

An introduction to conformal prediction sets

Draft notes of
E. Ročnin

- disclaimers:
- not exhaustive (very active area)
 - prediction sets for response \neq confidence region for parameter

Classical review paper:
[Angelopoulos and Bates (2022)]
↳ very nice historical notes on work

Road map:

I Setting

- ① Notation and aim
- ② Non-conformity score function

II Classical methods

- ① Naive method
- ② Split conformal method
- ③ Full conformal method
- ④ Cross-validated conformal method

III Some extensions

- ① Nested prediction sets
- ② Conditional guarantees
- ③ Beyond exchangeability
- ④ Transductive approach

I Setting

① Notation and aim

- Let (X_i, Y_i) random variables with
covariate \swarrow \searrow outcome

$X_i \in \mathbb{R}^d$ & possibly large (for simplicity, it could be image, graph,...)

$Y_i \in \mathcal{Y}$ e.g. $\mathcal{Y} = \{1, \dots, K\}$ classification or $\mathcal{Y} = \mathbb{R}$ regression

- (X_i, Y_i) are iid $\sim P$ (unknown) over several samples:

$\mathcal{D} = ((X_1, Y_1), \dots, (X_n, Y_n))$ full training sample observed

$\mathcal{D}_{\text{test}} = (X_{n+1}, Y_{n+1})$ test sample

observed \swarrow \searrow not observed

Aim: build $\hat{\mathcal{C}}_\alpha(\cdot; \mathcal{D}) \subset \mathcal{Y}$ such that

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_\alpha(X_{n+1}; \mathcal{D})) \geq 1 - \alpha \quad \text{"prediction set"}$$

⊕ true for any distribution P (or weak assumption like no ties)

⊕ true for any sample size $n \geq 1$

⊕ can be combined with any "machine learning" algorithm

⊖ $\hat{\mathcal{C}}_\alpha = \mathcal{Y}$ allowed (not necessarily informative)

② Non-conformity score function

Measurable function $S(\cdot; \mathcal{D}) : (x, y) \in \mathbb{R}^d \times \mathcal{Y} \rightarrow \mathbb{R}$

• regression $S(x, y; \mathcal{D}) = |y - \hat{\mu}(x, \mathcal{D})|$ "residual"

for $\hat{\mu}(x, \mathcal{D})$ estimator of $\mu(x) = \mathbb{E}[Y_1 | X_1 = x]$

• classification $S(x, y; \mathcal{D}) = 1 - \hat{\pi}_x(y; \mathcal{D})$ "prob that label y is not predicted in x "

for $\hat{\pi}_x(y; \mathcal{D})$ estimator of $\pi_x(y) = \mathbb{P}(Y_1 = y | X_1 = x)$

$S(x, y; \mathcal{D})$ is large if y does not conform to the prediction in x

II Classical methods

① Naive method

• Compute $S_i = S(X_i, Y_i; \mathcal{D})$, $1 \leq i \leq n$

↳ empirical distribution of 'typical' non conformity scores

• Take q_α "some quantile" of this distribution

• Choose $\hat{\mathcal{C}}_\alpha(X_{n+1}, \mathcal{D}) = \{y \in \mathcal{Y} : S(X_{n+1}, y; \mathcal{D}) \leq q_\alpha\}$
as prediction set

⚠ fail in general: under interpolation $S_i = 0$, $1 \leq i \leq n$
↳ only works asymptotically, with a good predictor! ⚠

The problem is that \mathcal{D} is used both for prediction and for evaluating the quality of the prediction

② Split conformal method

[Papadopoulos et al. (2002)] = inductive
= split conformal

Split \mathcal{D} into

$$\mathcal{D}_{\text{train}} = ((X_i, Y_i), i \in I_{\text{train}}) \text{ size } n_{\text{train}}$$

$$\mathcal{D}_{\text{cal}} = ((X_i, Y_i), i \in I_{\text{cal}}) \text{ size } n_{\text{cal}}$$

- Non-conformity score function $S(x, y; \mathcal{D}_{\text{train}})$
- Scores $S_i = S(X_i, Y_i; \mathcal{D}_{\text{train}})$, $i \in I_{\text{cal}}$
- Take $q_\alpha = (1-\alpha)$ -quantile of $(S_i, i \in I_{\text{cal}}) \cup \{+\infty\}$
 $= (1-\alpha)(1 + \frac{1}{n_{\text{cal}}})$ -quantile of $(S_i, i \in I_{\text{cal}})$
 $= S_{(\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil)}$

$$\text{for } S_{(1)} \leq \dots \leq S_{(n_{\text{cal}})} \leq S_{(n_{\text{cal}}+1)} = +\infty$$

