{s}(\hat{A}(X_i), Y_i) := \max\left(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right)$ 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 2: $\widehat{C}_\alpha(X_{n+1}) = [\widehat{QR}_{\mathrm{lower}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S});$

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

SCP is defined by the conformity score function

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

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

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

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

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

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:

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

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:

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

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

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

- Impact of the calibration set on the coverage
- On distribution-free X-conditional validity
- Y-conditional validity
- Case study: healthcare

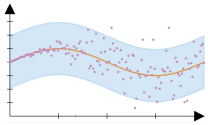
Beyond exchangeability

SCP: what choices for the regression scores?

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

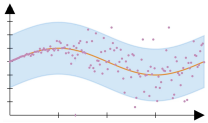
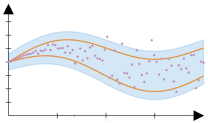
SCP: what choices for the regression scores?

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

	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		

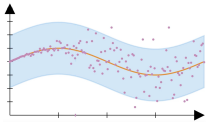
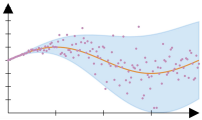
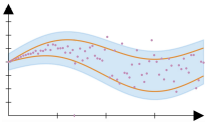
SCP: what choices for the regression scores?

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

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

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

Challenges that we are going to explore in the today's lecture

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

Challenges that we are going to explore in the today's lecture

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

Challenges that we are going to explore in the today's lecture

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

Challenges that we are going to explore in the today's lecture

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)

Theorem (Distribution conditional on the calibration data)**.**

If the scores are a.s. distinct, SCP outputs $\widehat{C}_\alpha$ such that for any distribution $\mathcal{D}$:

$$\mathbb{P}_{\mathcal{D}}\left(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) | (X_i, Y_i)_{i\in \mathrm{Cal}}\right) \sim \beta\left(k_\alpha, \#\mathrm{Cal} + 1 - k_\alpha\right),$$

with $k_\alpha = \lceil (1-\alpha)(\#\mathrm{Cal}+1) \rceil$.

From the β distribution, we get that it has

expectation $\dfrac{k_\alpha}{k_\alpha + \#\mathrm{Cal} + 1 - k_\alpha} = \dfrac{k_\alpha}{\#\mathrm{Cal}+1} = \dfrac{\lceil (1-\alpha)(\#\mathrm{Cal}+1) \rceil}{\#\mathrm{Cal}+1} \geq 1-\alpha,$

and variance $\dfrac{k_\alpha\left(\#\mathrm{Cal} + 1 - k_\alpha\right)}{\left(\#\mathrm{Cal}+1\right)^2\left(\#\mathrm{Cal}+2\right)} \approx \dfrac{\alpha(1-\alpha)}{\#\mathrm{Cal}+2}.$

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 $0 < \delta \leq 0.5$:

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

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

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

- Impact of the calibration set on the coverage
- On distribution-free X-conditional validity
- Y-conditional validity
- Case study: healthcare

Beyond exchangeability

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 $\mathcal{D}^{\text{exch}(n+1)}$,

we have that:

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

Informative conditional coverage as such is impossible

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

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

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

Informative conditional coverage as such is impossible

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

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

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

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

Informative conditional coverage as such is impossible

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

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

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

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

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

Informative conditional coverage as such is impossible

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

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

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

↪ distribution-free X-conditional hardness result apply beyond CP
↪ X-conditional estimators are overly large even on easy cases
↪ the lower bounds are tight

Example (Naive estimator).

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

Informative conditional coverage as such is impossible

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

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

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

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

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

$\hookrightarrow$ the lower bounds are tight

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

Getting closer to X-conditional coverage

	X-conditionally valid	
"Non smooth" distribution		

Getting closer to X-conditional coverage

	X-conditionally valid	
"Non smooth" distribution	X-cov.: ✓ Length: ✓	

Getting closer to X-conditional coverage

	X-conditionally valid	
"Non smooth" distribution	X-cov.: ✓ Length: ✓	
"Smooth" distribution	X-cov.: ✓ Length: ✗	

Getting closer to X-conditional coverage

	X-conditionally valid	Non X-conditionally valid
"Non smooth" distribution	X-cov.: ✓ Length: ✓	X-cov.: ✗ Length: `not relevant`
"Smooth" distribution	X-cov.: ✓ Length: ✗	

