An introduction to conformal prediction and distribution-free inference

CIRM tutorial (part 2)

Rina Foygel Barber (University of Chicago)
CIRM December 2024

http://rinafb.github.io/

Introduction

The conformal framework

Overview of Part 1:

- Conformal allows us to start with any algorithm,
  & calibrate it to achieve (marginal) predictive coverage
- Tradeoff between statistical & computational efficiency:
  Split CP, full CP, and CV-based versions
- Conformal + model-based methods $\rightsquigarrow$ "best of both worlds"

The conformal framework

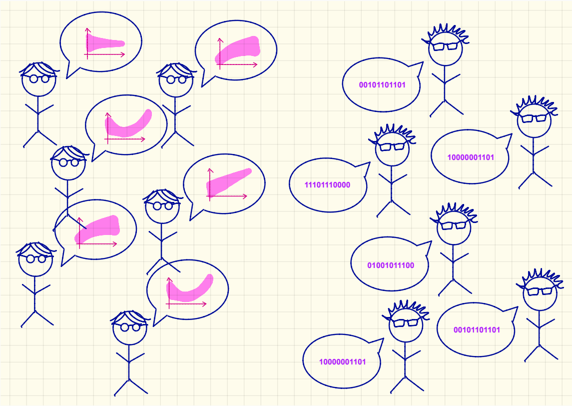
Part 2 will examine extensions:

- Beyond marginal coverage — conditional coverage guarantees
- Beyond the i.i.d. assumption — the streaming-data setting

Conditional coverage

Is marginal coverage enough?

Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1})\right\} \geq 1 - \alpha$

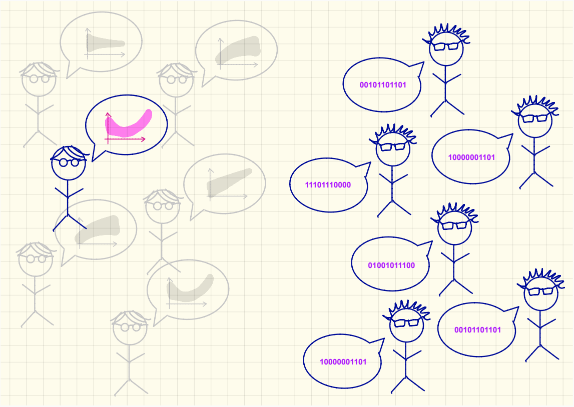


each researcher samples
a training data set,
and constructs $\mathcal{C}$

each individual in test set
samples features X
& asks for prediction

Is marginal coverage enough?

Training-conditional coverage: $\mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \;\middle|\; \{(X_i, Y_i)\}_{i\in[n]} \right\} \geq 1-\alpha$

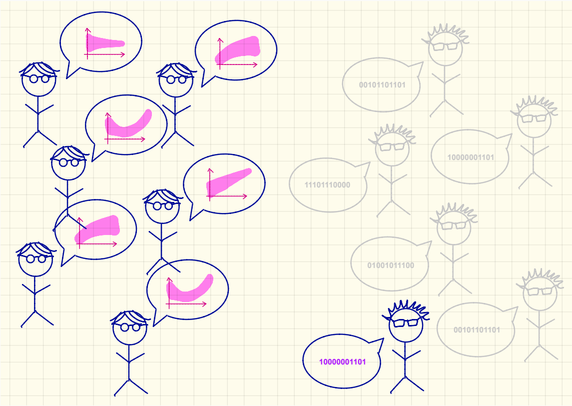


for this researcher...

...what will be the coverage
on average over the test set?

Is marginal coverage enough?

Test-conditional coverage: $\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1}\right\} \geq 1-\alpha$?

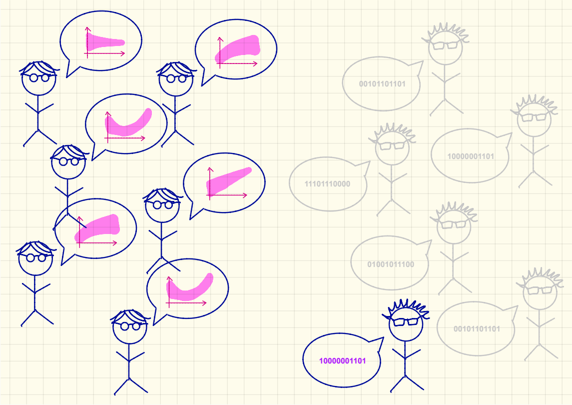


what will be the coverage on average over a draw of the training data...

...for this test individual?

Is marginal coverage enough?

Label-~~Test~~-conditional coverage: $\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid \cancel{X_{n+1}}\, Y_{n+1}\right\} \geq 1-\alpha$?



$\underbrace{\hspace{6em}}$ what will be the coverage on average over a draw of the training data...