Then the split conformal prediction set is given by

$$\hat{\mathcal{C}}_\alpha^{\text{split}}(X_{n+1}, \mathcal{D}) = \{y \in \mathcal{Y} : S(X_{n+1}, y; \mathcal{D}_{\text{train}}) \leq q_\alpha\}$$

Examples :

* regression : prediction interval

$\hat{\mu}(x, D_{\text{train}})$ estimator of $\mu(x) = E[Y_1 | X_1 = x]$ can be anything! (OLS, NN, RF, ...)

$$S_i = |Y_i - \hat{\mu}(X_i, D_{\text{train}})|, i \in I_{\text{cal}}$$

$$\hat{\mathcal{L}}_{\alpha}^{\text{split}}(X_{n+1}, D) = \{y \in \mathcal{Y} : |y - \hat{\mu}(X_{n+1}, D_{\text{train}})| \leq q_{\alpha}\}$$
$$= [\hat{\mu}(X_{n+1}, D_{\text{train}}) - q_{\alpha}, \hat{\mu}(X_{n+1}, D_{\text{train}}) + q_{\alpha}]$$

Illustration good predictor $\Leftrightarrow q_{\alpha}$ small \Leftrightarrow small interval

* classification : prediction label set

$\hat{\pi}_x(y; D_{\text{train}})$ estimator of $\pi_x(y) = P(Y_1 = y | X_1 = x)$ (RNN, NN, ...)

$$S_i = 1 - \hat{\pi}_{X_i}(Y_i; D_{\text{train}}), i \in I_{\text{cal}}$$

$$\hat{\mathcal{L}}_{\alpha}^{\text{split}}(X_{n+1}, D) = \{y \in \mathcal{Y} : 1 - \hat{\pi}_{X_{n+1}}(y; D_{\text{train}}) \leq q_{\alpha}\}$$
$$= \{y \in \mathcal{Y} : \hat{\pi}_{X_{n+1}}(y; D_{\text{train}}) \geq 1 - q_{\alpha}\}$$

Note that the threshold $1 - q_{\alpha}$ is not $1 - \alpha$

\hookrightarrow on garde les labels $y \in \mathcal{Y}$ les plus probables au vu des scores de references.

good classification $\Leftrightarrow q_{\alpha}$ small \Leftrightarrow few labels in $\hat{\mathcal{L}}_{\alpha}$

Proposition: $\forall P, \forall (n_{\text{train}}, n_{\text{cal}}),$

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_{\alpha}^{\text{split}}(X_{n+1}; \mathcal{D}_{\text{train}})) \geq 1 - \alpha$$

and if $\mathbb{P}(S_i = S_j) = 0 \quad \forall i \neq j \leq 1 - \alpha + \frac{1}{n_{\text{cal}} + 1}$

split conformal is valid

with accuracy $\approx \frac{1}{n_{\text{cal}}}$

Proof: $S_{n+1} = S(X_{n+1}, Y_{n+1}; \mathcal{D}_{\text{train}})$

$Y_{n+1} \notin \hat{\mathcal{C}}_{\alpha}(X_{n+1}; \mathcal{D}_{\text{train}}) \Leftrightarrow S_{(n\alpha)} < S_{n+1}, \quad n\alpha = \lceil (1-\alpha)(n_{\text{cal}}+1) \rceil$

$$\Leftrightarrow \sum_{i \in \mathcal{I}_{\text{cal}}} \mathbb{1}\{S_i < S_{n+1}\} \geq n\alpha$$

$n_{\text{cal}}+1$ elements in $\sum_{i \in \mathcal{I}_{\text{cal}} \cup \{n+1\}} \mathbb{1}\{S_i < S_{n+1}\} \geq (1-\alpha)(n_{\text{cal}}+1)$

$$\Leftrightarrow \sum_{i \in \mathcal{I}_{\text{cal}} \cup \{n+1\}} \mathbb{1}\{S_i \geq S_{n+1}\} \leq \alpha(n_{\text{cal}}+1)$$

