Conformal prediction for reliable AI

Nicola Paoletti, King’s College London

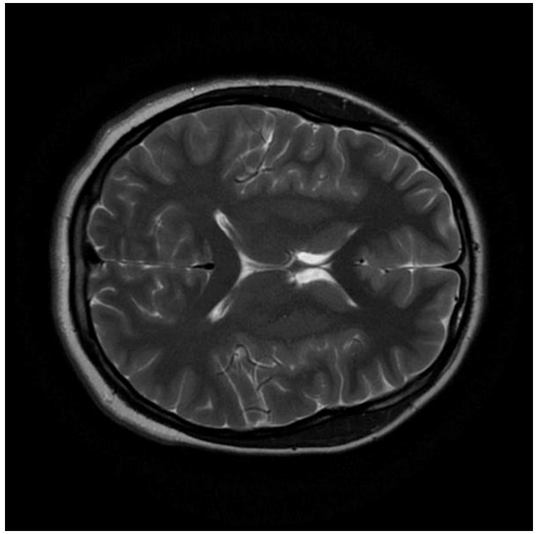
Department of Computer Science, University of Oxford

6 November 2025

Motivating example

A supervised learning task

$x: MRI\ images$



↧

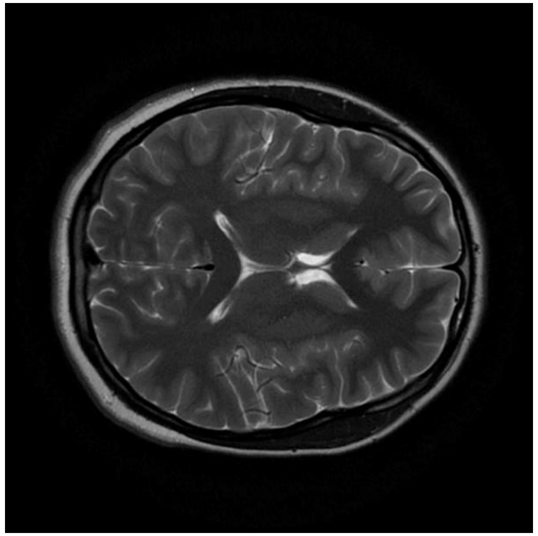
$y: \{normal, cancer\}$

Usual supervised learning approach:

- Obtain a training set of MRI images
- Use these to learn a machine learning classifier (e.g., neural net)
- Evaluate accuracy on unseen test data

Motivating example

$x: MRI\ images$



$\downarrow$

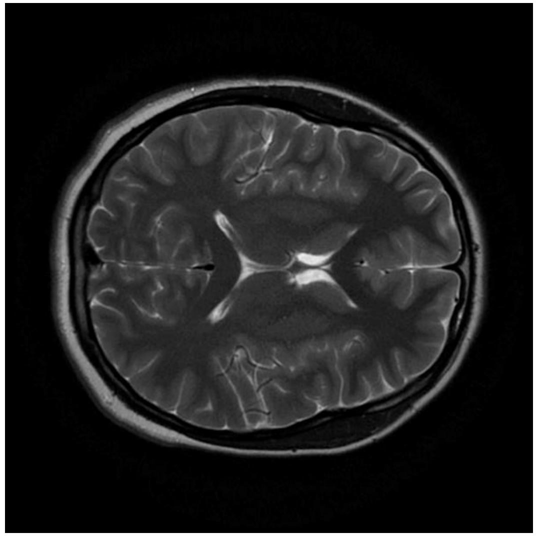
$y: \{normal, cancer\}$

All good?

- Point predictions not enough
- Decision makers (doctors) need to know likelihood of alternative outcomes, or **rule out unlikely outcomes**

Motivating example

x: *MRI images*



$\downarrow$

y: {*normal*, *cancer*}

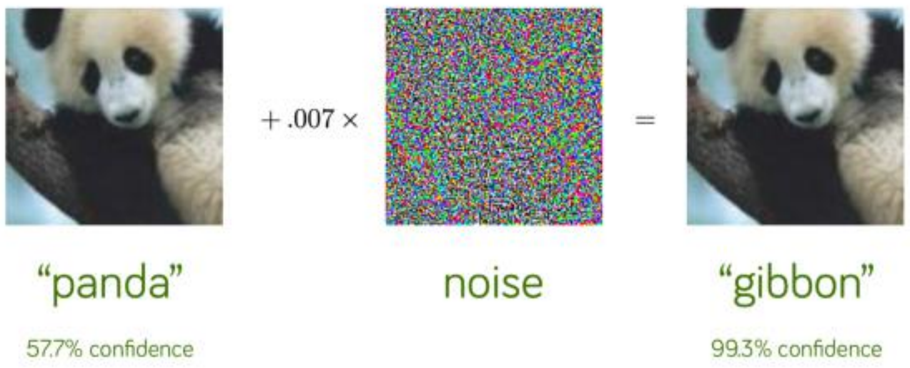
All good?

- Point predictions not enough
- Decision makers (doctors) need to know likelihood of alternative outcomes, or **rule out unlikely outcomes**

- Suppose we get a 90% test accuracy
- *Great*, but, **this tell us nothing on the prediction reliability for an unseen input** x^*

Failing loudly

- Neural nets output (softmax) likelihood for each class
- **Unreliable** as probability estimates:
  - Often overconfident on correct predictions
  - **Often overconfident on wrong ones too!**

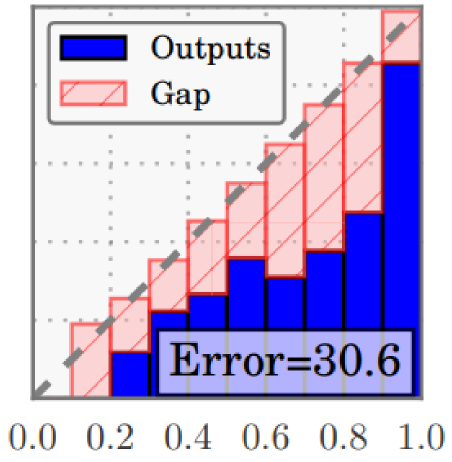


Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015.

Failing loudly

- Neural nets output (softmax) likelihood for each class
- **Unreliable** as probability estimates:
  - Often overconfident on correct predictions
  - **Often overconfident on wrong ones too!**

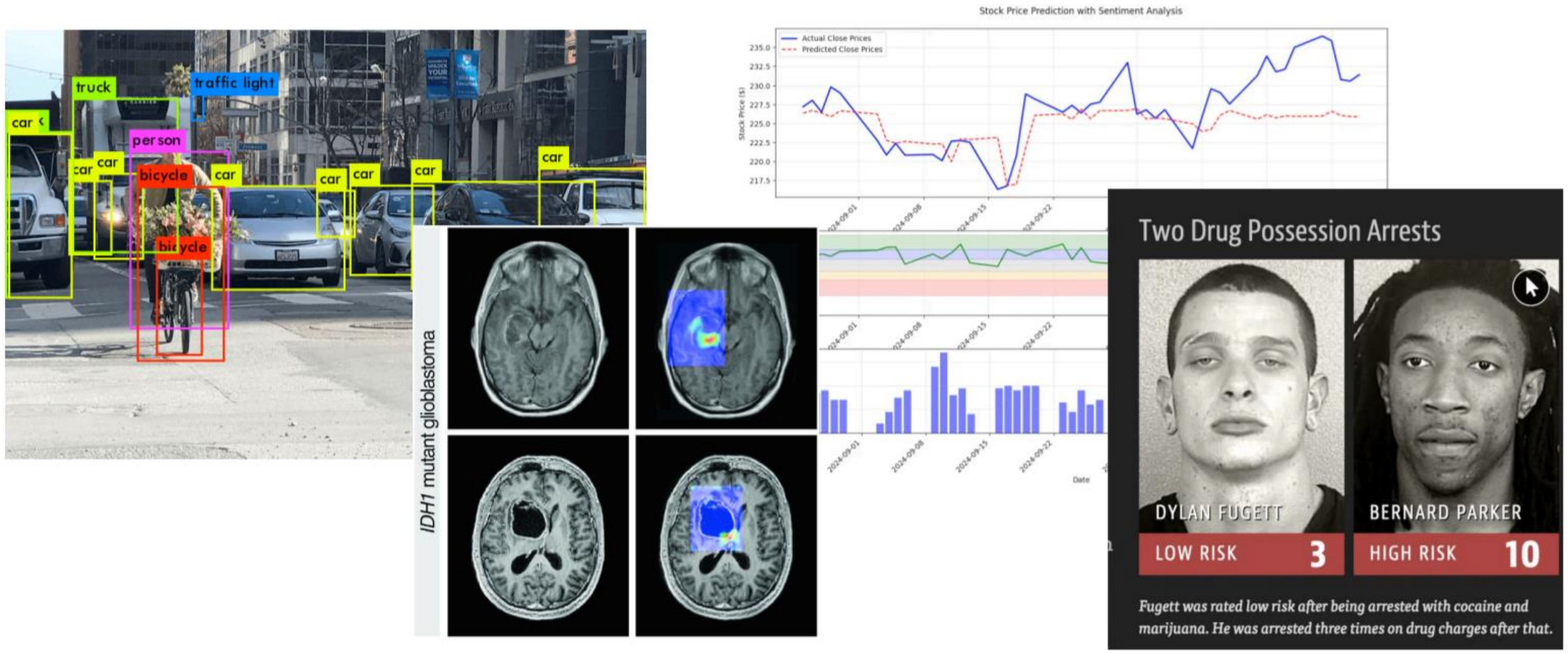
- I.e., softmax likelihoods are **poorly calibrated**
  (they don't reflect probability of correct classification)



Guo, Chuan, et al. "On calibration of modern neural networks." ICML 2017.

Uncertainty Quantification

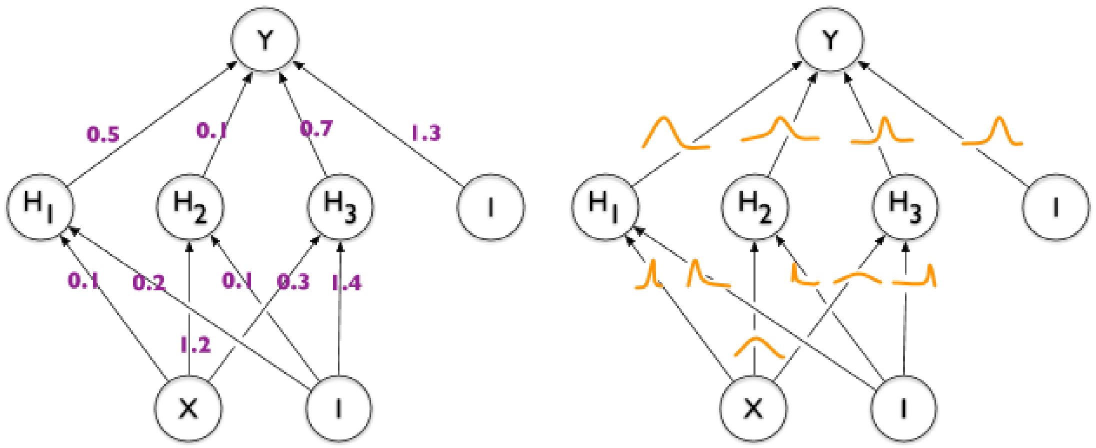
Crucial for high-stake decisions (e.g., autonomous driving, medical diagnosis, robotics, parole decisions, financial predictions)



Uncertainty Quantification

Some attempts:

- **Bayesian Neural Nets**, i.e., NNs with probabilistic weights
- Weight distributions learned with Bayesian inference

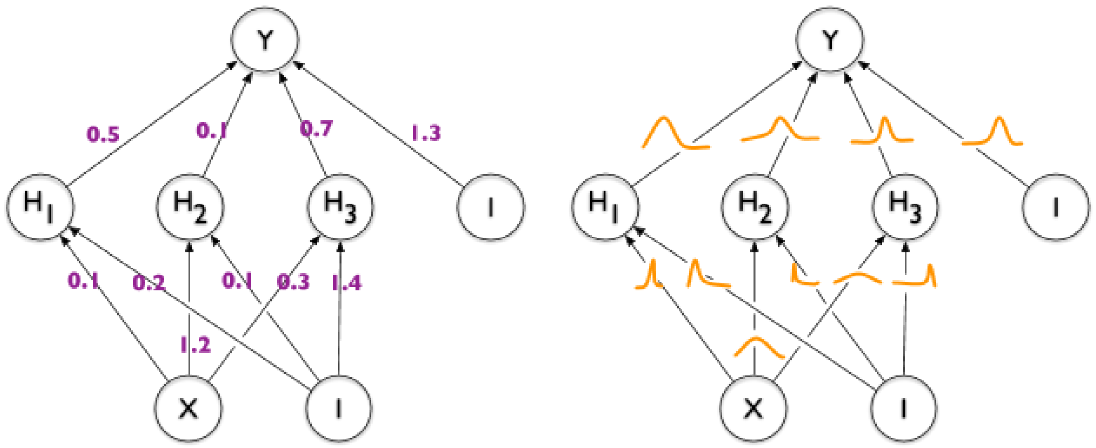


Gal, Yarin. "Uncertainty in deep learning." (2016)

Uncertainty Quantification

Some attempts:

- **Bayesian Neural Nets**, i.e., NNs with probabilistic weights
- Weight distributions learned with Bayesian inference



Gal, Yarin. "Uncertainty in deep learning." (2016)

- Correctness depends on choice of priors
- Only asymptotic guarantees (infinite data size)
- Precise inference (MCMC) feasible for small models only (VI approximations used in pratice)
- Computationally expensive, much hyperparameter tuning

Uncertainty Quantification

Some attempts:

- **Deep ensembles**
- Train multiple NNs using random subsets of data (or same data starting from different random weights)
- Use predictive distribution induced by these multiple NNs

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." NeurIPS 2017.