Getting closer to X-conditional coverage

	X-conditionally valid	Non X-conditionally valid
"Non smooth" distribution	X-cov.: ✓ Length: ✓	X-cov.: ✗ Length: `not relevant`
"Smooth" distribution	X-cov.: ✓ Length: ✗	X-cov.: $\approx$ Length: ✓

Getting closer to X-conditional coverage

	X-conditionally valid	Non X-conditionally valid
"Non smooth" distribution	X-cov.: ✓ Length: ✓	X-cov.: ✗ Length: `not relevant`
"Smooth" distribution	X-cov.: ✓ Length: ✗	X-cov.: $\approx$ Length: ✓

- 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

	X-conditionally valid	Non X-conditionally valid
"Non smooth" distribution	X-cov.: ✓ Length: ✓	X-cov.: ✗ Length: `not relevant`
"Smooth" distribution	X-cov.: ✓ Length: ✗	X-cov.: $\approx$ Length: ✓

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

Non exhaustive references.

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

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

Let $\delta > 0$ be a tolerance level.

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

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

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

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

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

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

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

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

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

Pre-defined partition of the $\mathcal{X}$ space

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

$\mathcal{G}$ a set of "groups" (i.e., define G a random variable taking its values in $\mathcal{G}$).

Definition (Distribution-free $\mathcal{G}$-conditional validity ($\mathcal{G}$CV)).

$\widehat{C}_\alpha$ achieves distribution-free $\mathcal{G}$-conditional validity if for any distribution $\mathcal{D}$ on $(\mathcal{X}, \mathcal{G}, \mathcal{Y})$, it holds that:

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

Theorem (General $\mathcal{G}$CV hardness result).

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

- *Regression*

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

- *Classification*

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

Theorem (General $\mathcal{G}$CV hardness result).

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

- *Regression*

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

- *Classification*

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

Theorem (General $\mathcal{G}$CV hardness result).

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

- *Regression*

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

- *Classification*

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

Theorem (General $\mathcal{G}$CV hardness result).

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

- *Regression*

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

- *Classification*

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

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

Theorem (General $\mathcal{G}$CV hardness result).

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

- *Regression*

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

- *Classification*

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

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

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

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

Theorem (General $\mathcal{G}$CV hardness result).

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

- *Regression*

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

- *Classification*

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

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

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

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

Restricting the link between G and (X or Y) does not allow informative $\mathcal{G}$CV

Analogous statements are also available for the classification framework.

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

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

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

Analogous statements are also available for the classification framework.

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

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

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

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

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

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

Analogous statements are also available for the classification framework.

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

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

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

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

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

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

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

Analogous statements are also available for the classification framework.

Implications for $\mathcal{G}$CV in practice

	$\mathcal{G}$-conditionally valid even when $G \not\perp\!\!\!\perp (X, Y)$	
"Non smooth" distribution		

Implications for $\mathcal{G}$CV in practice

	$\mathcal{G}$-conditionally valid even when $G \not\perp\!\!\!\perp (X, Y)$	
"Non smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✓	

Implications for $\mathcal{G}$CV in practice

	$\mathcal{G}$-conditionally valid even when $G \not\perp\!\!\!\perp (X, Y)$	
"Non smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✓	
"Smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✗	

Implications for $\mathcal{G}$CV in practice

	$\mathcal{G}$-conditionally valid even when $G \not\perp\!\!\!\perp (X, Y)$	$\mathcal{G}$-conditionally valid at most if $G \perp\!\!\!\perp (X, Y)$
"Non smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✓	$\mathcal{G}$-cov.: ✗ Length: `not relevant`
"Smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✗	

Implications for $\mathcal{G}$CV in practice

	$\mathcal{G}$-conditionally valid even when $G \not\perp\!\!\!\perp (X, Y)$	$\mathcal{G}$-conditionally valid at most if $G \perp\!\!\!\perp (X, Y)$
"Non smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✓	$\mathcal{G}$-cov.: ✗ Length: `not relevant`
"Smooth" distribution	$\mathcal{G}$-cov.: ✓ Length: ✗	$\mathcal{G}$-cov.: $\approx$ Length: ✓