= rank of S_{n+1} in $\{S_1, \dots, S_{n_{\text{cal}}}, S_{n+1}\}$

• But $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ iid

$\Rightarrow S_1, \dots, S_{n_{\text{cal}}}, S_{n+1}$ exchangeable | $\mathcal{D}_{\text{train}}$

if $S_i \neq S_j$ for $i \neq j$ almost surely then rank $\sim U(1, \dots, n_{\text{cal}}+1)$

and $\mathbb{P}(Y_{n+1} \notin \hat{\mathcal{C}}_{\alpha}(X_{n+1}; \mathcal{D}_{\text{train}})) = \frac{\lceil \alpha(n_{\text{cal}}+1) \rceil}{n_{\text{cal}}+1}$ okay!

$$\in \left[\alpha - \frac{1}{n_{\text{cal}}+1}, \alpha \right]$$

in general, we have

Lemma [Romano and Wolf (2005)] (empirical p-values are super-uniform)

For $(Y_1, \dots, Y_B, Y_{B+1})$ any real valued exchangeable random vector

$$\mathbb{P} \left(\frac{1}{B+1} \sum_{b=1}^{B+1} \mathbb{1} \{ Y_b \geq Y_{B+1} \} \leq \alpha \right) \leq \frac{[\alpha(B+1)]}{B+1} \leq \alpha$$

↳ \hat{p} (score to \geq observed score) 'empirical p-value'

Proof: see for instance [Arlot et al. (2010)] proof Lemma 5.2

Here the empirical p-value is: $\hat{p}(Y_{n+1}) = \frac{1}{n+1} \sum_{i \in I_{\text{cal}} \cup \{n+1\}} \mathbb{1} \{ S_i \geq S_{n+1} \}$

This shows $\mathbb{P} (Y_{n+1} \notin \hat{G}_\alpha(X_{n+1}; D_{\text{train}})) \leq \frac{[\alpha(n+1)]}{n+1} \leq \alpha$ \square

Split conformal \oplus easy to compute
 \ominus splitting a bit arbitrary and sample waste

③ Full conformal method [Vovk et al (2005)]

 Keep $D = ((X_1, Y_1), \dots, (X_n, Y_n))$ full training sample and train with $D \cup \{(X_{n+1}, Y_{n+1})\}$

• non conformity score function $S(x, y; D \cup \{(X_{n+1}, Y_{n+1})\})$ which is assumed permutation invariant w.r.t. sample

• scores $S_i(Y_{n+1}) = S(X_i, Y_i; D \cup \{(X_{n+1}, Y_{n+1})\})$, $1 \leq i \leq n+1$

$\hat{G}_\alpha(X_{n+1}, D) = \{y \in \mathcal{Y} : S_i(y) \leq S_{(\lceil (1-\alpha)(n+1) \rceil)}(y)\}$

Proposition: $\forall P, \forall n,$

$$\mathbb{P}(Y_{n+1} \in \overset{\text{full}}{S}_\alpha(X_{n+1}; \mathcal{D}_{\text{train}})) \geq 1 - \alpha$$

$$\text{and if } \mathbb{P}(S_i = S_j) = 0 \quad \forall i \neq j \leq 1 - \alpha + \frac{1}{n+1}$$

full conformal is valid

Proof by the previous proof, just have to show that

$(S_1, \dots, S_n, S_{n+1})$ are exchangeable

clear because $\forall \sigma$ permutation of $\{1, \dots, n+1\}$

$$(S_{\sigma(i)}, 1 \leq i \leq n+1) = (S(X_{\sigma(i)}, Y_{\sigma(i)}; ((X_j, Y_j), 1 \leq j \leq n+1)), 1 \leq i \leq n+1)$$

$$= (S(X_{\sigma(i)}, Y_{\sigma(i)}; ((X_j, Y_j), 1 \leq j \leq n+1)), 1 \leq i \leq n+1)$$

↑
permutation invariance of score function

$$\mathcal{N}(S(X_i, Y_i; (X_j, Y_j), 1 \leq j \leq n+1), 1 \leq i \leq n+1)$$

↑
because $(X_1, Y_1) \dots (X_{n+1}, Y_{n+1})$ are exchangeable

□

Full conformal

- ⊕ no sample splitting and 'cleaner'
- ⊖ difficult to compute

Possible to 'conformalize'. Ridge [Noureltdinov et al (2001)]

. LASSO [Lei (2017)]

↳ see also the works of Eugène Ndiaye (Apple Paris)