Uncertainty Quantification

Some attempts:

- **Deep ensembles**
- Train multiple NNs using random subsets of data (or same data starting from different random weights)
- Use predictive distribution induced by these multiple NNs

- No correctness guarantees
- Computationally expensive

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." NeurIPS 2017.

Conformal Prediction (CP)

- **Distribution-free**
  (no assumptions on priors or data-generating distribution)
- **Finite-sample guarantees** (as opposed to asymptotic)
- **Works with any ML model**
- Complements point predictions with **prediction regions guaranteed to include (unknown) ground truth with given probability**
  - Probabilities are well-calibrated (90% means 90%)

Vladimir Vovk (Royal Holloway)

Emmanuel Candes (Stanford)

- *Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer, 2005.*
- *Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint (2021).*

Outline

- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

Outline

- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

CP - a bird's eye view

Input:

- trained ML model $\hat{f}$
- held out **calibration data** $Z = \{(x_i, y_i)\}_{i=1}^{n} \sim \mathcal{Z}$
  - $\mathcal{Z}$ is the **unknown** data-generating distribution
- (non-conformity) **score function** $S(x, y)$
  - a quantitative notion of prediction error committed by $\hat{f}$
  - arbitrary, but should quantify "discrepancy" between y and $\hat{f}(x)$
- (arbitrary) error probability $\alpha \in (0,1)$

Output:

prediction region $C_\alpha(x^*)$ for test point (x^*, y^*) such that

$$\mathbb{P}\left(y^* \in C_\alpha(x^*)\right) \geq 1 - \alpha$$

CP - a bird's eye view

Input:
- trained ML model $\hat{f}$
- held out **calibration data** $Z = \{(x_i, y_i)\}_{i=1}^{n} \sim \mathcal{Z}$
  - $\mathcal{Z}$ is the **unknown** data-generating distribution
- (non-conformity) **score function** $S(x, y)$
  - a quantitative notion of prediction error committed by $\hat{f}$
  - arbitrary, but should quantify "discrepancy" between y and $\hat{f}(x)$

- Works with any distribution $\mathcal{Z}$ (assumed unknown)
- Works with any data size n
- Only assumption is $\boldsymbol{Z} \cup \{(\boldsymbol{x}^*, \boldsymbol{y}^*)\}$ **exchangeable** (weaker than iid)

$$\mathbb{P}\left(y^* \in C_\alpha(x^*)\right) \geq 1 - \alpha$$

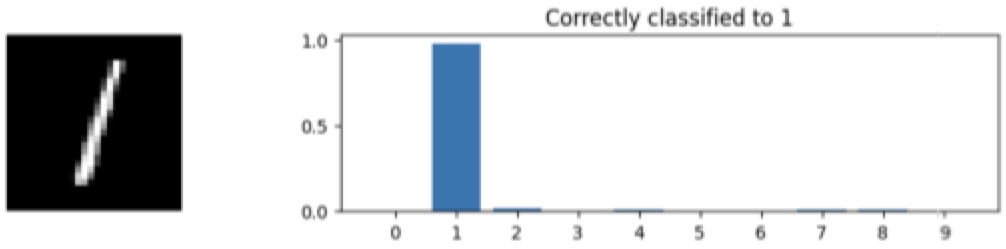
CP algorithm

- *Intuition:* include in $C_\alpha(x^*)$ all outputs (whose scores) appear likely w.r.t. calibration data
- **Step 0: define score function**
- CP guarantees hold for any choice of $S(x, y) \in \mathbb{R}$
  - But only reasonable $S(x, y)$ (see below) yield efficient (small/informative) regions

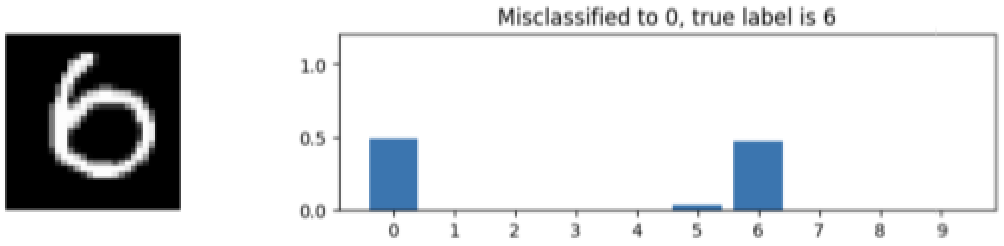
Common choices are:

- $S(x, y) = |\hat{f}(x) - y|_p$ for regression
- $S(x, y) = 1 - \hat{f}_y(x)$ for classification ($\hat{f}_y(x) \in [0,1]$ is likelihood predicted for class y)

CP algorithm



← Low (good) score



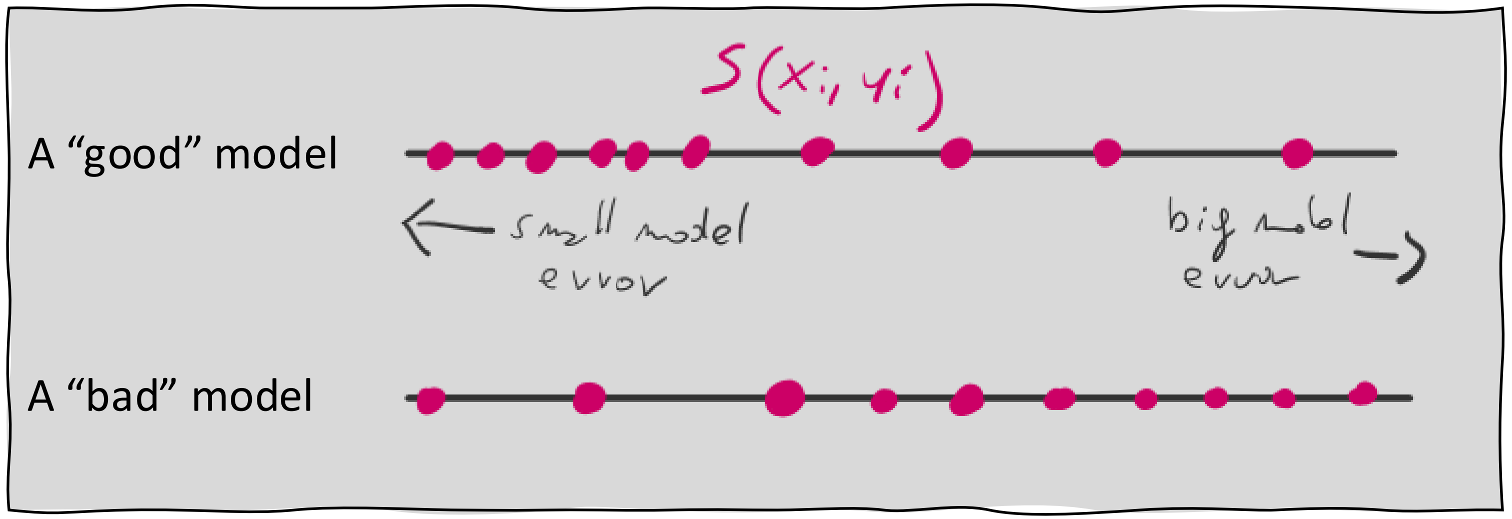
← High (bad) score

Common choices are:

- $S(x, y) = |\hat{f}(x) - y|_p$ for regression
- $S(x, y) = 1 - \hat{f}_y(x)$ for classification ($\hat{f}_y(x) \in [0,1]$ is likelihood predicted for class y)

CP algorithm

- **Step 1: construct calibration distribution**
  - empirical distribution of scores of correct outputs for all $(x_i, y_i) \in Z$



Formally, the calib distribution is

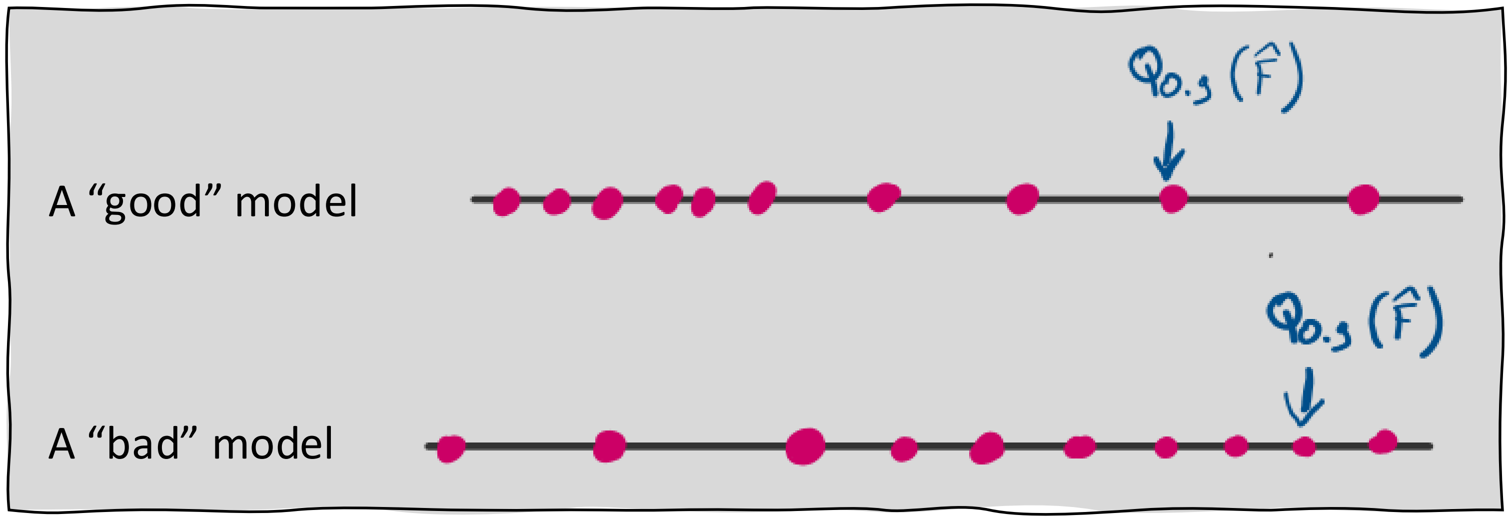
$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n} \delta_{s_i}$$