$\underbrace{\hspace{6em}}$ ...for this test individual?

Conditioning on the test data

The marginal coverage guarantee: $\underbrace{\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1})\right\}}_{\text{averaged over training + test data}} \geq 1 - \alpha$

How can we avoid the following scenario?

- Coverage is 90% on average
- But, coverage for patients $>$ 65 years old, is only 10%
- Or, coverage for patients with poor outcomes, is only 10%

Hardness of test-conditional coverage

Let $\mathcal{C}$ be *any* procedure satisfying test-conditional coverage,

$$\mathbb{P}_P\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1}\right\} \geq 1-\alpha \text{ almost surely, for all } P$$

$(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}) \overset{\text{iid}}{\sim} P$

Theorem[1,2]

Let P be any distribution with a marginal P_X that is nonatomic.

Then, if $\mathcal{Y} = \mathbb{R}$,

$\mathbb{P}_P\{X = x\} = 0$ for all $x \in \mathcal{X}$

$$\mathbb{E}\left[\text{length}(\mathcal{C}(X_{n+1})\right] = \infty.$$

[1] Vovk 2012, *Conditional validity of inductive conformal predictors*

[2] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Hardness of test-conditional coverage

Let $\mathcal{C}$ be *any* procedure satisfying test-conditional coverage.

Key lemma[3,4]

Let P be any distribution with a marginal P_X that is nonatomic. Then for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathbb{P}\{y \in \mathcal{C}(x)\} \geq 1 - \alpha.$$

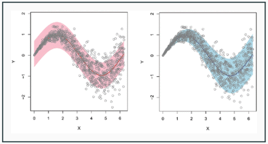
[3] Vovk 2012, *Conditional validity of inductive conformal predictors*

[4] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

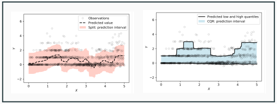
Empirical vs theoretical

For all conformal methods so far (split CP/full CP/jack+/...)

- Distribution-free marginal coverage theory for *any* score s
- Distribution-free conditional coverage is impossible for *any* score s (when X is nonatomic)
- Empirically, the choice of s has substantial impact on conditional coverage



(figure from Lei et al 2018)



(figure from Romano et al 2019)

Empirical vs theoretical

Aim: to find a relaxation of test-conditional coverage that...

- Is interpretable & meaningful
- Is possible to achieve distribution-free (& without high computational cost)
- Does not lead to overly conservative methods in the continuous case

Approximate test-conditional coverage

$(1-\alpha, \delta)$-conditional coverage[5]

For any distribution P & any $\mathcal{X}_0 \subseteq \mathcal{X}$ with $P_X(\mathcal{X}_0) \geq \delta$,

$$\mathbb{P}_P\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_0\right\} \geq 1-\alpha.$$

Intuition: no large regions in feature space with low coverage

[5] B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*

Approximate test-conditional coverage

$(1-\alpha, \delta)$-conditional coverage[5]

For any distribution P & any $\mathcal{X}_0 \subseteq \mathcal{X}$ with $P_X(\mathcal{X}_0) \geq \delta$,

$$\mathbb{P}_P\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_0\right\} \geq 1-\alpha.$$

Intuition: no large regions in feature space with low coverage

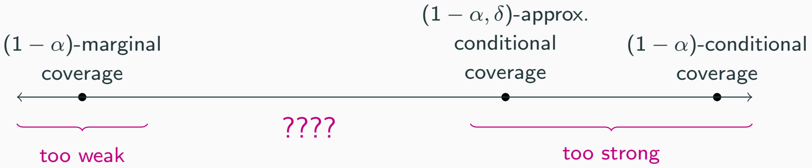
Trivial solutions, e.g.,

- Any method with $1-\alpha\delta$ marginal coverage (e.g., split CP)

Theorem—any $\mathcal{C}$ satisfying $(1-\alpha, \delta)$-conditional coverage, returns intervals at least as large as a trivial solution

[5] B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*

Approximate test-conditional coverage



Bin-conditional coverage

A possible relaxation — coverage conditional on bins:[6,7,8]

Partition $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$,

& require $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_k\} \geq 1 - \alpha$ for each k

[6] Vovk 2012, *Conditional validity of inductive conformal predictors*

[7] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

[8] Vovk et al 2005, *Algorithmic Learning in a Random World*

Bin-conditional coverage

A possible relaxation — coverage conditional on bins:[6,7,8]

Partition $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$,

& require $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_k\} \geq 1 - \alpha$ for each k

- For each k, data points $\{(X_i, Y_i) : X_i \in \mathcal{X}_k\}$ are exchangeable
- $\rightsquigarrow$ can run CP separately for each bin k,
  to guarantee coverage conditional on $X_{n+1} \in \mathcal{X}_k$