④ Cross-validated conformal method [Barber et al (2021)] = Jackknife +

- Aim: \oplus computable
 \oplus no sample splitting

but ... \ominus loss of a factor 2 in the coverage

- score function $S(x, y; D')$ permutation invariant in sample D'
- scores $S_i = S(X_i, Y_i; D_{-i})$, $1 \leq i \leq n$ 'leave one out' = 'jackknife'
 $\uparrow (X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n)$
- $\hat{p}(Y_{n+1}) = \frac{1}{n+1} \left(1 + \sum_{i=1}^n \mathbb{1}\{S_i \geq S(X_{n+1}, Y_{n+1}; D_{-i})\} \right)$
- $\hat{C}_\alpha^{cv}(X_{n+1}; D) = \{y \in Y : \hat{p}(y) > \alpha\}$

Théorème: $\forall p, \forall n, \mathbb{P}(Y_{n+1} \in \hat{C}_\alpha^{cv}(X_{n+1}, D)) \geq 1 - 2\alpha + \frac{1}{n+1}$
 $\geq 1 - 2\alpha$

cv conformal is valid up to a factor 2

Proof Key point:

Lemma: for any binary antisymmetric matrix A $p \times p$ with null diagonal

$A_{ij} \in \{0, 1\}$ $A_{ij} = 1 - A_{ji}$ $A_{ii} = 0$

$S(A) = \{ 1 \leq i \leq p : \sum_{j=1}^p A_{ij} \geq p(1-\alpha) \}$ nb of lines with that many "1"

is of cardinal $\leq 2\alpha p - 1$

Proof: $\forall i \in S(A)$, we have

$$1 + \sum_{j \in S(A) \setminus \{i\}} (1 - A_{ij}) = \sum_{j \in S(A)} (1 - A_{ij}) \leq \sum_{j=1}^p (1 - A_{ij}) = p - \sum_{j=1}^p A_{ij} \leq p\alpha$$

for any pair $(R, \ell) \subset S(A)$ with $R \neq \ell$ let $(i, j) \subset S(A)$ st $\{i, j\} = \{R, \ell\}$

This makes at most $|S(A)|(p\alpha - 1)$ choices for a pair of $S(A)$ $A_{ij} = 0$

$$\text{Thus } |S(A)|(|S(A)| - 1)/2 \leq |S(A)|(p\alpha - 1) \quad \square$$

Then let $D' = D \cup \{(X_{n+1}, Y_{n+1})\}$

$\{i, j \in \{1, \dots, n+1\}, D'_{-i-j}$ the set D' with (X_i, Y_i) and (X_j, Y_j) have been removed

$$R_{ij} = S(X_i, Y_i; D'_{-i-j}) \quad \text{"residual in } i \text{ with sample } -i-j\text{"}$$

$$A_{ij} = \mathbb{1}\{R_{ij} > R_{ji}\} \quad \text{"score better in } i \text{ than } j\text{"}$$

A $(n+1) \times (n+1)$ binary anti-symmetric matrix

Also, by exchangeability of $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$, we have

$$\forall i, \exists A_{ij} - A_{ji}, 1 \leq j \leq n+1 \sim \exists R_{n+1, \pi(j)} - R_{\pi(j), n+1}, 1 \leq j \leq n+1$$

Thus $\sum_{j=1}^{n+1} A_{ij} \sim \sum_{j=1}^{n+1} A_{n+1, j}$

$\pi: i \leftrightarrow n+1$
not ordered set

Hence $\mathbb{P}(Y_{n+1} \notin \hat{C}_\alpha^{CV}(X_{n+1}, D)) = \mathbb{P}\left(\sum_{j=1}^n \mathbb{1}\{S(X_j, Y_j; D'_{-n+1-j}) < S(X_{n+1}, Y_{n+1}; D'_{-n+1-j})\}\right)$

$D_{-i} = D'_{-n+1-i}$

$$= \mathbb{P}\left(\sum_{j=1}^{n+1} A_{n+1, j} \geq (n+1)\alpha\right)$$

put \mathbb{E} outside

$$\stackrel{(*)}{=} \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(\sum_{j=1}^{n+1} A_{i, j} \geq (n+1)\alpha\right) = \mathbb{E}\left[\frac{|S(A)|}{n+1}\right] \leq \frac{2\alpha(n+1) - 1}{n+1}$$

lemma \uparrow

$i \in S(A)$ lemma \square

Examples regression $S(X_i, Y_i; D_{-i}) = |Y_i - \hat{\mu}(X_i; D_{-i})|$

$\hat{p} > \alpha$

$\Leftrightarrow 1 + \sum_{i=1}^n \mathbb{1} \{ \hat{\mu}(X_{n+1}, D_{-i}) - S_i \leq Y_{n+1} \leq \hat{\mu}(X_{n+1}, D_{-i}) + S_i \} > (n+1)\alpha$