Achieving Y-conditional validity in classification

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

Achieving Y-conditional validity in classification

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

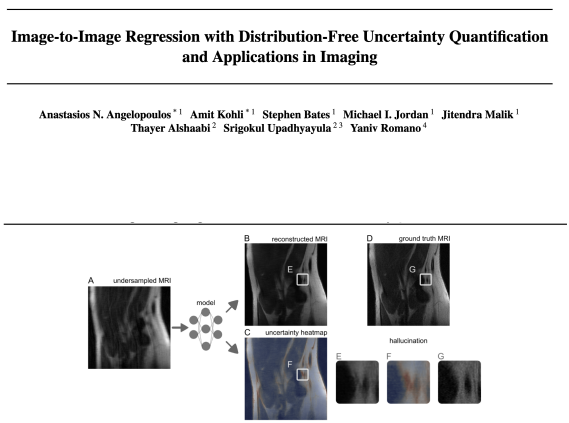


Figure 1. **An algorithmic MRI reconstruction with uncertainty.** A rapidly acquired but undersampled MR image of a knee (A) is fed into a model that predicts a sharp reconstruction (B) with calibrated uncertainty (C). In (C), red means high uncertainty and blue means low uncertainty. Wherever the reconstruction 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. (2022)

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

Method:

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

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

Method:

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

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

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.

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

Method:

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

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

     $\to$ 3 regressors are used

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

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

Method:

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

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

   $\to$ 3 regressors are used

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

Guarantee:

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

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

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

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. (2022)

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.

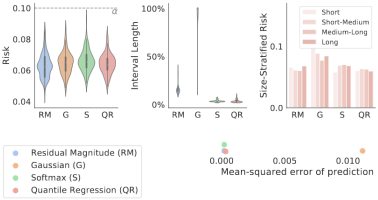
- Hard problem (impossibility results!)
- Introduce metrics to see *if* and *on which underlying regressors* such problem happens.

How do you understand that?

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

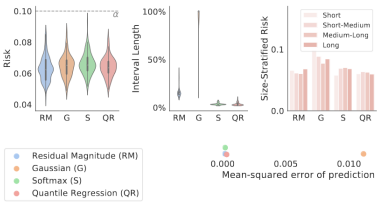
Example of such metrics (see also Feldman et al., 2021) **:**

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



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

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



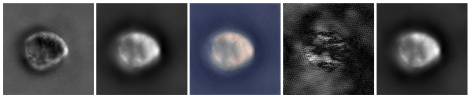


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth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

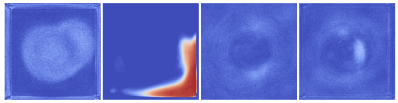
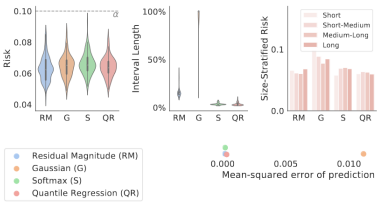


Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnitude, gaussian, softmax, and quantile regression.

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

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

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



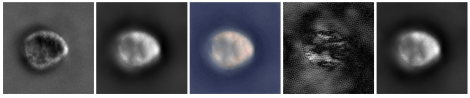


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth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

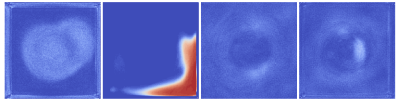


Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnitude, gaussian, softmax, and quantile regression.

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

Take aways:

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

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

Beyond exchangeability

- Some (short) literature review
- Case study: electricity price forecasting

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

Beyond exchangeability

Some (short) literature review

Case study: electricity price forecasting

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

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

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

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. $\omega_i = \omega(X_i) = \dfrac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$
  3. outputs $\widehat{C}_\alpha(X_{n+1}) =$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Generalizations

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

Generalizations

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

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:
  - Chernozhukov et al. (2018)
    $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓
  - Barber et al. (2022)
    $\hookrightarrow$ Quantifies the coverage loss depending on the strength of exchangeability violation
    $$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha - \underbrace{\text{average violation of exchangeability by each calibration point}}$$
    $\hookrightarrow$ proposed algorithm: reweighting again!
    e.g., in a temporal setting, give higher weights to more recent points.

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

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

  often relaxed in:

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

Online setting

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

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

  often relaxed in:

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

$\rightsquigarrow$ More during the case study!

Recent developments (as of late 2023...)