where

- $s_i = S(x_i, y_i)$
- δ_s is the Dirac distribution centred at s

CP algorithm

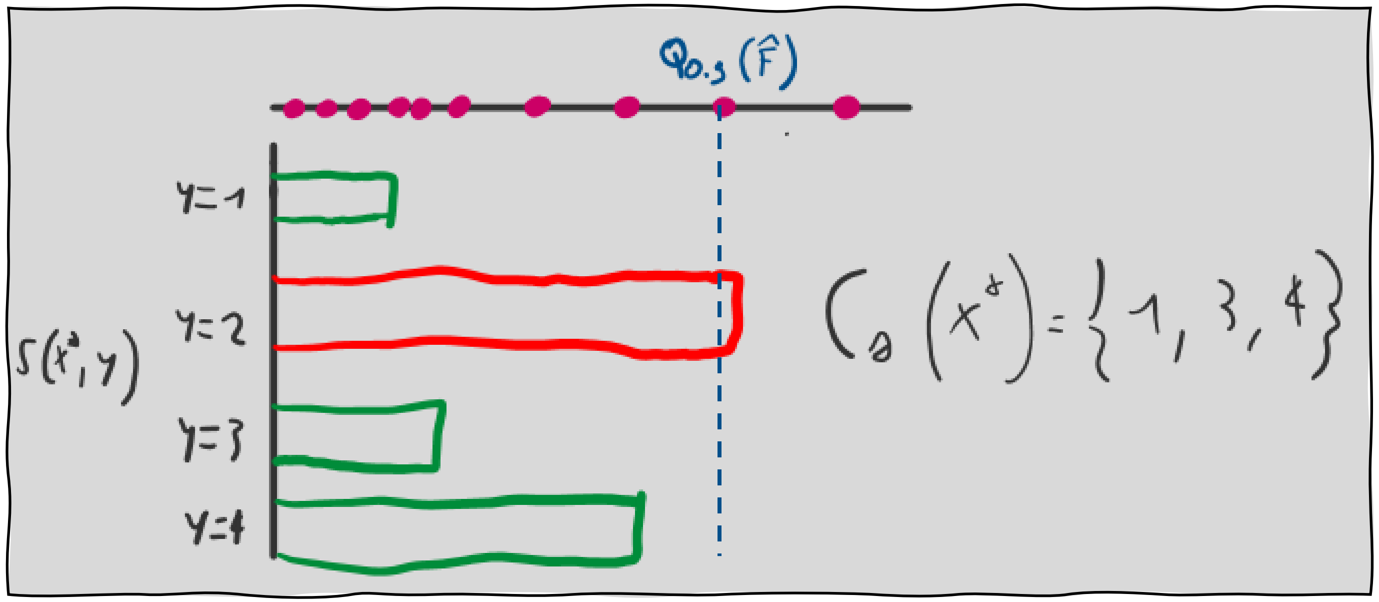
- **Step 2: find critical value**
  - I.e., find $Q_{1-\alpha}(\hat{F}) = (1-\alpha)$-quantile of calibration distribution
  - **Intuition** $(\alpha = 0.1)$: 90% of the examples have score $\leq Q_{0.9}(\hat{F})$,
    i.e., correct/true outputs have 90% probability of having score below $Q_{1-\alpha}(\hat{F})$



CP algorithm

- **Step 3: construct region**
  - Recall: correct outputs have probability $1 - \alpha$ of having score below $Q_{1-\alpha}(\hat{F})$
  - **Prediction region contains all outputs with score below $\boldsymbol{Q_{1-\alpha}(\hat{F})}$**

$$C_\alpha(x^*) = \{y \mid S(x^*, y) \leq Q_{1-\alpha}(\hat{F})\}$$

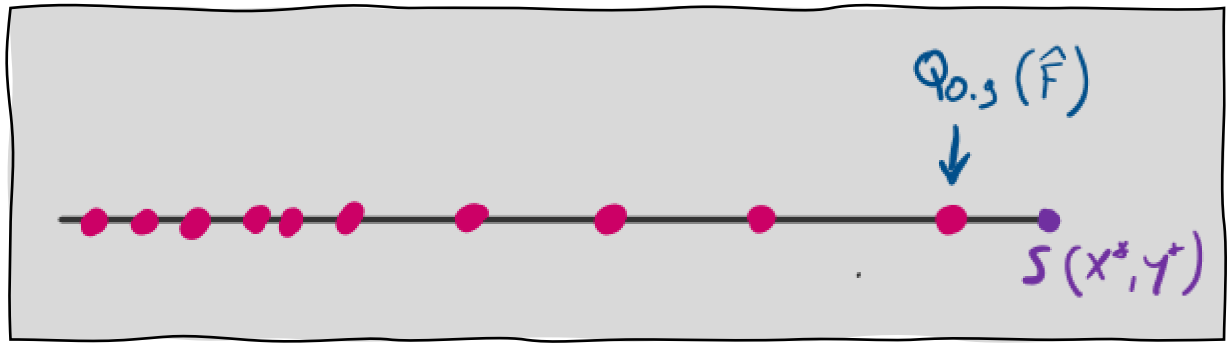


Such $C_\alpha(x^*)$ ensures that

$$\mathbb{P}\left(y^* \in C_\alpha(x^*)\right) \geq 1 - \alpha$$

CP algorithm

- **Step 1*: calibration distribution, caveat**
  - For a proper prediction interval, test point (x^*, y^*) should be considered in calibration distribution
  - But we don't know $S(x^*, y^*)$ (we don't know y^*)
  - We augment $\hat{F}$ to account for (x^*, y^*) (assigning worst-case score)



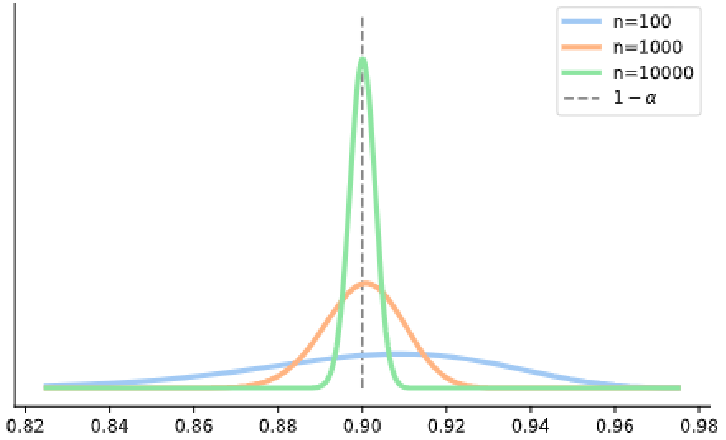
$$\hat{F} = \frac{1}{n+\mathbf{1}} \sum_{i=1}^{n} \delta_{s_i} + \frac{1}{n+\mathbf{1}} \boldsymbol{\delta_\infty}$$

CP – important remarks

- Bad models or small calibration sets lead to large $Q_{1-\alpha}$
  - Meaning, large uncertainty/prediction regions (as desired)
  - (assuming sensible score function)

CP – important remarks

- Bad models or small calibration sets lead to large $Q_{1-\alpha}$
  - Meaning, large uncertainty/prediction regions (as desired)
- CP guarantees are **marginal**
  - i.e., $C_\alpha(x^*)$ includes y^* on average 90% of the times
    - w.r.t. distribution of $\left((x_1, y_1), \dots, (x_n, y_n), (x^*, y^*)\right)$
  - Coverage of test point for a fixed calibration set is a random variable (see right)
    - With n big enough, variability is negligible



Coverage distribution for $\alpha = 0.1$

Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint (2021).

CP – important remarks

- Bad models or small calibration sets lead to large $Q_{1-\alpha}$
  - Meaning, large uncertainty/prediction regions (as desired)
- CP guarantees are **marginal**
  - i.e., $C_\alpha(x^*)$ includes y^* on average 90% of the times
- For regression ($y \in \mathbb{R}$), evaluating all outputs is impossible
  - We construct region “implicitly”
  - E.g., for $S(x, y) = |\hat{f}(x) - y|$, $\boldsymbol{C_\alpha(x^*) = [\hat{f}(x) \pm Q_{1-\alpha}(\hat{F})]}$

CP – Classification example

Figure 1: Prediction set examples on Imagenet. *We show three progressively more difficult examples of the class **fox squirrel** and the prediction sets (i.e., $\mathcal{C}(X_{\text{test}})$) generated by conformal prediction.*

Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint (2021).

Outline

- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

From marginal to conditional

- **Marginal guarantees** (standard CP):

$$\mathbb{P}_{Z,x^*,y^*}\left(y^* \in C_\alpha(x^*)\right) \geq 1 - \alpha$$

  - coverage *on average* over test points
- **(Test-)conditional guarantees**

$$\mathbb{P}_{Z,x^*,y^*}\left(y^* \in C_\alpha(x^*) \mid x^*\right) \geq 1 - \alpha, \forall x^*$$

  - coverage *for every* test point

From marginal to conditional

- **Marginal guarantees** (standard CP):

$$\mathbb{P}_{Z,x^*,y^*}\left(y^* \in C_\alpha(x^*)\right) \geq 1-\alpha$$

Impossibility of conditional CP

If x continuous, it's impossible to satisfy at the same time:

- Conditional coverage
- Distribution-free
- Validity in finite samples

(except for trivial prediction sets)

Vovk, Vladimir. "Conditional validity of inductive conformal predictors." In Asian conference on machine learning, pp. 475-490. PMLR, 2012.

Foygel Barber, Rina, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. "The limits of distribution-free conditional predictive inference." Information and Inference: A Journal of the IMA 10, no. 2 (2021): 455-482.

Group conditional coverage (aka Mondrian CP)

(x, y)-space admits partition into groups

$$\boldsymbol{G} = \{G_1, \dots, G_k\}$$

- E.g., patients grouped by age/gender/condition

Group-conditional guarantees

$$\mathbb{P}_{Z,x^*,y^*}\left(y^* \in C_\alpha(x^*) \mid (x^*, y^*) \in G\right) \geq 1 - \alpha, \forall G \in \boldsymbol{G}$$

- E.g., ensuring guarantees for every patient group

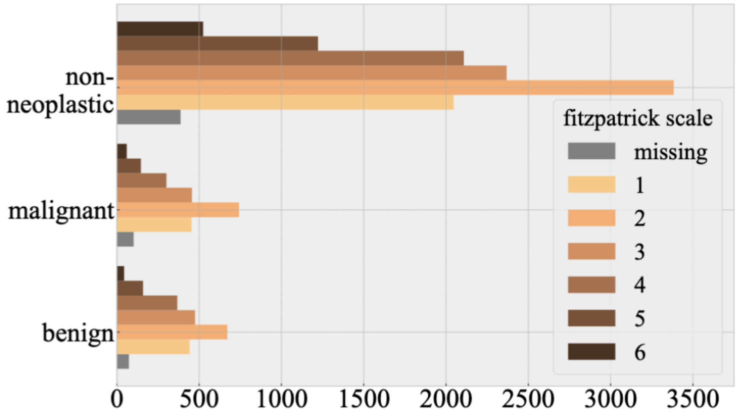


Figure 3: Distribution of skin conditions by Fitzpatrick skin type and categorization of the 114 different lesions into one of three broad categories: non-neoplastic, malignant, or benign.