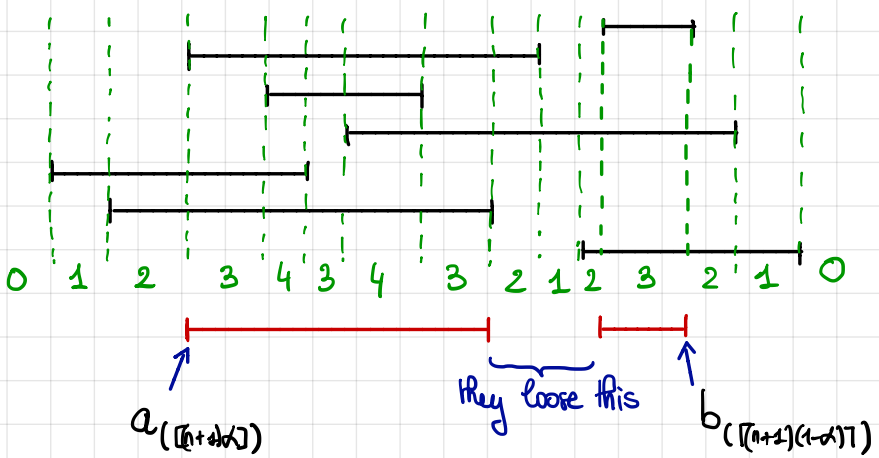
$(n \leq \infty) \Leftrightarrow n \leq \infty$

$\Leftrightarrow \sum_{i=1}^n \mathbb{1} \{ \underbrace{\hat{\mu}(X_{n+1}, D_{-i}) - S_i}_{= a_i} \leq Y_{n+1} \leq \underbrace{\hat{\mu}(X_{n+1}, D_{-i}) + S_i}_{= b_i} \} \geq [(n+1)\alpha]$

$\Rightarrow \left\{ \begin{array}{l} \sum_{i=1}^n \mathbb{1} \{ Y_{n+1} \leq b_i \} \geq [(n+1)\alpha] \\ \sum_{i=1}^n \mathbb{1} \{ a_i \leq Y_{n+1} \} \geq [(n+1)\alpha] \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} Y_{n+1} \leq b_{(\lceil (n+1)(1-\alpha) \rceil)} \\ Y_{n+1} \geq a_{(\lfloor (n+1)\alpha \rfloor)} \end{array} \right.$

[Barber et al (2021)] choose $\hat{L}_\alpha = [a_{(\lfloor (n+1)\alpha \rfloor)}, b_{(\lceil (n+1)(1-\alpha) \rceil)}]$

Illustration $(n+1)\alpha = 3, n = 7, [a_i, b_i], 1 \leq i \leq n$



classification: need to invest $\hat{\Pi}_{X_{n+1}}(y; D_{-i})$ not explicit in general

III Some extensions

① Nested prediction sets [Gupta et al (2022)]

⚠ shape of prediction interval in regression very 'basic' ⚠
 $\hookrightarrow \hat{\mu}(x) \pm q_\alpha$

\rightarrow general approach encompassing many shapes of prediction sets

$\{\mathcal{F}_t(x; \mathcal{D}_{\min})\}_{t \in \mathcal{T}} \subset \mathcal{Y}$ nested sequence of prediction sets
 \downarrow \mathcal{TCR} $\forall t \leq t', \mathcal{F}_{t'}(x; \mathcal{D}_{\min}) \subseteq \mathcal{F}_t(x; \mathcal{D}_{\min})$

score function $S(x, y; \mathcal{D}_{\min}) = \inf \{t \in \mathcal{T} : y \in \mathcal{F}_t(x; \mathcal{D}_{\min})\}$
 \hookrightarrow smallest index of the set capturing y

scores $S_i = S(X_i, Y_i; \mathcal{D}_{\min})$ and usual split conformal (...)
 \rightarrow also works with full / cv conformal (proofs for free!)