- Consider splitting strategies that respect the temporal structure

Non exhaustive references.

Recent developments (as of late 2023...)

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

Non exhaustive references.

Recent developments (as of late 2023...)

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

Non exhaustive references.

Recent developments (as of late 2023...)

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

Non exhaustive references.

Recent developments (as of late 2023...)

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

Non exhaustive references.

Recent developments (as of late 2023...)

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

Non exhaustive references.

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

Beyond exchangeability

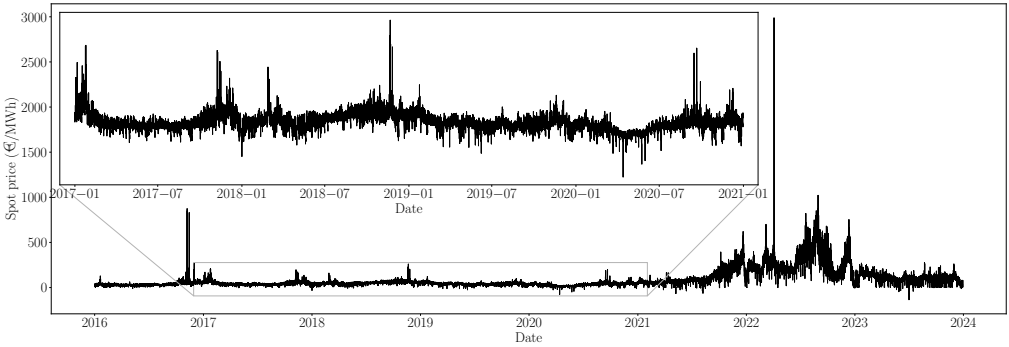
Some (short) literature review

Case study: electricity price forecasting

Forecasting French spot electricity prices

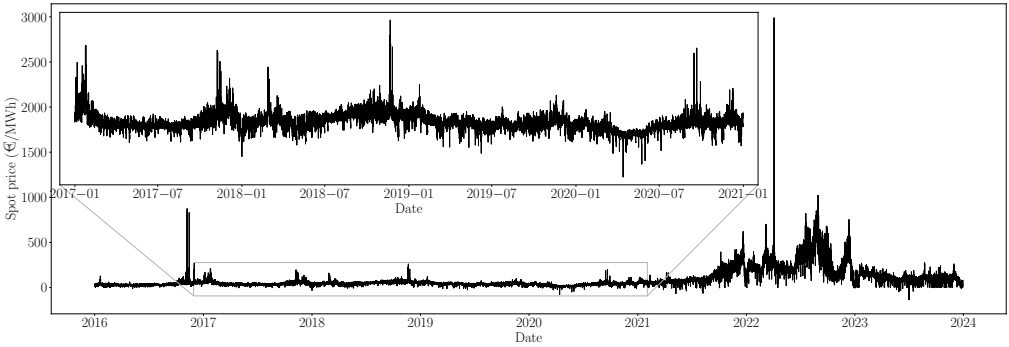
Hourly day-ahead market prices (between producers and suppliers)

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



Forecasting French spot electricity prices

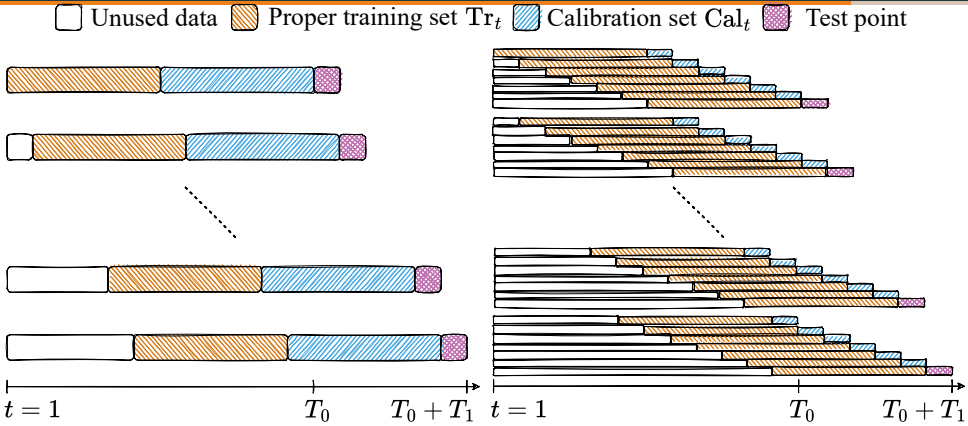
Hourly day-ahead market prices (between producers and suppliers)



To which extent are they forecastable?