[6] Vovk 2012, *Conditional validity of inductive conformal predictors*

[7] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

[8] Vovk et al 2005, *Algorithmic Learning in a Random World*

Bin-conditional coverage

A best-of-both-worlds guarantee:[9]

- The distribution-free guarantee:

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_k\} \geq 1 - \alpha, \ \forall k$$

[9] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Bin-conditional coverage

A best-of-both-worlds guarantee:[9]

- The distribution-free guarantee:

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_k\right\} \geq 1-\alpha, \ \forall k$$

- If bins have vanishing diameter, + additional assumptions (e.g., continuity of $x \mapsto$ (distrib. of $Y \mid X = x$)):

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} = x\right\} \to 1-\alpha, \ \forall x$$

[9] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Localized conformal prediction

A different relaxation — localized guarantees, i.e., conditions of the type

$$\mathbb{P}_P\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \approx x\right\} \gtrapprox 1 - \alpha$$

e.g., coverage conditional on $X_{n+1} \in \mathbb{B}(x, r_n)$, where $r_n \to 0$

Localized conformal prediction

A possible approach?

To guarantee $\approx$ coverage over balls $X_{n+1} \in \mathbb{B}(x, r_n)$...

compute $\widehat{q}$ using only calibration points $\|X_i - X_{n+1}\| \leq r_n$?

Localized conformal prediction

A possible approach?

To guarantee $\approx$ coverage over balls $X_{n+1} \in \mathbb{B}(x, r_n)$...
compute $\widehat{q}$ using only calibration points $\|X_i - X_{n+1}\| \leq r_n$?

$$\mathcal{C}(X_{n+1}) = \left\{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \mathrm{Quantile}_{(1-\alpha)(1+1/|\mathcal{I}_n|)}\left(\{S_i\}_{i \in \mathcal{I}_n}\right)\right\}$$

where

$$\mathcal{I}_n = \{i : n_0 < i \leq n, \|X_i - X_{n+1}\| \leq r_n\}$$

Localized conformal prediction

What we expect:

- Marginal coverage (like any conformal method)
- & approx. conditional coverage (maybe need smoothness?)

[10] Guan 2023, *Localized conformal prediction: A generalized inference framework for conformal prediction*

Localized conformal prediction

What we expect:

- Marginal coverage (like any conformal method)
- & approx. conditional coverage (maybe need smoothness?)

What we see (in the worst case):[10]

- Even marginal coverage can fail!

[10] Guan 2023, *Localized conformal prediction: A generalized inference framework for conformal prediction*

Localized conformal prediction

The LCP method[11]

1. Construct score function s using pretraining data $Z_1, \ldots, Z_{n_0}$
2. Let $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a kernel, and define weights

$$w_i = \frac{H(X_{n+1}, X_i)}{\sum_{j=n_0+1}^{n+1} H(X_{n+1}, X_j)}$$

3. Compute weighted quantile $\widehat{q}_\alpha$ at level $1 - \alpha$
4. For test point $n + 1$ return prediction interval

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \widehat{q}_\alpha \ \}$$

[11] Guan 2023, *Localized conformal prediction: A generalized inference framework for conformal prediction*

Localized conformal prediction

The LCP method[11]**—with recalibration step**

1. Construct score function s using pretraining data $Z_1, \ldots, Z_{n_0}$
2. Let $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a kernel, and define weights

$$w_i = \frac{H(X_{n+1}, X_i)}{\sum_{j=n_0+1}^{n+1} H(X_{n+1}, X_j)}$$

3. Compute weighted quantile $\widehat{q}_\alpha$ at level $\cancel{1-\alpha}$ $1 - \widetilde{\alpha}(y)$
4. For test point $n+1$ return prediction interval

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq \widehat{q}_{\widetilde{\alpha}(y)} \right\}$$

[11] Guan 2023, *Localized conformal prediction: A generalized inference framework for conformal prediction*

Localized conformal prediction

Examples of the kernel H:

- Box kernel: $H(x, x') = \mathbb{1}_{\|x-x'\| \leq h_n}$
- Exponential kernel: $H(x, x') = e^{-\|x-x'\|/h_n}$
- Gaussian kernel: $H(x, x') = e^{-\|x-x'\|^2/2h_n^2}$

Randomly-localized conformal prediction

RLCP[12]

1. Construct score function s using pretraining data $Z_1, \ldots, Z_{n_0}$
2. Sample $\widetilde{X}_{n+1} \sim H(X_{n+1}, \cdot)$
3. Define weights

$$\widetilde{w}_i = \frac{H(X_i, \widetilde{X}_{n+1})}{\sum_{j=n_0+1}^{n+1} H(X_j, \widetilde{X}_{n+1})}$$

4. Compute weighted quantile $\widehat{q}$ at level $1 - \alpha$
5. For test point $n+1$ return prediction interval

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \leq \widehat{q}\}$$

[12] Hore & B. 2023 *Conformal prediction with local weights: randomization enables robust guarantees*

Theoretical guarantees for RLCP

Theorem: marginal coverage for RLCP

For the RLCP method,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$$

Theoretical guarantees for RLCP

Theorem: marginal coverage for RLCP

For the RLCP method,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$$

The marginal coverage theorem follows from:

Theorem: key property of RLCP

For the RLCP method,

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \;\middle|\; \widetilde{X}_{n+1}\right\} \geq 1 - \alpha$$

Theoretical guarantees for RLCP

Returning to the goal of $\approx$ test-conditional coverage...