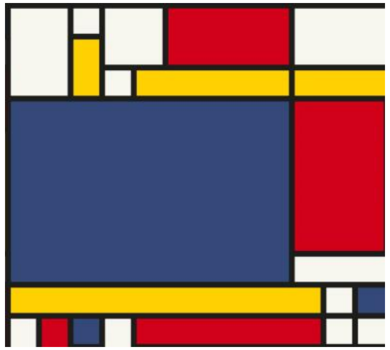
- *Lu, Charles, et al. "Fair conformal predictors for applications in medical imaging." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 11. 2022.*
- *Toccaceli, Paolo, and Alexander Gammerman. "Combination of inductive mondrian conformal predictors." Machine Learning 108.3 (2019): 489-510.*

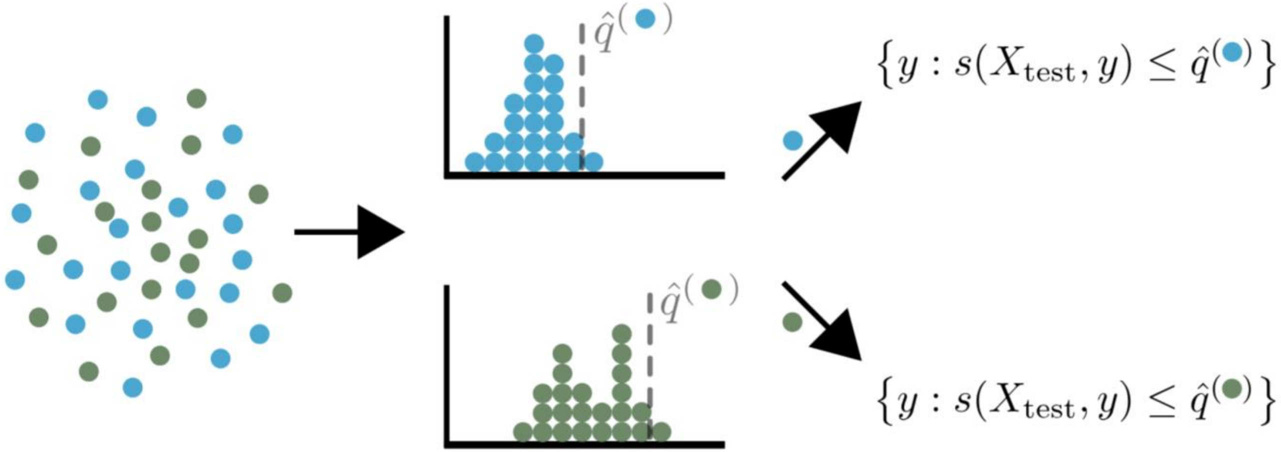
Group conditional coverage (aka Mondrian CP)

$$\mathbb{P}_{Z,x^*,y^*}\left(y^* \in C_\alpha(x^*) \mid (x^*, y^*) \in G\right) \geq 1 - \alpha, \forall G \in \boldsymbol{G}$$

Approach:

- Partition calibration set w.r.t. G
- Compute group conditional quantiles $q^{G_1}, q^{G_2}, \ldots$
- Apply right quantile based on test point membership





Outline

- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

CP and distribution shifts

- CP only relies on exchangeability
- Violated when test distribution $P^*_{X,Y}$ changes w.r.t. calibration distribution $P_{X,Y}$
  - **more frequent than not**
- $P_{X,Y} = P_X \times P_{Y|X} \neq P^*_{X,Y} = P^*_X \times P^*_{Y|X}$

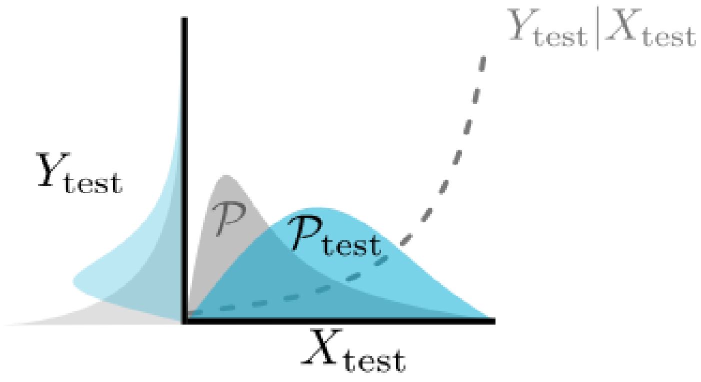
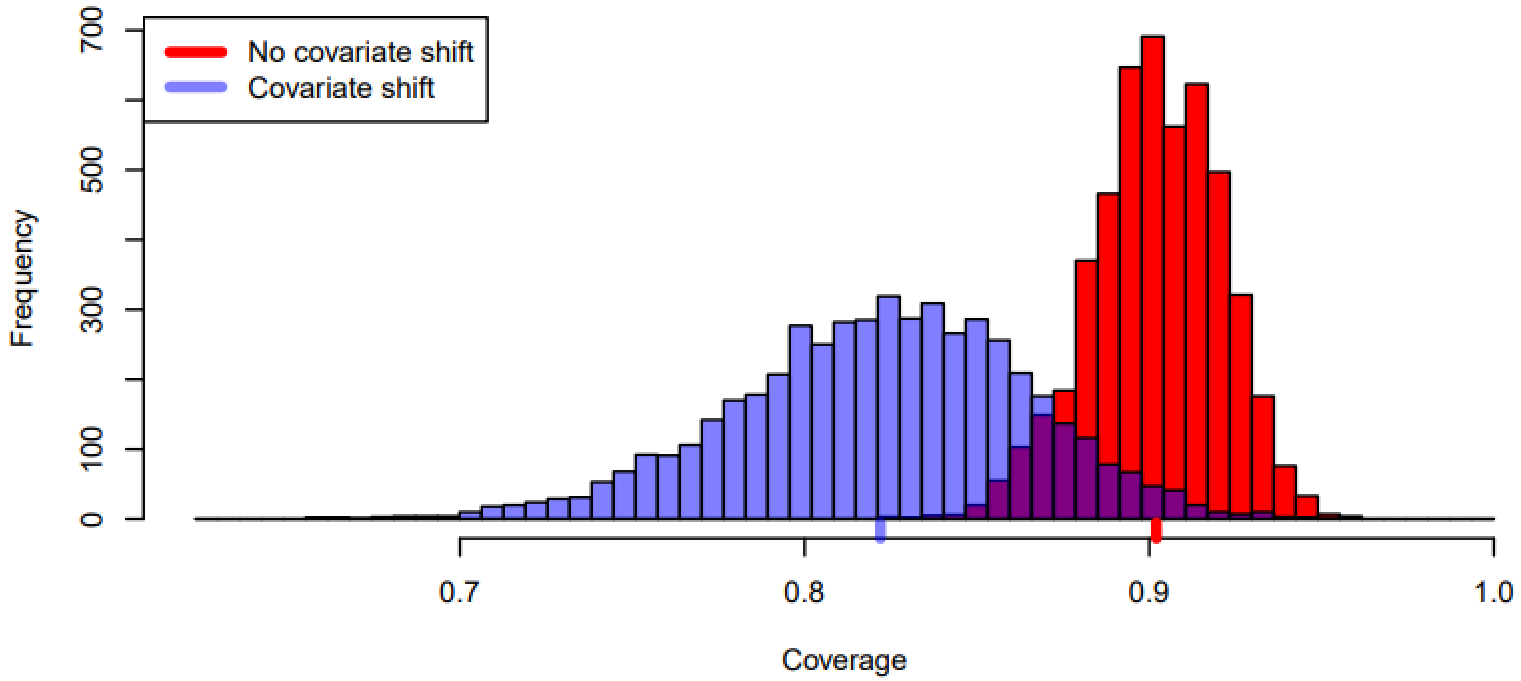


Illustration of covariate shift
(from “Gentle introduction…”)

- **Covariate shift:** P_X changes, $P_{Y|X}$ stays the same
- **Concept drift:** $P_{Y|X}$ changes, P_X remains the same

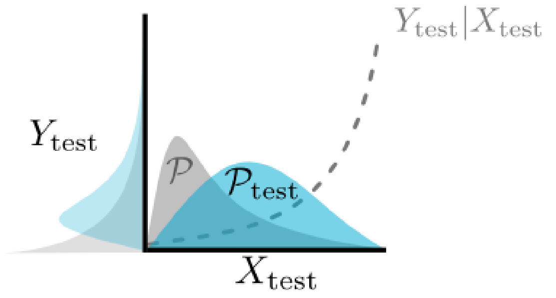
CP and distribution shifts

CP coverage can be jeopardized by shifts



R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in NeurIPS 2019, pp. 2530–2540.

Solution: weighted exchangeability

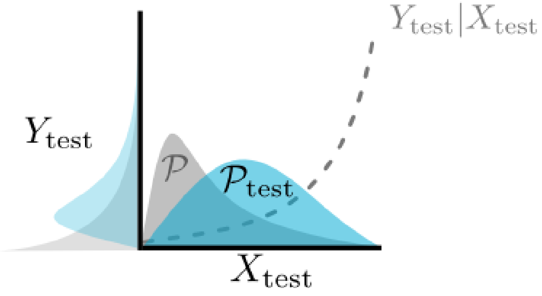


- Manipulate probabilities of calibration data as if it came from $P^*_{X,Y}$
  - In this way, we **restore exchangeability**
- **How:** use density ratio $w(x, y) = \mathrm{d}P^*_{X,Y}(x, y)/\mathrm{d}P_{X,Y}(x, y)$

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in NeurIPS 2019, pp. 2530–2540.

Solution: weighted exchangeability

- Manipulate probabilities of calibration data as if it came from $P^*_{X,Y}$
  - In this way, we **restore exchangeability**
- **How:** use density ratio $w(x,y) = \mathrm{d}P^*_{X,Y}(x,y)/\mathrm{d}P_{X,Y}(x,y)$



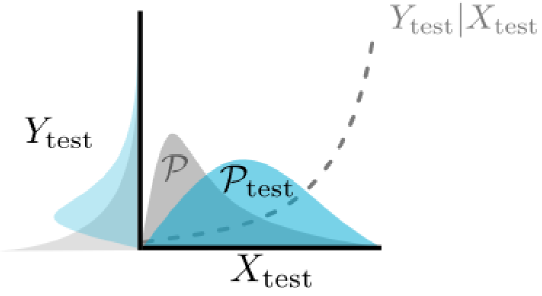
$$\hat{F} = \frac{1}{n+1}\sum_{i=1}^{n} \delta_{S(x_i,y_i)} + \frac{1}{n+1}\delta_\infty \longrightarrow \hat{F}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{p}_{\boldsymbol{x}_i,\boldsymbol{y}_i} \cdot \delta_{S(x_i,y_i)} + \boldsymbol{p}_{\boldsymbol{x},\boldsymbol{y}} \cdot \delta_\infty$$

$$p_{x_i,y_i} = \frac{w(x_i,y_i)}{\sum_{j=1}^{n} w(x_j,y_j) + w(x,y)} \; ; \; p_{x,y} = \frac{w(x,y)}{\sum_{j=1}^{n} w(x_j,y_j) + w(x,y)}$$

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in NeurIPS 2019, pp. 2530–2540.

Solution: weighted exchangeability

- Manipulate probabilities of calibration data as if it came from $P^*_{X,Y}$
  - In this way, we **restore exchangeability**



- **How:** use density ratio $w(x,y) = \mathrm{d}P^*_{X,Y}(x,y)/\mathrm{d}P_{X,Y}(x,y)$

$$\hat{F} = \frac{1}{n+1}\sum_{i=1}^{n} \delta_{S(x_i,y_i)} + \frac{1}{n+1}\delta_\infty \longrightarrow \hat{F}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{p}_{\boldsymbol{x}_i,\boldsymbol{y}_i} \cdot \delta_{S(x_i,y_i)} + \boldsymbol{p}_{\boldsymbol{x},\boldsymbol{y}} \cdot \delta_\infty$$

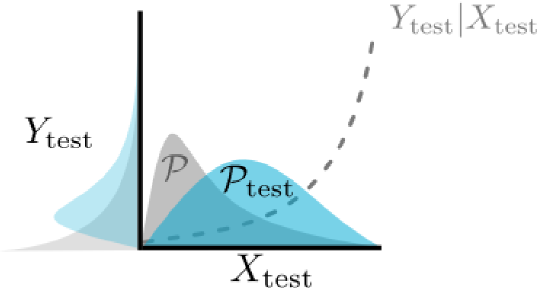
$$\boldsymbol{C}_\alpha(\boldsymbol{x}^*) = \{\boldsymbol{y} \mid \boldsymbol{S}(\boldsymbol{x}^*,\boldsymbol{y}) \leq \boldsymbol{Q}_{1-\alpha}(\widehat{\boldsymbol{F}}(\boldsymbol{x}^*,\boldsymbol{y}))\}$$

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in NeurIPS 2019, pp. 2530–2540.

Solution: weighted exchangeability

Challenge:

- requires reweighting $\hat{F}$ for every test input x^* and candidate output y
- need to enumerate and test candidate outputs individually → tricky for regression ($y \in \mathbb{R}$)



$$\hat{F}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{p}_{\boldsymbol{x}_i, \boldsymbol{y}_i} \cdot \delta_{S(x_i, y_i)} + \boldsymbol{p}_{\boldsymbol{x}, \boldsymbol{y}} \cdot \delta_{\infty}$$

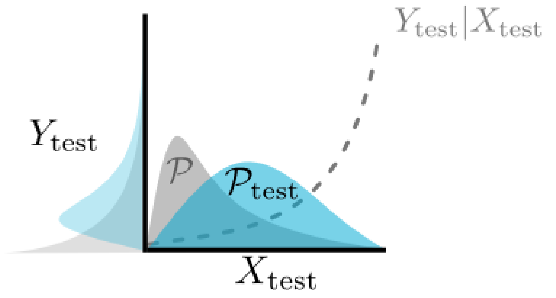
$$\boldsymbol{C}_{\alpha}(\boldsymbol{x}^*) = \{\boldsymbol{y} \mid \boldsymbol{S}(\boldsymbol{x}^*, \boldsymbol{y}) \leq \boldsymbol{Q}_{1-\alpha}(\hat{\boldsymbol{F}}(\boldsymbol{x}^*, \boldsymbol{y}))\}$$

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, “Conformal prediction under covariate shift,” in NeurIPS 2019, pp. 2530–2540.