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

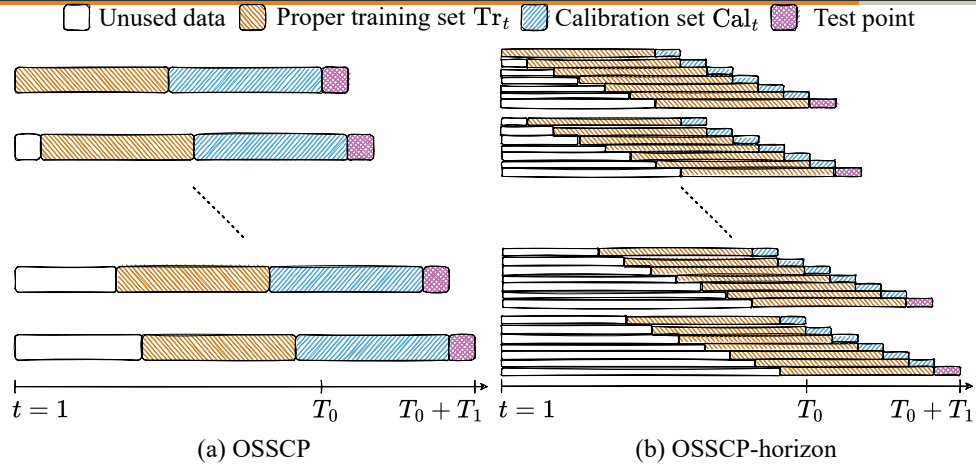
Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP

(b) OSSCP-horizon

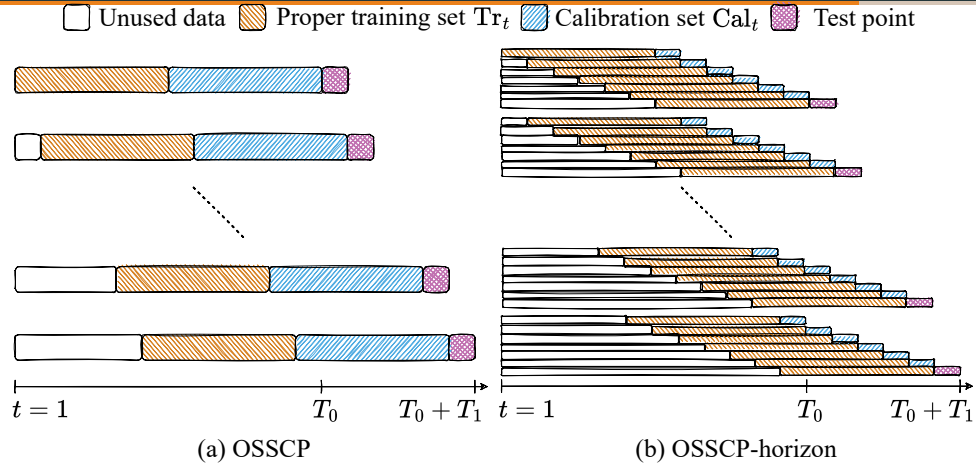
Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



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

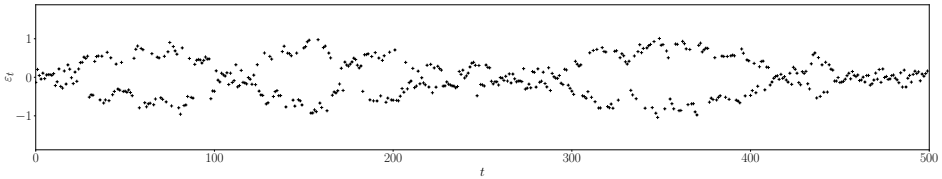
Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