Theorem: asymptotic local coverage for RLCP

For RLCP with $H(x, x') \propto \mathbb{1}_{\|x-x'\| \leq h_n}$, if $h_n, r_n \to 0$, $h_n/r_n \to 0$,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in \mathbb{B}(x, r_n)\} \geq 1 - \alpha - \mathrm{o}(1)$$

as long as P_X has a density which is continuous and positive at x

Coverage relative to a class of functions

Another relaxation — require coverage with respect to a class of test functions[13]

For a class of functions $\mathcal{F} = \{f : \mathcal{X} \to [0, \infty)\}$, require

$$\mathbb{E}\left[f(X_{n+1}) \cdot \left(\mathbb{1}_{Y_{n+1} \in \mathcal{C}(X_{n+1})} - (1 - \alpha)\right)\right] \geq 0 \text{ for all } f \in \mathcal{F}$$

- Test-conditional coverage $\leftrightarrow$ all measurable functions
- Marginal coverage $\leftrightarrow$ one function, $f(x) \equiv 1$
- Bin-conditional coverage $\leftrightarrow$ functions $f(x) = \mathbb{1}_{x \in \mathcal{X}_k}$

[13] Gibbs et al 2023, *Conformal Prediction With Conditional Guarantees*

Conformal prediction in the online setting

Prediction in a streaming setting

Conformal prediction is often studied for a single test point Z_{n+1}.

In practice, we want to predict "in real time":

$$Z_1 \quad Z_2 \quad Z_3 \quad \dots \quad Z_{t-1} \quad \underbrace{Z_t}_{\text{predict } Y} \quad Z_{t+1} \quad \dots$$

$$Z_1 \quad Z_2 \quad Z_3 \quad \dots \quad Z_{t-1} \quad Z_t \quad \underbrace{Z_{t+1}}_{\text{predict } Y} \quad Z_{t+2} \quad \dots$$

$$Z_1 \quad Z_2 \quad Z_3 \quad \dots \quad Z_{t-1} \quad Z_t \quad Z_{t+1} \quad \underbrace{Z_{t+2}}_{\text{predict } Y} \quad Z_{t+3} \quad \dots$$

$$\vdots$$

Prediction in a streaming setting

$\mathcal{C}_t$ = the prediction interval constructed for prediction at time t

Conformal method guarantee for each t:

$$\mathbb{P}\{Y_t \in \mathcal{C}_t(X_t)\} \geq 1 - \alpha$$

Is this sufficient for practical purposes?

Prediction in a streaming setting

$\mathcal{C}_t$ = the prediction interval constructed for prediction at time t

Conformal method guarantee for each t:

$$\mathbb{P}\{Y_t \in \mathcal{C}_t(X_t)\} \geq 1 - \alpha$$

Is this sufficient for practical purposes?

A high-dependence scenario... what if:

- with probability $1 - \alpha$, for all t, $Y_t \in \mathcal{C}_t(X_t)$
- with probability α, for all t, $Y_t \notin \mathcal{C}_t(X_t)$

Conformal prediction as a hypothesis test

Recall construction of full conformal prediction:

Test value $y \in \mathcal{Y}$

- Fit model to training+test data
$$\widehat{\mu}^y = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$S_i^y = |Y_i - \widehat{\mu}^y(X_i)|, \; i = 1, \ldots, n, \quad S_{n+1}^y = |y - \widehat{\mu}^y(X_{n+1})$$
- Check if $S_{n+1}^y \leq \mathrm{Quantile}_{(1-\alpha)(1+1/n)}(S_1^y, \ldots, S_n^y)$

$y \rightsquigarrow \{\mathrm{Yes}, \mathrm{No}\}$

if Yes: add y to $\mathcal{C}(X_{n+1})$ if No: discard y

Conformal prediction as a hypothesis test

Recall construction of full conformal prediction:

Test value $y \in \mathcal{Y}$

- Fit model to training+test data
$$\widehat{\mu}^y = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$S_i^y = |Y_i - \widehat{\mu}^y(X_i)|, \ i = 1, \ldots, n, \quad S_{n+1}^y = |y - \widehat{\mu}^y(X_{n+1})$$
- Check if $S_{n+1}^y \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1^y, \ldots, S_n^y)$