Reference	$\mathcal{F}_t(x)$	\mathcal{T}	Estimates
Lei et al. (2018)	$[\hat{\mu}(x) - t, \hat{\mu}(x) + t]$ usual	$[0, \infty)$	$\hat{\mu}$
Lei et al. (2018)	$[\hat{\mu}(x) - t\hat{\sigma}(x), \hat{\mu}(x) + t\hat{\sigma}(x)]$ locally weighted	$[0, \infty)$	$\hat{\mu}, \hat{\sigma}$
Romano et al. (2019)	$[\hat{q}_{\alpha/2}(x) - t, \hat{q}_{1-\alpha/2}(x) + t]$ conformalized quantile	$(-\infty, \infty)$	$\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$
Kivaranovic et al. (2019)	$(1+t)[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)] - t\hat{q}_{1/2}(x)$	$(-\infty, \infty)$	$\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}, \hat{q}_{1/2}$
Sesia and Candès (2020)	$[\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)] \pm t(\hat{q}_{1-\alpha/2}(x) - \hat{q}_{\alpha/2}(x))$	$(-1/2, \infty)$	$\hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}$
Chernozhukov et al. (2019)	$[\hat{q}_t(x), \hat{q}_{1-t}(x)]$	$(0, 1/2)$	$\{\hat{q}_\alpha\}_{\alpha \in [0, 1]}$
Izbicki et al. (2019)	$\{y : \hat{f}(y x) \geq \check{t}_\delta(x)\}^2$	$[0, 1]$	\hat{f}

② Conditional guarantees

→ conditional coverage is always more informative

* conditional on the full training sample D

[Vovk (2012)], [Bian and Barber (2023)], [Liang and Barber (2023)]

$$\text{Let } \alpha_P(D) = \mathbb{P}(Y_{n+1} \notin \hat{C}_\alpha^{\text{split}}(X_{n+1}) \mid D)$$

$$\text{Aim: } \mathbb{P}(\alpha_P(D) \leq \alpha + \varepsilon) \geq 1 - \delta$$

Hoeffding: $\varepsilon = \sqrt{\frac{\log 4/\delta}{2n_{\text{cal}}}}$ works (concentration of $S_{(\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil)}$)

Is such result possible for full conformal/cv?

- not for any predictor
- yes under a 'stability condition' for the predictor

* conditional on the new covariate X_{n+1}

Ideally, the prediction set should be tailored to the covariate X_{n+1}
↳ valid for all individuals with the same covariate X_{n+1}

$$\forall P, \forall n, \mathbb{P}(Y_{n+1} \in \hat{C}_\alpha^{\uparrow}(X_{n+1}) \mid X_{n+1} = x) \stackrel{(*)}{\geq} 1 - \alpha$$

[Vovk (2012)], [Lei and Wasserman (2014)] for P -almost all x

Theorem: if $(*)$ holds, and in the regression case, we have

$\hat{C}_\alpha(\cdot)$ 'degenerated':

$$\forall P, \mathbb{P}_P(|\hat{C}_\alpha(x)| = \infty) \geq 1 - \alpha \text{ for } P_{X_{n+1}} \text{ almost all } x \text{ not in the atoms of } P_{X_{n+1}}$$

Proof: if $\exists P: |\hat{C}_\alpha| < +\infty$ with $\mu_{P_{X_{n+1}}} \geq \alpha$ build P' with γ lower contradicting $(*)$

[Barber et al. (2020)] with a weaker aim only possible to say trivial things

Definition (conditional coverage)

$\hat{C}_{\alpha, S}$ does $(1-\alpha, S)$ -CC, if

$$\mathbb{P}(Y_{t+1} \in \hat{C}_{\alpha, S}(X_{t+1}) \mid X_{t+1} \in \mathcal{X}) \geq 1-\alpha$$

for all P and $\mathcal{X} \subset \mathbb{R}^d$ with $P_{\mathcal{X}}(\mathcal{X}) \geq S$

Theorem: $\forall \hat{C}_{\alpha, S}$ satisfying $(1-\alpha, S)$ -CC, we have

$$\mathbb{E}[\Lambda(\hat{C}_{\alpha, S}(X_{t+1}))] \geq (1-\alpha) \inf_{\hat{C}} \mathbb{E}[\Lambda(\hat{C})]$$

\hat{C} with marginal cov $\geq 1-\alpha S$

↳ marginal cov at $1-\alpha S$ is essentially the best that you can do

↳ if (*) then $\mathbb{E}[\Lambda(\hat{C}_{\alpha}(X_{t+1}))] \geq \sup_S \{ \dots \}$ large!