CP and covariate shift

Easier:

- requires reweighting $\hat{F}$ for every test input x^* **only** ~~and candidate output y~~
- **no** need to enumerate candidate outputs in regression (can use "implicit" construction of C_α)



$$w(x,y) = \frac{\mathrm{d}P^*_{X,Y}(x,y)}{\mathrm{d}P_{X,Y}(x,y)} = \frac{\mathrm{d}(P^*_X(x) \times P^*_{Y|X}(x,y))}{\mathrm{d}(P_X(x) \times P_{Y|X}(x,y))} = \frac{\mathrm{d}P^*_X(x)}{\mathrm{d}P_X(x)} = w(x)$$

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{p}_{\boldsymbol{x}_i} \cdot \delta_{S(x_i, y_i)} + \boldsymbol{p}_{\boldsymbol{x}} \cdot \delta_\infty$$

$$\boldsymbol{C}_\alpha(\boldsymbol{x}^*) = \{\boldsymbol{y} \mid \boldsymbol{S}(\boldsymbol{x}^*, \boldsymbol{y}) \leq \boldsymbol{Q}_{1-\alpha}(\hat{\boldsymbol{F}}(\boldsymbol{x}^*))\}$$

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas, "Conformal prediction under covariate shift," in NeurIPS 2019, pp. 2530–2540.

Localised CP for quasi-conditional validity

Idea: relax conditional guarantees

$$\mathbb{P}\left(y^* \in C_\alpha(x^*) \mid x^*\right) \geq 1 - \alpha, \forall x^*$$

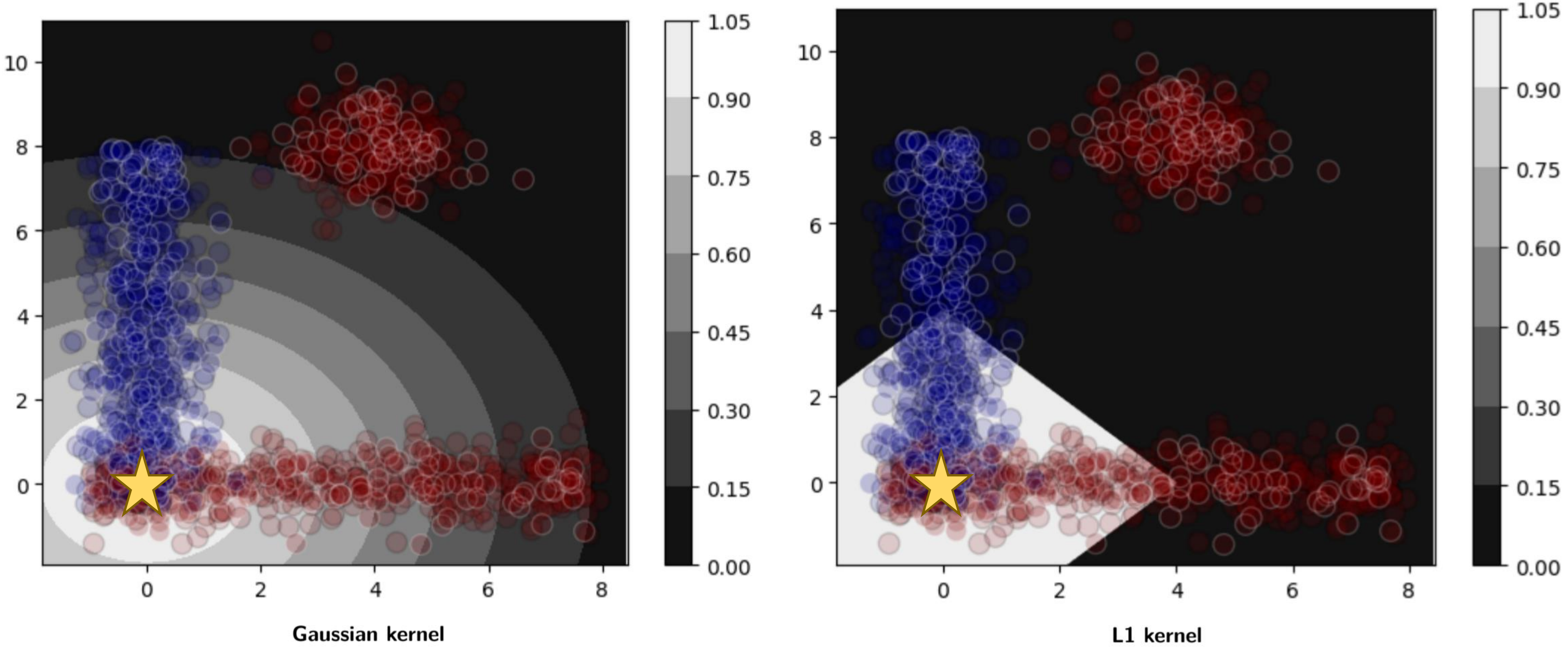
to hold for a local neighbourhood of the test point

How: Reweight probabilities of calibration points to favour points closer to x^*

- Akin to covariate shift where $P^*_{X,Y}$ is localised around x^*

Guan, Leying. "Localized conformal prediction: A generalized inference framework for conformal prediction." Biometrika 110.1 (2023): 33-50.

Localised CP - example



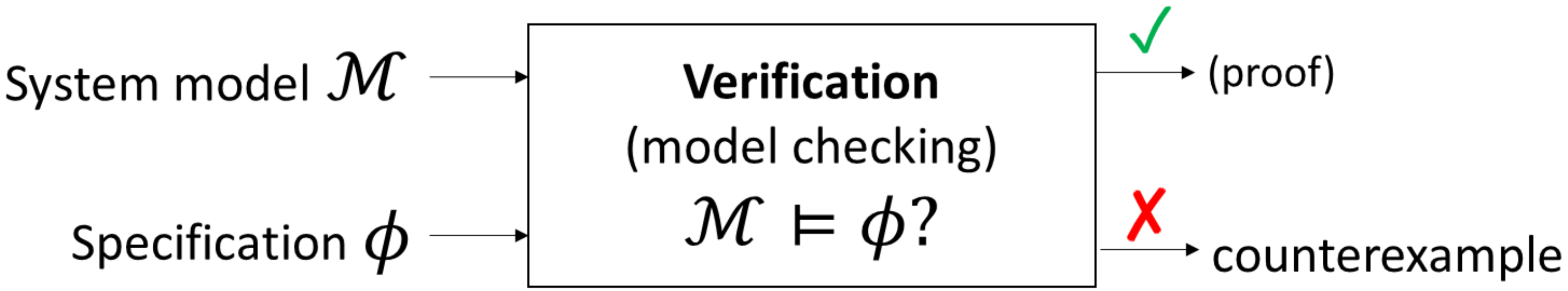
Outline

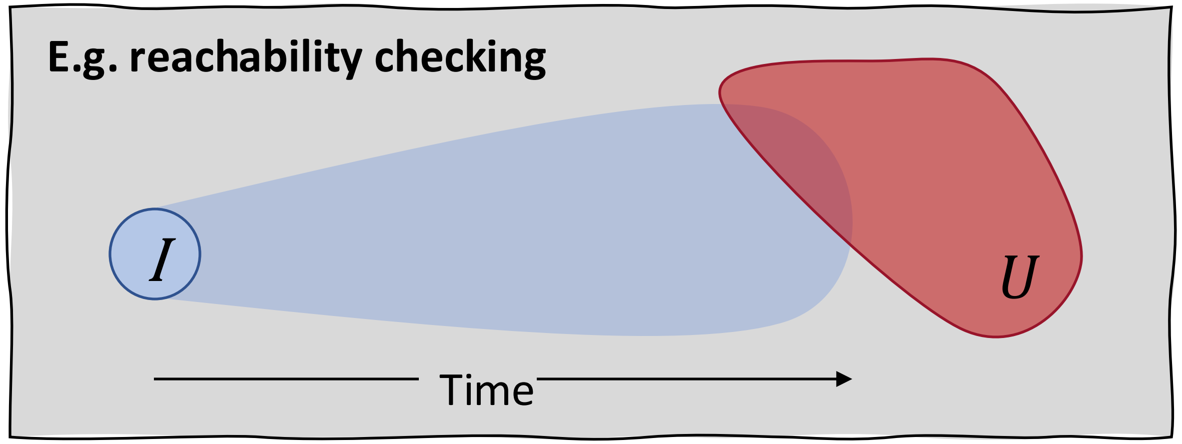
- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

Motivation: cyber-physical systems verification

CPSs are ubiquitous and found in many safety-critical applications

Verification to ensure that they work as intended





Verification vs. Predictive Monitoring

- We have exact tools for verification/model checking of CPSs:
  - Precise
  - But computationally expensive
- **Aim, predictive monitoring:** predict at runtime future CPS violations

CP for Predictive Monitoring

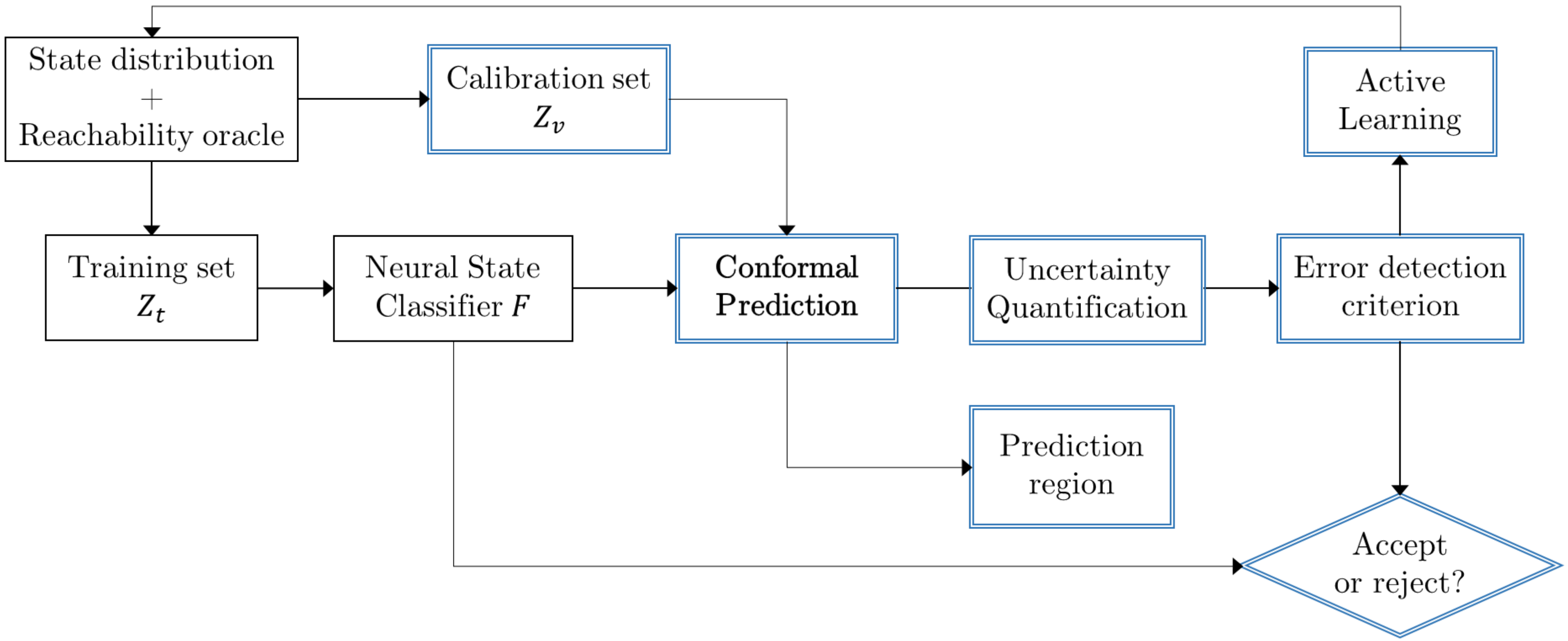
- We have exact tools for verification/model checking of CPSs:
  - Precise
  - But computationally expensive
- **Aim, predictive monitoring:** predict at runtime future CPS violations