$y \rightsquigarrow \{\text{Yes, No}\}$

if Yes: add y to $\mathcal{C}(X_{n+1})$ if No: discard y

Reinterpret this as a hypothesis test:

$H_{0,y}$: Data points $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ are exchangeable

Conformal prediction as a hypothesis test

Full conformal prediction (p-value version)[14]

For each $y \in \mathcal{Y}$ define a conformal p-value:

$$p(y) = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\{S_i^y \geq S_{n+1}^y\}}{n+1}$$

where

$$S_i^y = s^y(X_i, Y_i), i = 1, \ldots, n, \quad S_{n+1}^y = s^y(X_{n+1}, y),$$

for fitted score function $s^y = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\big)$

Then define prediction interval: $\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : p(y) > \alpha\}$

[14] Vovk et al 2005, *Algorithmic Learning in a Random World*

The conformal p-value

The marginal coverage guarantee (under exchangeability):

$$\mathbb{P}\left\{Y_{n+1} \notin \mathcal{C}(X_{n+1})\right\} = \underbrace{\mathbb{P}\left\{p(Y_{n+1} \leq \alpha\right\} \leq \alpha}_{\text{i.e., } p(Y_{n+1}) \text{ is a valid p-value}}$$

Conformal p-values are independent

Theorem: conformal p-values in streaming time[15]

Assume scores $\{s_t(Z_i)\}_{i=1,\ldots,t}$ are distinct at each t (no ties).

Then:

- For each t, $p_t \sim \text{Uniform}\{\frac{1}{t}, \frac{2}{t}, \ldots, 1\}$
- And, $p_1, p_2, \ldots$ are mutually independent

Implication:

$\sum_{i=1}^t \mathbb{1}\{Y_i \notin \mathcal{C}_i(X_i)\} = \sum_{i=1}^t \mathbb{1}\{p_i \leq \alpha\} \leq \text{Binomial}(t, \alpha)$

(& this also holds if ties allowed)

$\rightsquigarrow$ the high-dependence scenario cannot occur

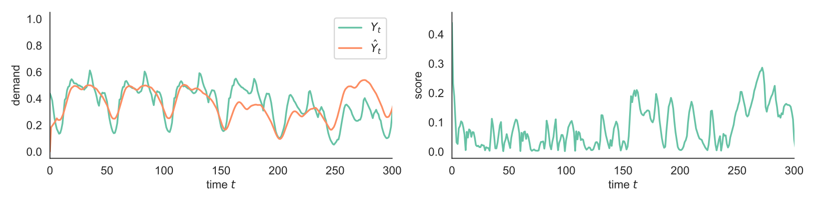
[15] Vovk et al 2003, *Testing Exchangeability On-Line*

The online setting: arbitrary data streams

Removing the exchangeability assumption

The conformal framework allows us to use *any* model,
while ensuring validity with respect to *any* distribution....

But, we assume the data is i.i.d. (or exchangeable)—
does not allow for drift, dependence, changepoints, ...



(figure shows `Elec` data set[16] — tracking electricity demand in Australia)

[16] Harries 1999, *Splice-2 comparative evaluation: Electricity pricing*

Removing the exchangeability assumption

Without assuming exchangeability (or a bounded/known violation) impossible to guarantee coverage at a fixed time t:

If we observe $\Big((X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}), X_t\Big)$, the i.i.d. setting

$$(X_1, Y_1), \ldots, (X_t, Y_t) \overset{\text{iid}}{\sim} P$$

is indistinguishable from

$$(X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1}) \overset{\text{iid}}{\sim} P,\ (X_t, Y_t) \sim P_X \times Q_{Y|X}$$

Removing the exchangeability assumption

To make guarantees possible—relax the notion of valid coverage:

Coverage holds at every fixed time t	$\longrightarrow$	Coverage holds on average over all times t

- If a changepoint at time t causes the method to lose coverage, can compensate by being more conservative at later times $t' > t$ to maintain average coverage

Adaptive conformal inference

Method: adaptive conformal inference[17,18]

At each time t, via conformal or some other method, construct

1. A score function $s_t(x, y)$
2. Estimated quantiles $\widehat{q}_t(1-a)$ for $s_t(X, Y)$, for $a \in [0, 1]$

To allow values $a \notin [0, 1]$ define

$$\widehat{q}_t(1-a) = +\infty, \quad a < 0$$
$$\widehat{q}_t(1-a) = -\infty, \quad a > 1$$

[17] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[18] Gibbs & Candès 2022, *Conformal Inference for Online Prediction with Arbitrary Distribution Shifts*

Adaptive conformal inference

Adaptive conformal inference[19,20]