Definition (group conditional coverage)

\hat{C}_{α} such that $\mathbb{P}(Y_{t+1} \in \hat{C}_{\alpha}(X_{t+1}) \mid X_{t+1} \in G) \geq 1-\alpha$

for all P and $G \in \mathcal{G}$ (subsets of \mathbb{R}^d)

[Vovk et al. (2019)] Mondrian conformal prediction \rightarrow non overlapping groups

[Jung et al. (2023)], [Gibbs et al. (2023)] More general solutions

* conditional on the new label Y_{n+1}

Investigated in classification \rightarrow 'class specific coverage'

⚠ classical coverage provide more coverage for frequent classes ⚠

if $Y = \{ \text{Common Cold, Flu, Covid} \}$ maybe always predict
 $\hat{Y}_x = \{ \text{Common cold, Flu} \}$ has correct coverage

[Lei (2014)], [Sadinle et al. (2018)], [Derhacopian et al (202?)]

$$\forall P, \forall y \in Y, \mathbb{P}(Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) | Y_{n+1} = y) \geq 1 - \alpha$$

- all classes treated equally, so 'fair'
- kind of frequentist guarantee because unknown quantity y is fixed
 \approx confidence region for the parameter y

Classical solution is to calibrate by separating the possible labels

③ Beyond exchangeability

* Covariate shift (= domain shift)

What if we calibrate with wrong data? calibrate with 'dogs'
evaluate on 'cats'

[Tibshirani et al. (2019)]

$$\begin{cases} (X_i, Y_i) \text{ iid } \sim P = P_X \times P_{Y|X} \\ (X_{n+1}, Y_{n+1}) \sim \tilde{P} = \tilde{P}_X \times P_{Y|X} \end{cases} \quad \rightarrow \text{covariate shift}$$

Solution is to use weighted quantile / p-values with $w(x) = \frac{d\tilde{P}_X(x)}{dP_X(x)}$
 \rightarrow to compensate the covariate shift

Namely $\{Y_{n+1} \in \hat{\mathcal{C}}_\alpha^{\text{wsplit}}(X_{n+1})\} = \{\hat{p}_w > \alpha\}$

with
$$\hat{p}_w = \frac{\omega(X_{n+1}) + \sum_{i \in \mathcal{I}_{cal}} \omega(X_i) \mathbb{1}\{S_i \geq S_{n+1}\}}{\sum_{i \in \mathcal{I}_{cal}} \omega(X_i) + \omega(X_{n+1})}$$

$\downarrow S(X_i, Y_i; P_{train})$
 $\downarrow S(X_{n+1}, Y_{n+1}; P_{train})$

→ still provide coverage $1 - \alpha$ (weighted exchangeability)

→ if weights unknown and estimated, remainder term [Jim and Cambas (2023)]

* 'Almost' exchangeable core [Barber et al. (2023)]

• (X_i, Y_i) 's indep but $d_{TV}(\mathcal{L}(X_i, Y_i), \mathcal{L}(X_{n+1}, Y_{n+1}))$ not zero

• weights $\{(\omega_i)_{i \in \mathcal{I}_{cal}}\}_{\omega_{n+1} = 1}$ fixed in $[0, 1]^{n_{cal}}$, $\tilde{\omega}_i = \frac{\omega_i}{\sum_{i \in \mathcal{I}_{cal}} \omega_i + 1}$ $i \in \mathcal{I}_{cal} \cup \{n+1\}$

• weighted split conformal as above with

$$\hat{p}_w = \tilde{\omega}_{n+1} + \sum_{i \in \mathcal{I}_{cal}} \tilde{\omega}_i \mathbb{1}\{S_i \geq S_{n+1}\}$$

weights to flip the d_{TV}

Theorem:

$$\mathbb{P}(Y_{n+1} \in \hat{\mathcal{C}}_\alpha^{\text{wsplit}}(X_{n+1})) \geq 1 - \alpha - 2 \sum_{i=1}^n \tilde{\omega}_i d_{TV}(\mathcal{L}(X_i, Y_i), \mathcal{L}(X_{n+1}, Y_{n+1}))$$

↪ remainder

Example: $d_{TV}(\mathcal{L}(X_i, Y_i), \mathcal{L}(X_{n+1}, Y_{n+1})) \leq \varepsilon (n+1-i)$ 'time shift'

$$\omega_i = \rho^{n+1-i}$$

$$\Rightarrow \text{remainder} \leq 2\varepsilon \sum_{i=1}^n (n+1-i) \rho^{n+1-i} = 2\varepsilon \sum_{i=1}^n i \rho^i \leq \frac{2\varepsilon \rho}{(1-\rho)^2}$$