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

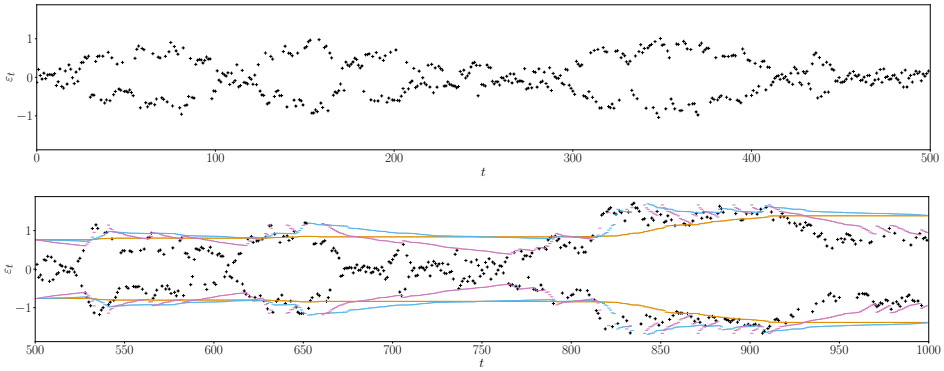
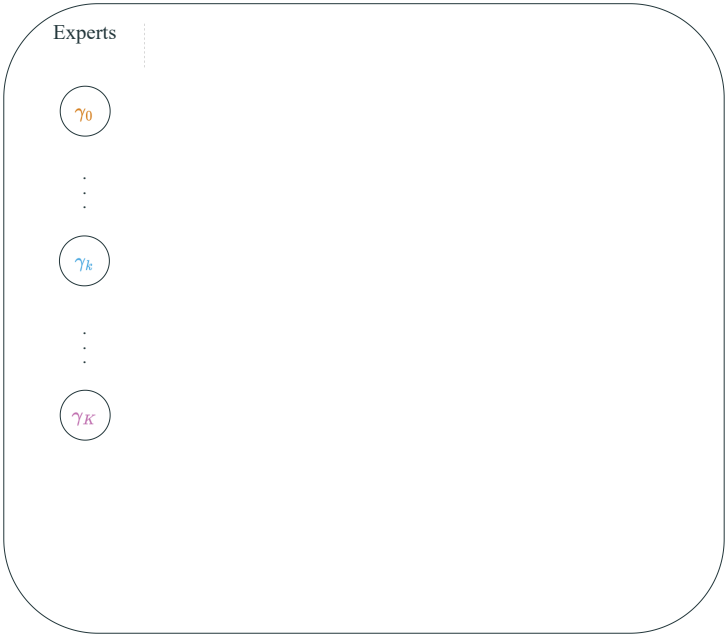
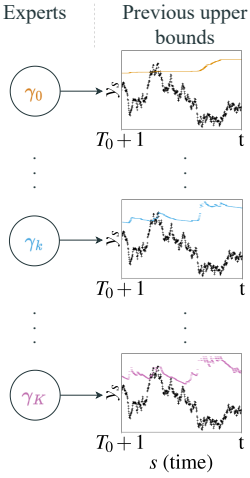


Figure 3: Visualisation of ACI with different values of γ ($\gamma = 0$, $\gamma = 0.01$, $\gamma = 0.05$)

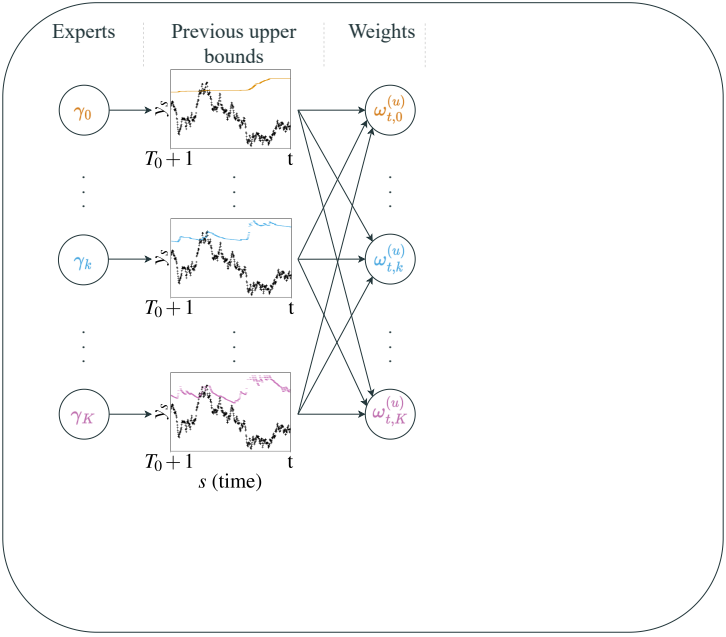
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