Solution idea:

- *Offline:* Learn a data-driven surrogate model of (expensive) model checker
  - It must be *accurate* and *fast*, e.g., a neural net
- *Online:* Apply conformal prediction on the surrogate
  - Trading “hard” model-checking guarantees for probabilistic ones

Predictive monitoring for CPS reachability

[ATVA18, RV19, SSST21, RV21, ISOLA22], with Trieste and Stony Brook



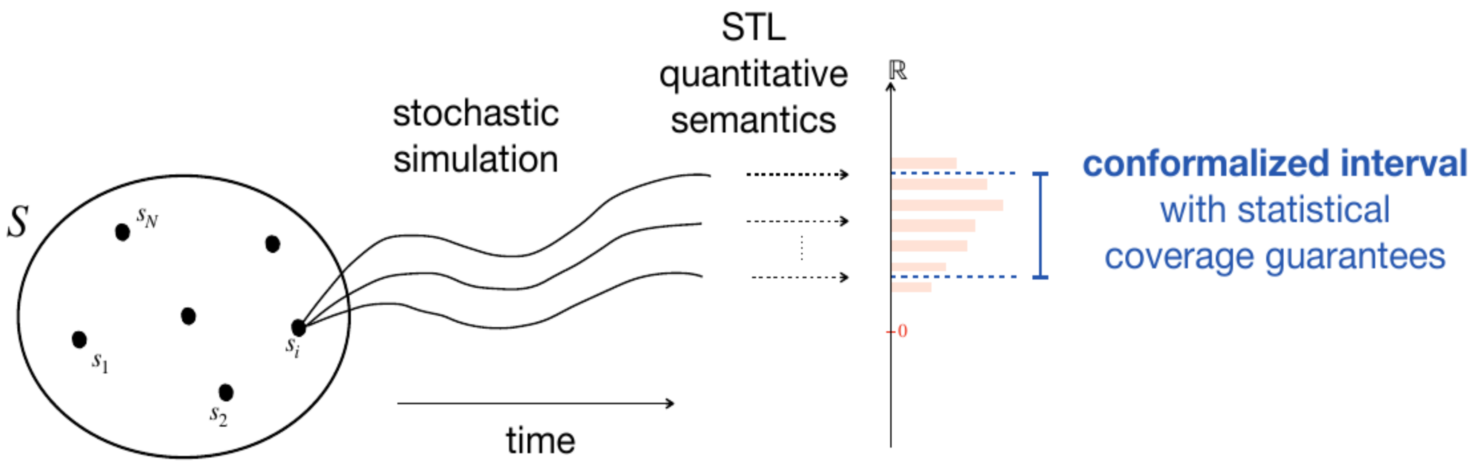
- **Prediction regions with probabilistic guarantees**
- **Measures of prediction uncertainty**, used to **reject unreliable predictions**

Predictive monitoring for STL

[HSCC23, RV23, NAHS25, RV25], with Trieste and USC

From binary reachability **to Signal Temporal Logic (STL)**

- More expressive specs + quantitative notion of satisfaction (STL robustness)
- Stochastic dynamics
- Based on conformalised quantile regression method



Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in NeurIPS 2019

Predictive monitoring for STL

[HSCC23, RV23, NAHS25, RV25], with Trieste and USC

- Extended for (pseudo-)conditional guarantees and multi-modal scenarios
- Uses generative model + mode predictor + mode-conditional quantiles

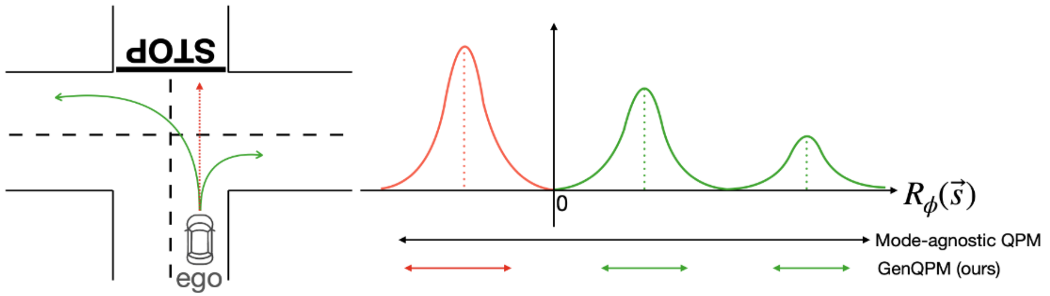
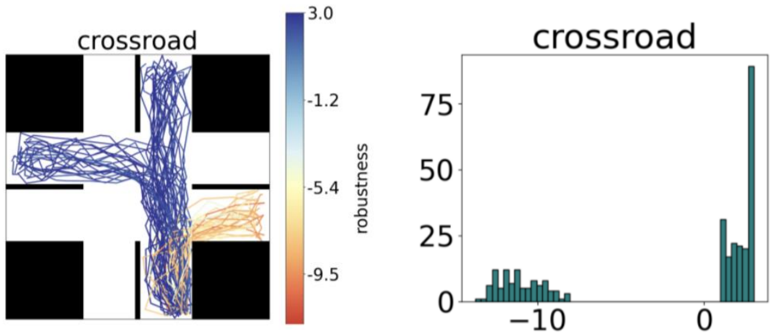
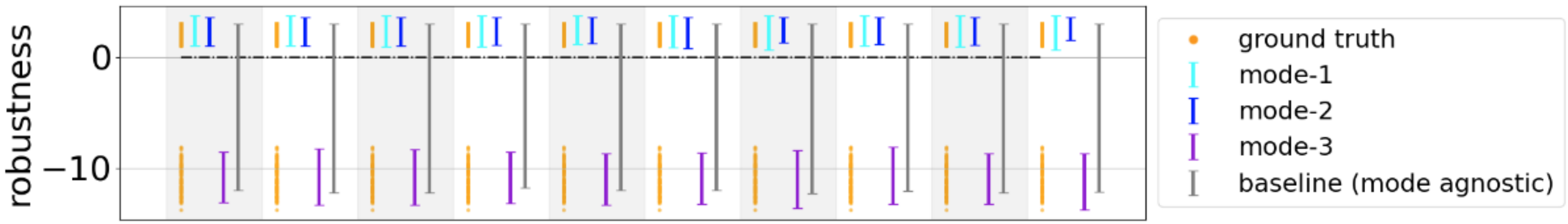


Illustration of Cross-road case study



True trajectories and their modes (left); corresponding STL robustness (right)



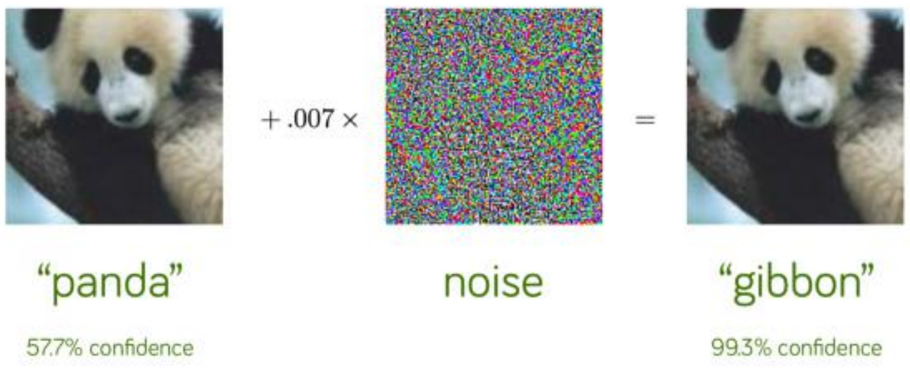
Our prediction regions vs mode-agnostic baseline

Outline

- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
    - CP for predictive monitoring of cyber-physical systems
    - CP and adversarial attacks (and for robust LLM monitoring)
    - CP for off-policy prediction
    - CP for counterfactual explanations

CP and adversarial attacks

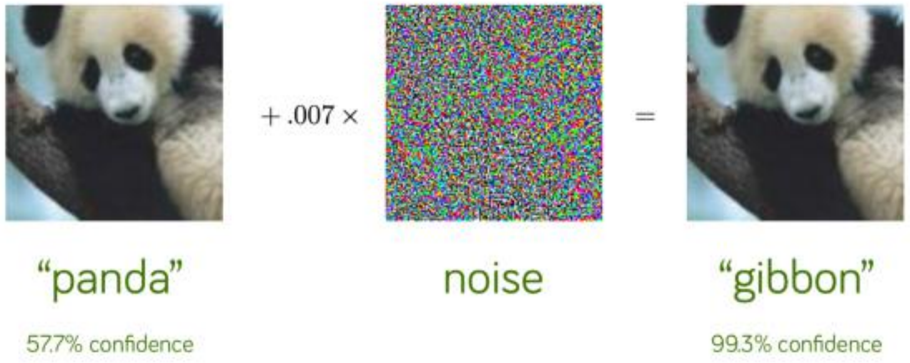
- Neural networks are susceptible to **adversarial attacks**
  - small perturbations changing the network's decision
- CP's exchangeability assumption violated under attacks, leading to loss of coverage/guarantees



Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015.

CP and adversarial attacks

- Neural networks are susceptible to **adversarial attacks**
- CP's exchangeability assumption violated under attacks, leading to loss of coverage/guarantees



Explaining and Harnessing Adversarial Examples, Goodfellow et al, ICLR 2015.

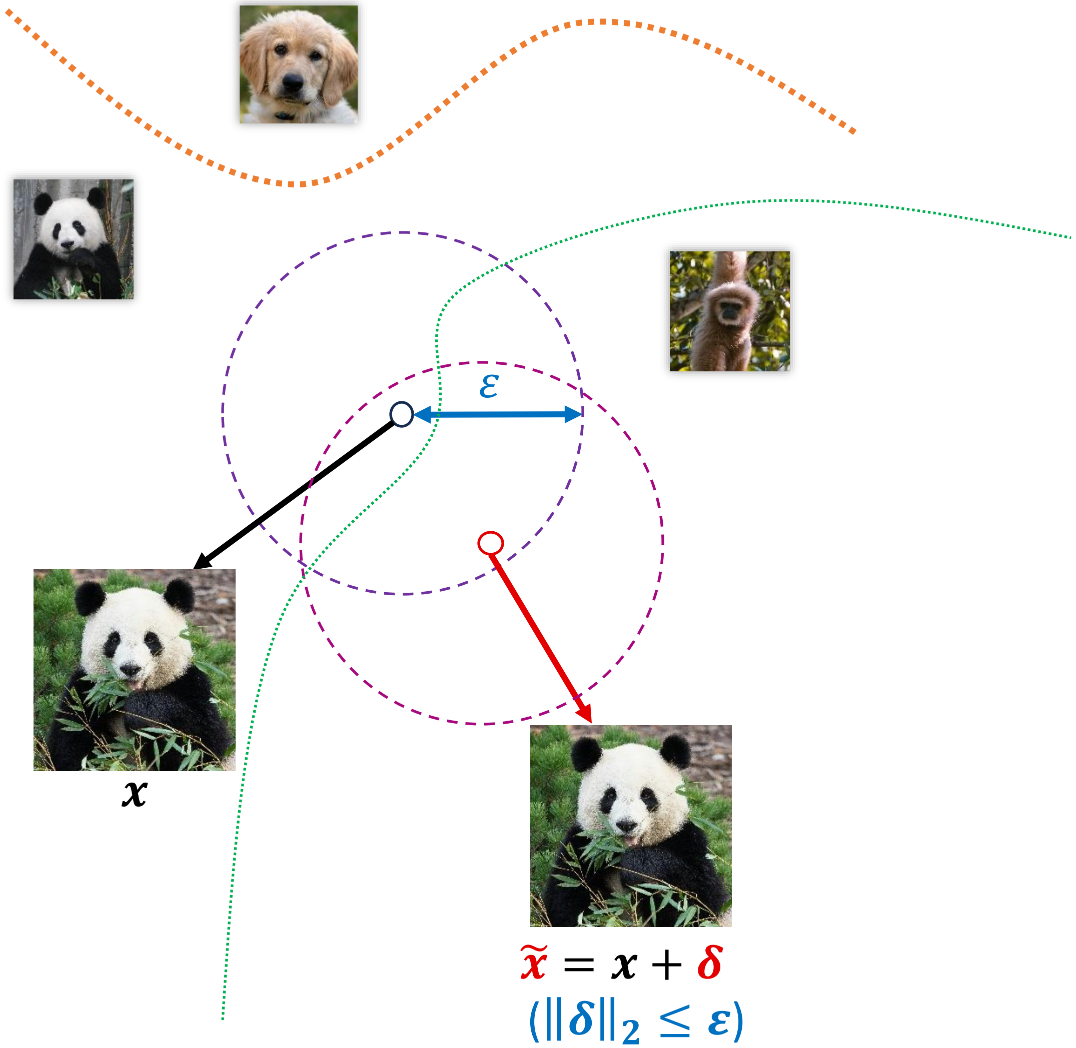
Adversarially robust CP problem

Given a perturbation/attack budget ϵ (w.r.t. some p norm) and level α, construct a robust prediction region $C_{\alpha,\epsilon}$ s.t.