④ Transductive approach [Vovk (2013)]

[Bates et al (2023)], [Marandon et al (2023)], [Gajin et al (2023)]

$D_{\text{test}} = ((X_{n+1}, Y_{n+1}), \dots, (X_{n+m}, Y_{n+m}))$ test sample of size m
 observed \rightarrow not observed

Multiple prediction sets $(\hat{C}_{i,t})_{1 \leq i \leq m}$ such that

$$\text{FCP}(t) = \frac{1}{m} \sum_{i=1}^m \mathbb{1} \{ Y_{n+i} \notin \hat{C}_{i,t} \}$$

is 'small' in probability

$$= \frac{1}{m} \sum_{i=1}^m \mathbb{1} \{ \hat{p}_i \leq t \}$$

to study

Split conformal: $\hat{p}_i = \frac{1}{n_{\text{cal}}+1} \left(1 + \sum_{j \in I_{\text{cal}}} \mathbb{1} \{ S_j \geq S_{n+i} \} \right)$, $1 \leq i \leq m$

$(\hat{p}_i, 1 \leq i \leq m)$

- positive dependence [Bates et al (2023)]
- full distribution known } [Gajin et al (2023)]
- concentration see also [Kasperov (2023)]

Theorem (DKW type concentration) Assume noises in S_i 's a.s.

$$\mathbb{P} \left(\sup_{t \in [0,1]} (\text{FCP}(t) - t) > \lambda \right) \leq \left(1 + \frac{2\sqrt{2\pi} \tau}{\sqrt{n_{\text{cal}} + m}} \right) e^{-2\tau \lambda^2}$$

for $\tau := \frac{n_{\text{cal}} m}{n_{\text{cal}} + m} \in \left[\frac{1}{2} n_{\text{cal}} \wedge m, n_{\text{cal}} \wedge m \right]$ and $\lambda > 0$

\Rightarrow with proba $\geq 1 - \delta$, $\forall \alpha$, $\text{FCP}(\alpha) \leq \alpha + \underbrace{\text{remainder}}_{\approx \sqrt{\frac{\log \tau}{\tau}}}$

References: (order of appearance)

- [Angelopoulos and Bates (2022)] A gentle introduction to conformal prediction (...)
- [Papadopoulos et al. (2002)] Inductive confidence machines for regression
- [Romano and Wolf (2005)] Exact and approximate stopdown methods for MT
- [Arlot et al. (2010)] Some nonasympt results on resampling in high dim (I)
- [Vovk et al. (2005)] Algorithmic learning in a random world
- [Nouretdinov et al (2002)] Ridge regression confidence machine
- [Lei (2017)] Fast exact conformalization of lasso using piecewise linear (...)
- [Barber et al (2021)] Predictive inference with the jackknife +
- [Gupta et al (2022)] Nested conformal prediction and quantile out of bags (...)
- [Vovk (2012)] Conditional validity of inductive conformal predictor
- [Bian and Barber (2023)] Training-conditional coverage for distribution free (...)
- [Liang and Barber (2023)] Algo stability implies training-conditional coverage (...)
- [Lei and Wasserman (2014)] Distribution free prediction bands for NP regression
- [Barber et al. (2020)] The limits of distribution-free conditional predictive inf.
- [Vovk et al. (2013)] Mondrian confidence machine
- [Jung et al. (2023)] Batch multivalued conformal prediction
- [Gibbs et al. (2023)] Conformal prediction with conditional guarantees
- [Lei (2014)] Classification with confidence
- [Sadinle et al. (2018)] Least ambiguous set-valued classifier with bounded error levels

- [Derfascobian et al (202?)] Adaptive prediction sets with class cond. cov.
- [Tibshirani et al. (2019)] Conformal prediction under covariate shift
- [Sin and Candès (2023)] Model-free selective inf. under cov. shift
via weighted conformal p-values
- [Barber et al. (2023)] Conformal prediction beyond exchangeability
- [Vovk (2013)] Transductive conformal predictor
- [Pates et al (2023)] Testing for outliers with conformal p-values
- [Marandon et al (2023)] Adaptive novelty detection with FDR guarantees
- [Gayin et al (2023)] Transductive conformal inference with adaptive scores
- [Haugens (2023)] On the universal distribution of the coverage in split conformal prediction