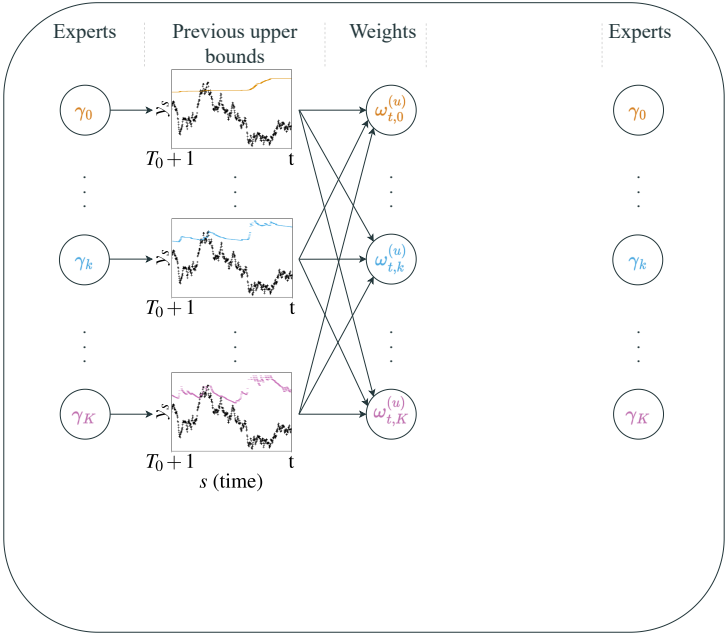
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



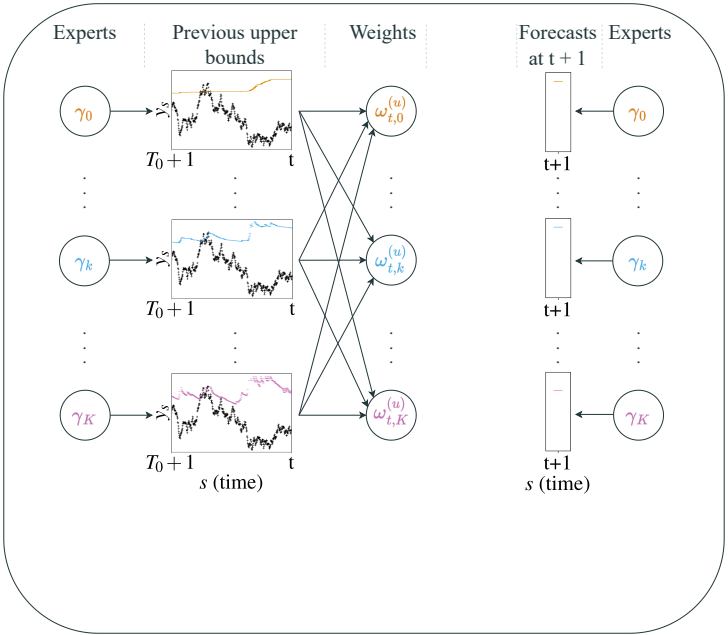
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



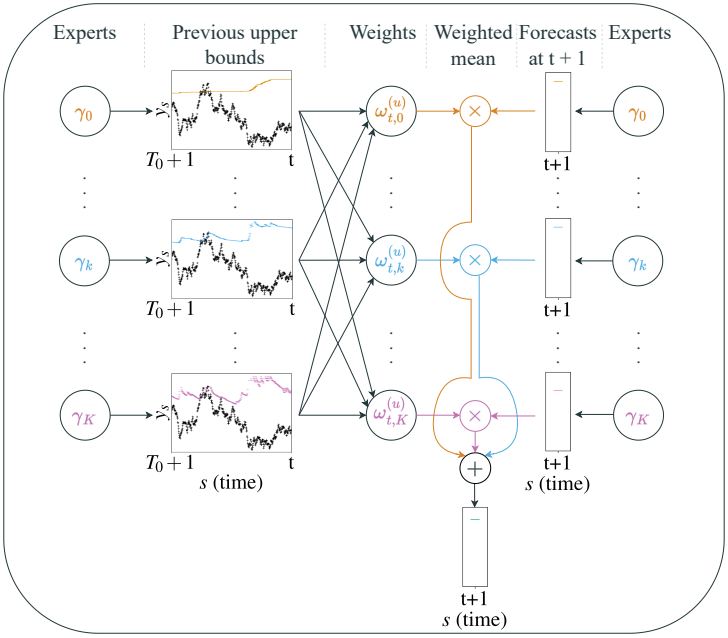
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