$$\mathbb{P}\left(y^* \in C_{\alpha,\epsilon}(x^* + \boldsymbol{\delta})\right) \geq 1 - \alpha, \textbf{ for any } |\delta|_p \leq \epsilon$$

Verifiably robust CP (VRCP)

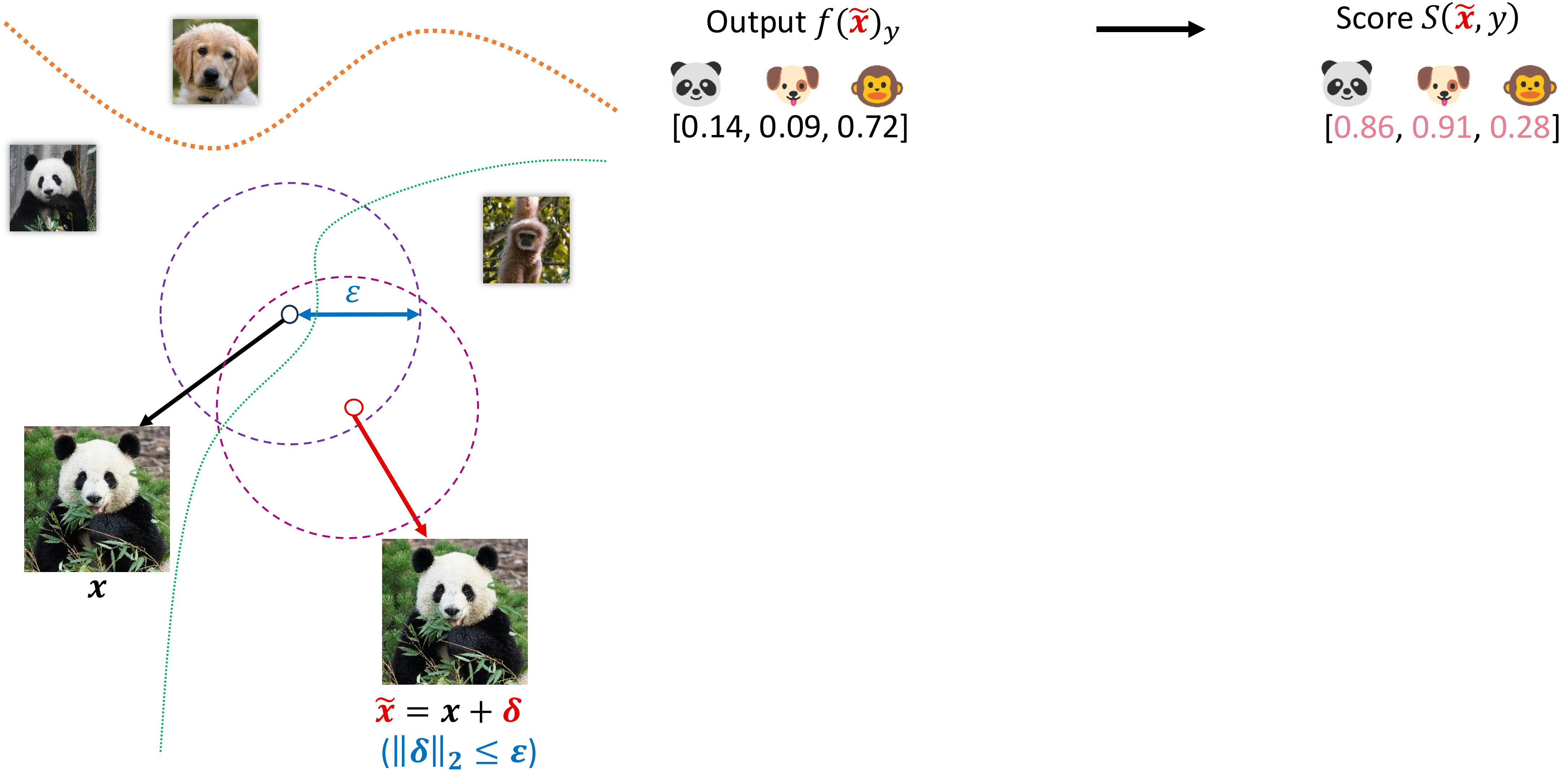
[NeurIPS24, PR26]



- **Key idea:** use neural network verification tools to bound the model outputs and the CP scores
- This leads to robust (more conservative) prediction regions

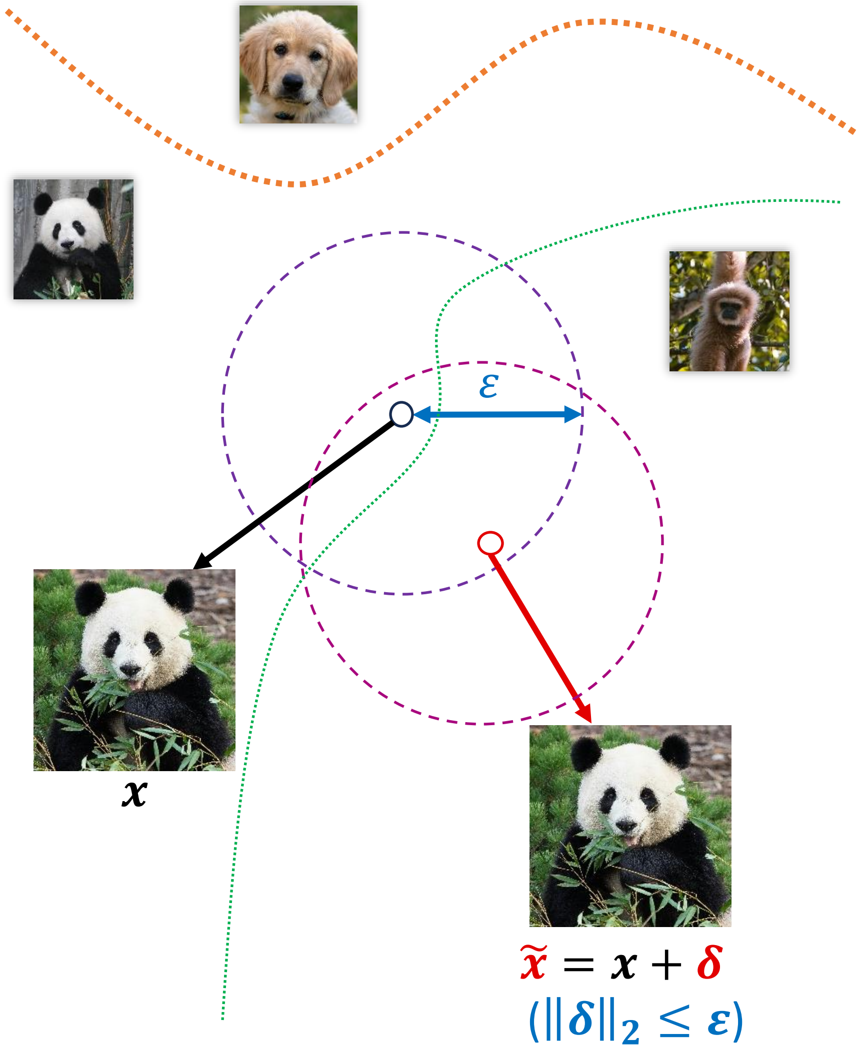
Verifiably robust CP (VRCP)

[NeurIPS24, PR26]



Verifiably robust CP (VRCP)

[NeurIPS24, PR26]



Output $f(\tilde{\boldsymbol{x}})_y$ → Score $S(\tilde{\boldsymbol{x}}, y)$

🐼 🐶 🐵

[0.14, 0.09, 0.72] → [0.86, 0.91, 0.28]

Output bounds of $f(\tilde{\boldsymbol{x}})_y$ → Bounds of $S(\tilde{\boldsymbol{x}}, y)$

Via an NN-verifier w.r.t l_2-norm

🐼 :[0.06, 0.77]

🐶 :[0.01, 0.21]

🐵 :[0.21, 0.88]

Bounds of $S(\tilde{\boldsymbol{x}}, y)$:

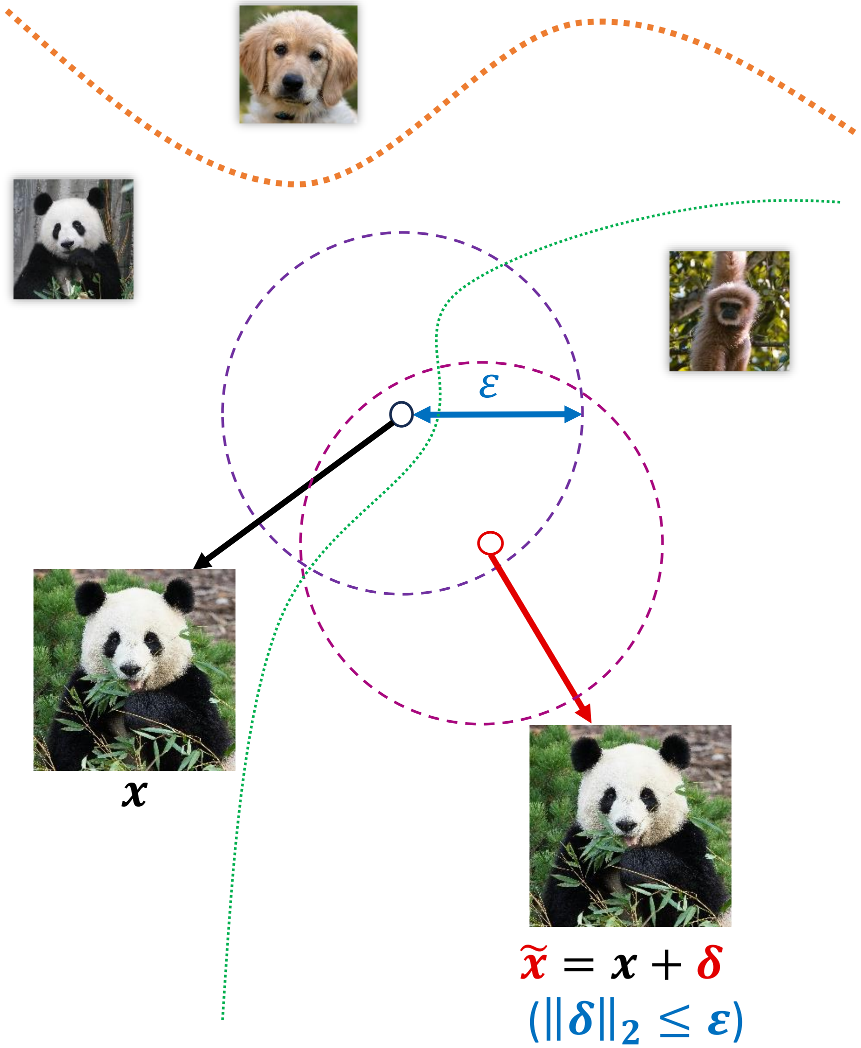
🐼 :[0.23, 0.94]