1. Initialize at some $\alpha_1 \in [0,1]$, and return

$$\mathcal{C}_1(X_1) = \{y \in \mathcal{Y} : s_1(X_1, y) \leq \widehat{q}_1(1 - \alpha_1)\}$$

[19] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[20] Gibbs & Candès 2022, *Conformal Inference for Online Prediction with Arbitrary Distribution Shifts*

Adaptive conformal inference

Adaptive conformal inference[19,20]

1. Initialize at some $\alpha_1 \in [0,1]$, and return
$$\mathcal{C}_1(X_1) = \{y \in \mathcal{Y} : s_1(X_1, y) \leq \widehat{q}_1(1-\alpha_1)\}$$
2. For each $t \geq 1$, update
$$\alpha_{t+1} = \alpha_t - \eta(\mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha)$$
and return
$$\mathcal{C}_{t+1}(X_{t+1}) = \{y \in \mathcal{Y} : s_{t+1}(X_{t+1}, y) \leq \widehat{q}_{t+1}(1-\alpha_{t+1})\}$$

[19] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[20] Gibbs & Candès 2022, *Conformal Inference for Online Prediction with Arbitrary Distribution Shifts*

Adaptive conformal inference

How ACI maintains coverage over time—intuition:

- If we undercover over a long stretch of time,

  α_t will decrease to compensate
- If we overcover over a long stretch of time,

  α_t will increase to compensate

Adaptive conformal inference

Lemma: bounded thresholds

For all $t \geq 1$,

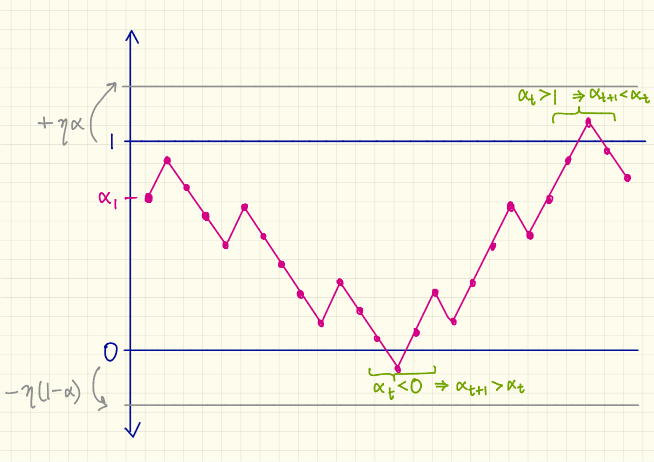
$$-\eta(1-\alpha) \leq \alpha_t \leq 1 + \eta\alpha$$

Adaptive conformal inference

Lemma: bounded thresholds

For all $t \geq 1$,

$$-\eta(1-\alpha) \leq \alpha_t \leq 1 + \eta\alpha$$



Adaptive conformal inference

Theorem: regret bound[21]

For any initial threshold $\alpha_1 \in [0, 1]$,

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha \right| \leq \frac{\max\{\alpha_1, 1 - \alpha_1\} + \eta}{\eta T}$$

This is a deterministic result:

- Data may have any distribution (or may be nonrandom)
- Score functions s_t may be fixed or arbitrarily data-dependent

[21] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

Proof of theorem:

By def. of update rule,

$$\alpha_{T+1} = \alpha_1 - \sum_{t=1}^{T} \eta(\mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha)$$

Adaptive conformal inference

Proof of theorem:

By def. of update rule,

$$\alpha_{T+1} = \alpha_1 - \sum_{t=1}^{T} \eta(\mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha)$$

Rearranging terms,

$$\sum_{t=1}^{T} \mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} = T\alpha + \eta^{-1} \cdot \underbrace{(\alpha_{T+1} - \alpha_1)}_{\text{bounded by Lemma}}$$

Adaptive conformal inference

The update can be on any tuning parameter—can update the thresholds directly

Quantile tracker[22]

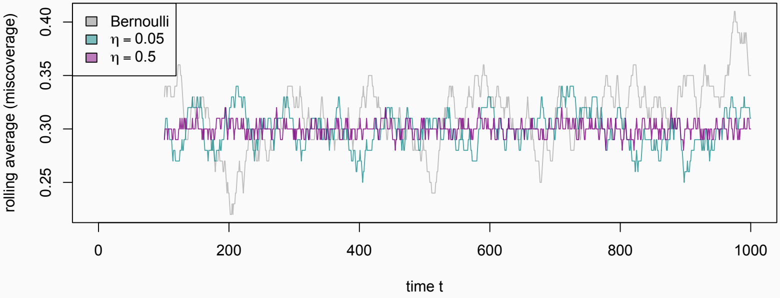
1. Assume all score functions return output in $[0, B]$
2. Initialize at some $q_1 \in [0, B]$, and return
$$\mathcal{C}_1(X_1) = \{y \in \mathcal{Y} : s_1(X_1, y) \leq q_1\}$$
3. For each $t \geq 1$, update
$$q_{t+1} = q_t + \eta(\mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha)$$
and return
$$\mathcal{C}_{t+1}(X_{t+1}) = \{y \in \mathcal{Y} : s_{t+1}(X_{t+1}, y) \leq q_{t+1}\}$$

[22] Angelopoulos et al 2023, *Conformal PID Control for Time Series Prediction*

Reconsidering a constant step size

What is the effect of using a constant step size η?