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

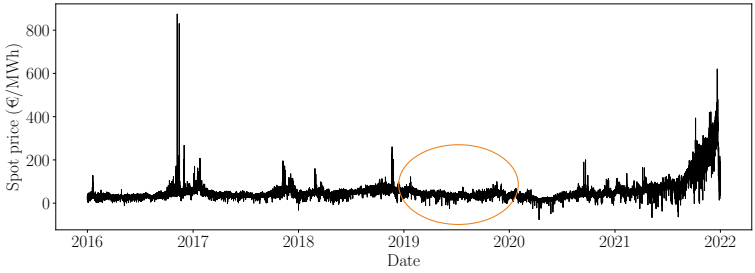


AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



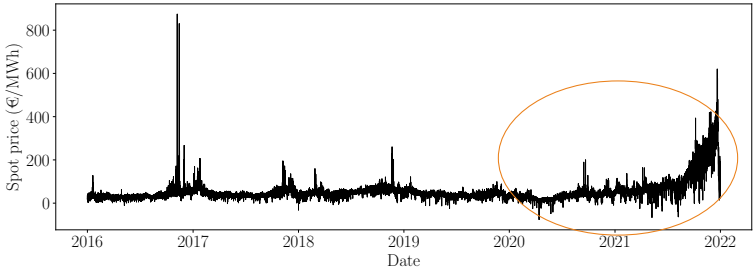
Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- 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[1] behave differently.



[1] 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)

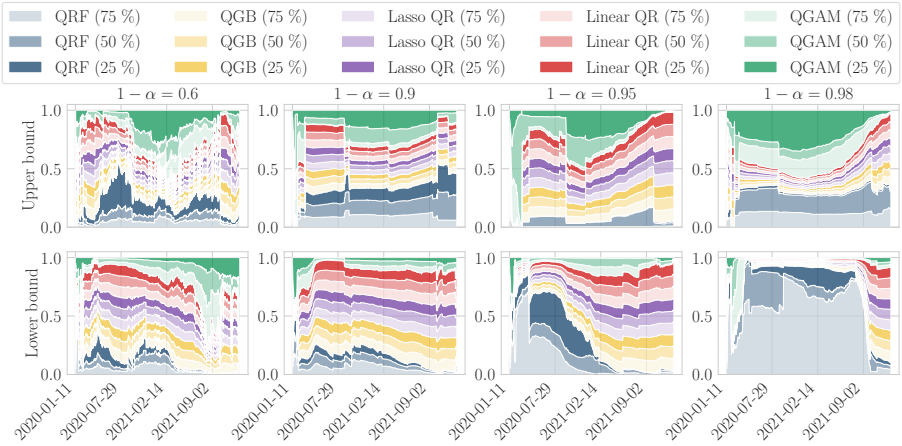
Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ 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.



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

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

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

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

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

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?

Where are we now?

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

Where are we now?

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)

Where are we now?

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)

Angelopoulos, A. N., Candès, E. J., and Tibshirani, R. J. (2023). Conformal pid control for time series prediction. arXiv: 2307.16895.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. (2022). 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. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2).

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (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. (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34.

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).

Ding, T., Angelopoulos, A., Bates, S., Jordan, M., and Tibshirani, R. J. (2023). Class-conditional conformal prediction with many classes. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64555–64576. Curran Associates, Inc.

Dutot, G., Zaffran, M., Féron, O., and Goude, Y. (2024). Adaptive probabilistic forecasting of french electricity spot prices.

Feldman, S., Bates, S., and Romano, Y. (2021). Improving Conditional Coverage via Orthogonal Quantile Regression. *arXiv:2106.00394 [cs]*. arXiv: 2106.00394.

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).

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., 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).

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.

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.

Zaffran, M., Dieuleveut, A., Josse, J., and Romano, Y. (2024). Predictive uncertainty quantification with missing values. Preprint submitted to *Journal of Machine Learning Research*, arXiv arXiv:2405.15641.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.