- Constant $\eta > 0$ ensures rapid corrections for undercoverage
- However, also *overcorrects* for errors that occur simply by random chance



Reconsidering a constant step size

Theorem: variability with a constant step size[23]

Assume $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for any P, and $\underbrace{\text{scores are trained online}}_{s_t \text{ may depend on } (X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})}$.

[23] Angelopoulos, B., & Bates 2024, *Online conformal with decaying step size*

Reconsidering a constant step size

Theorem: variability with a constant step size[23]

Assume $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for any P, and $\underbrace{\text{scores are trained online}}_{s_t \text{ may depend on } (X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})}$.

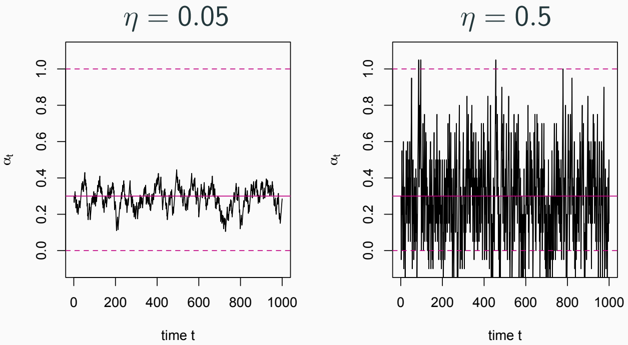
If scores $s_t(X_t, Y_t)$ are continuous & $\alpha \in \mathbb{Q}$,

$$\liminf_{T\to\infty} \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\alpha_t \le 0}}_{\text{how often } \mathcal{C}_t(X_t) = \mathcal{Y}} > 0 \text{ and } \liminf_{T\to\infty} \underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\alpha_t \ge 1}}_{\text{how often } \mathcal{C}_t(X_t) = \varnothing} > 0$$

[23] Angelopoulos, B., & Bates 2024, *Online conformal with decaying step size*

Reconsidering a constant step size

Illustration: the oracle setting



Reconsidering a constant step size

Online conformal inference with time-varying step size[24]

1. Initialize at some $q_1 \in [0, B]$, and return

$$\mathcal{C}_1(X_1) = \{y \in \mathcal{Y} : s_1(X_1, y) \leq q_1\}$$

2. For each $t \geq 1$, update

$$q_{t+1} = q_t + \eta_t \cdot (\mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha)$$

and return

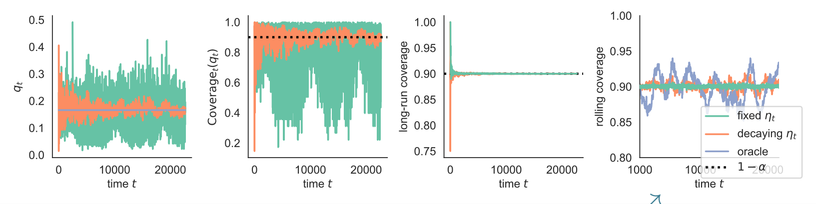
$$\mathcal{C}_{t+1}(X_{t+1}) = \{y \in \mathcal{Y} : s_{t+1}(X_{t+1}, y) \leq q_{t+1}\}$$

[24] Angelopoulos, B., & Bates 2024, *Online conformal with decaying step size*

Experiments

`Elec` data set[25] (time series)

- Prediction interval constructed with residual score $|Y_t - \widehat{Y}_t|$
- Prediction $\widehat{Y}_t$ given by average of data from 24–48 hours ago



rolling average (window = 1000)

[25] Harries 1999, *Splice-2 comparative evaluation: Electricity pricing*

Assumption-free theory

Theorem: regret bound (for decreasing η_t)

Let $s_1, s_2, \cdots : \mathcal{X} \times \mathcal{Y} \to [0, B]$, and $q_1 \in [0, B]$.

If $\eta_1 \geq \eta_2 \geq \cdots > 0$, then

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha \right| \leq \frac{B + \eta_1}{\eta_T T}$$

Assumption-free theory

Theorem: regret bound (for decreasing η_t)

Let $s_1, s_2, \dots : \mathcal{X} \times \mathcal{Y} \to [0, B]$, and $q_1 \in [0, B]$.

If $\eta_1 \geq \eta_2 \geq \dots > 0$, then

$$\left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{Y_t \notin \mathcal{C}_t(X_t)\} - \alpha \right| \leq \frac{B + \eta_1}{\eta_T T}$$

- Allowing η_t to increase if needed can accelerate adaptivity to changes (theory extends to this case)

Theory for i.i.d. data

The regret bounds hold for

- *any* data stream (random or deterministic)
- *any* sequence of bounded s_t's (random or deterministic)

In the i.i.d. setting, can give stronger results:

Theory for i.i.d. data

The regret bounds hold for

- *any* data stream (random or deterministic)
- *any* sequence of bounded s_t's (random or deterministic)

In the i.i.d. setting, can give stronger results:

Theorem: the i.i.d. data setting

Let $(X_t, Y_t) \overset{\text{iid}}{\sim} P$ for any P, s_t's trained online, and

$$\sum_{t \geq 1} \eta_t = \infty, \quad \sum_{t \geq 1} \eta_t^2 < \infty$$

Then the following holds almost surely:

$$\text{If } s_t \xrightarrow{\text{d}} s_* \text{ then } q_t \to \underbrace{\text{Quantile}_{1-\alpha}(s_*(X, Y))}_{\text{(assuming this quantile is unique)}}$$

Theory for i.i.d. data

Interpretation—ensure robustness without hurting performance

model-based method:
parametric or
nonparametric

Theory for i.i.d. data

Interpretation—ensure robustness without hurting performance

$$\left.\begin{array}{c}\text{model-based method:}\\ \text{parametric or}\\ \text{nonparametric}\end{array}\right\} \longrightarrow \begin{array}{c}\text{conformal prediction:}\\ \text{distribution-free, but}\\ \text{assumes exchangeability}\end{array}$$

Theory for i.i.d. data

Interpretation—ensure robustness without hurting performance

$$\left.\begin{array}{c}\text{model-based method:}\\\text{parametric or}\\\text{nonparametric}\end{array}\right\} \longrightarrow \begin{array}{c}\text{conformal prediction:}\\\text{distribution-free, but}\\\text{assumes exchangeability}\end{array}\right\} \longrightarrow \begin{array}{c}\text{adaptive conformal:}\\\text{removes exchangeability}\\\text{assumption}\end{array}$$

Summary

Summary

Part 1 — the conformal prediction framework:

- Distribution-free predictive coverage under exchangeability
- Pairs with any existing model / algorithm

Summary

Part 1 — the conformal prediction framework:

- Distribution-free predictive coverage under exchangeability
- Pairs with any existing model / algorithm

Part 2 — extensions of the conformal framework to handle:

- Relaxations of conditional coverage guarantees
- Distribution shift
- The online setting (i.i.d. or with distribution drift)

Summary

Many additional extensions & topics in the literature, including:

- Other notions of conditional coverage (e.g., training-conditional)[26]
- Other notions of risk (beyond coverage/non-coverage)[27]
- Weighted conformal prediction to handle distribution shift[28] (applications to causal inference[29], survival analysis[30], ...)
- Relaxations or extensions of exchangeability (e.g., hierarchical sampling structures)[31]
- Distribution-free calibration[32]

[26]Vovk 2012, *Conditional validity of inductive conformal predictors*; Bian & B. 2021, *Training-conditional coverage for distribution-free predictive inference*; Liang & B. 2023, *Algorithmic stability implies training-conditional coverage for distribution-free prediction methods*

[27]Angelopolous et al 2022, *Conformal Risk Control*

[28]Tibshirani, B., Candès, Ramdas 2019, *Conformal Prediction Under Covariate Shift*

[29]Lei & Candès 2021, *Conformal inference of counterfactuals and individual treatment effects*

[30]Candès, Lei, Ren 2021, *Conformalized survival analysis*; Gui, Hore, Ren, & B. 2022, *Conformalized survival analysis with adaptive cutoffs*

[31]B., Candès, Ramdas, & Tibshirani 2023, *Conformal prediction beyond exchangeability*; Prinster et al 2024 *Conformal Validity Guarantees Exist for Any Data Distribution*; Lee, B., & Willett 2023 *Distribution-free inference with hierarchical data*

[32]Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Summary

Books & additional resources:

- *Algorithmic Learning in a Random World*, Vovk, Gammerman, Shafer 2005
- *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*, Angelopoulos & Bates 2021
- *Theoretical Foundations of Conformal Prediction*, Angelopoulos, B., Bates 2024+
- Lecture notes by Ryan Tibshirani: `https://www.stat.berkeley.edu/~ryantibs/statlearn-s23/lectures/conformal.pdf`
- Tutorial videos & slides on my website: `https://rinafb.github.io/talks/`

Thank you!