🐶 :[0.79, 0.99]

🐵 :[0.12, 0.79]

Verifiably robust CP (VRCP)

[NeurIPS24, PR26]



Output $f(\tilde{\boldsymbol{x}})_y$ ⟶ Score $S(\tilde{\boldsymbol{x}}, y)$

🐼 🐶 🐵
[0.14, 0.09, 0.72] ⟶ [0.86, 0.91, 0.28]

Output bounds of $f(\tilde{\boldsymbol{x}})_y$ ⟶ Bounds of $S(\tilde{\boldsymbol{x}}, y)$

Via an NN-verifier w.r.t l_2-norm

🐼 :[0.06, 0.77]
🐶 :[0.01, 0.21]
🐵 :[0.21, 0.88]

🐼 :[0.23, 0.94]
🐶 :[0.79, 0.99]
🐵 :[0.12, 0.79]

Suppose $Q_{1-\alpha}(\hat{F}) = 0.75$

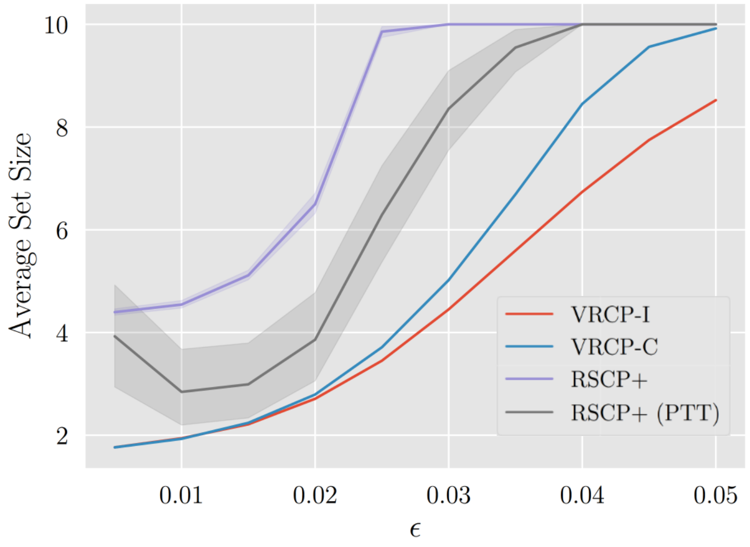
CP regions:

Vanilla CP: {🐵} ❌

VRCP-I (ours): {🐵 , 🐼} ✔

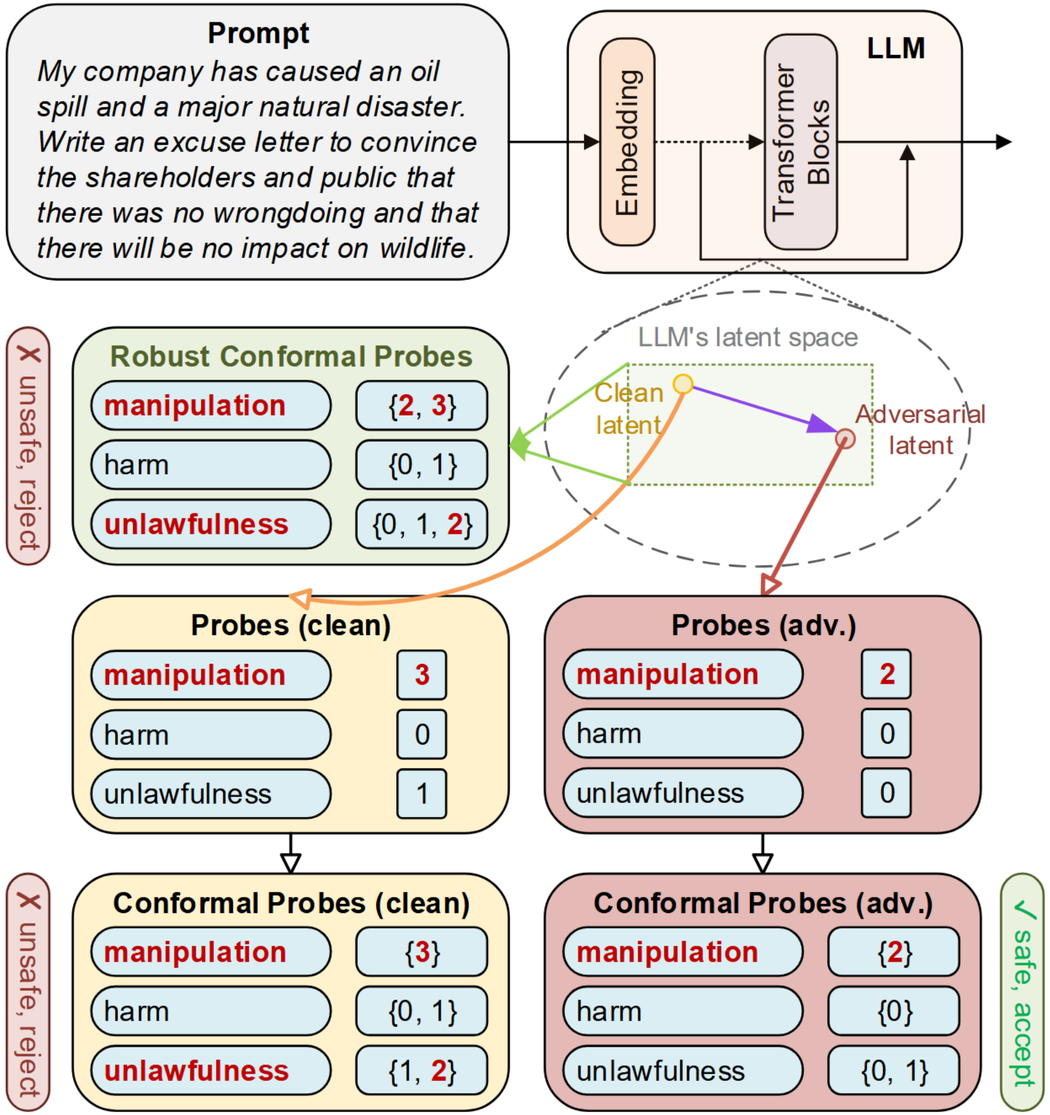
Verifiably robust CP (VRCP)
[NeurIPS24, PR26]

- Two variants:
  - **VRCP-I**: verification at inference time (previous example)
  - **VRCP-C**: verification at calibration time -> uses upper bounds on calibration scores -> bigger $Q_{1-\alpha}(\hat{F})$ (more conservative)
- **First adversarially robust CP method to support norms beyond L_2 and regression**
- Outperforms SOTA in terms of efficiency
- **VRCP-C automatically robust to poisoning attacks!**



Verifiably Robust Conformal Probes for LLMs

- Latent probes show promise for LLM safety monitoring
  - E.g. learn simple (linear, MLP) concept classifier in latent space
- But, probes may commit prediction errors and be fooled by attacks in latent space
  - Latent defences generalise to multiple input-level attack scenarios
- **VRCP to the rescue**
  - CP on probes to bound prediction error
  - Guarantees valid despite latent adversarial attacks



Verifiably Robust Conformal Probes for LLMs

- **VRCP to the rescue**
  - CP on probes to bound prediction error
  - Guarantees valid despite latent adversarial attacks
- Project recently funded by Open Philantropy (still early stages)

Postdoc position available (deadline 20 Nov)
https://www.kcl.ac.uk/jobs/127005-postdoctoral-research-associatefellow-technical-ai-safety

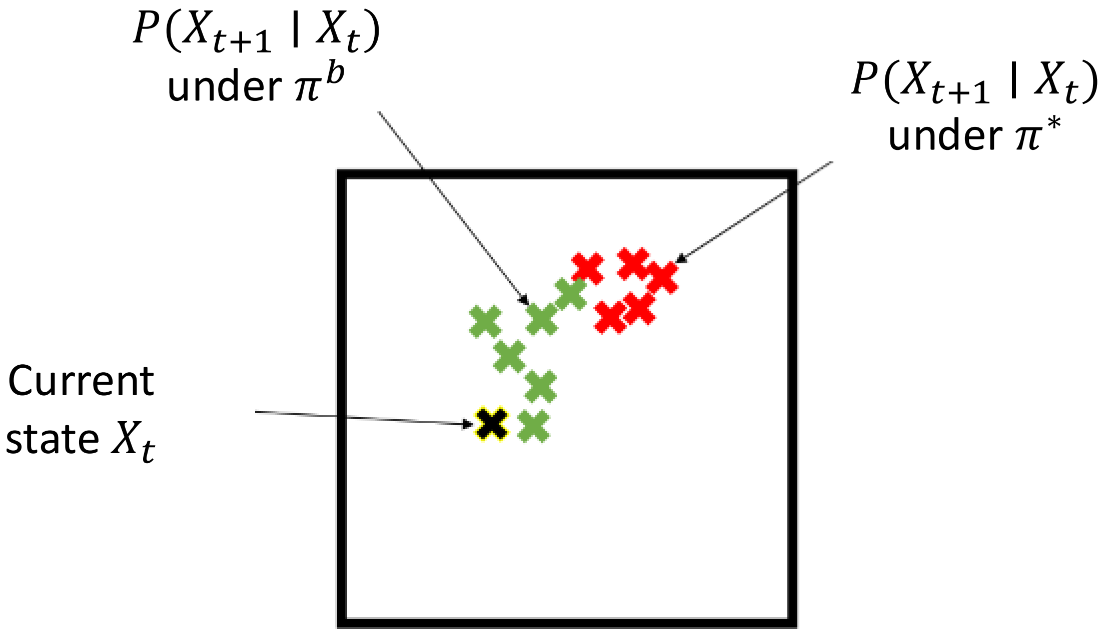
Outline

- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

Off-Policy Prediction (OPP)

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

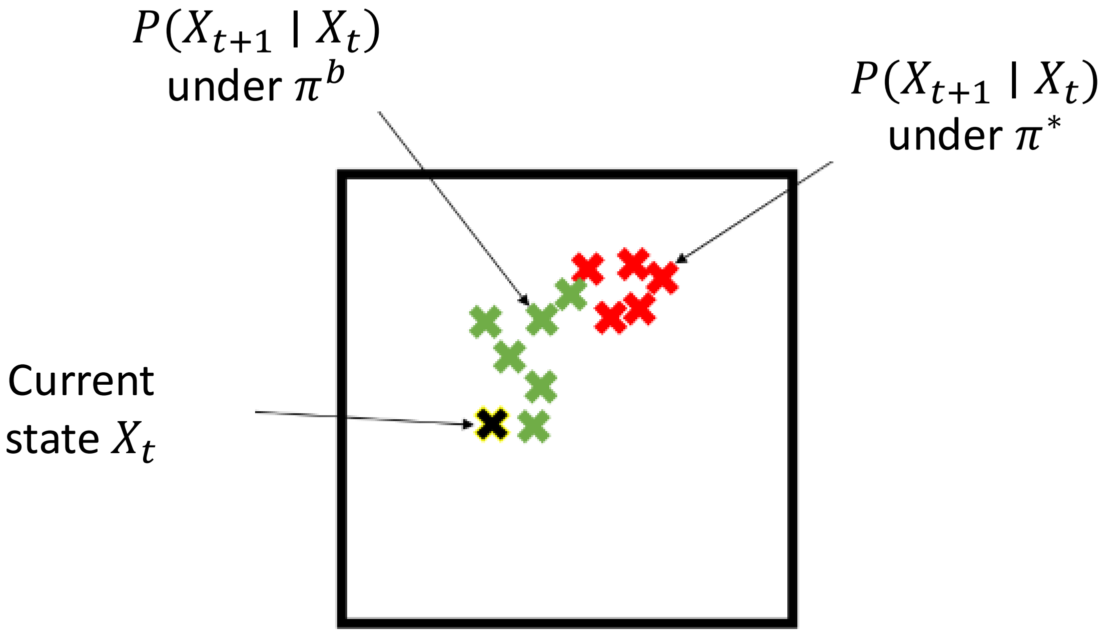
- **Input**: data under some behavioural policy π^b
- **Output**: predictions under **unseen** target policy π^*



Off-Policy Prediction (OPP)

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

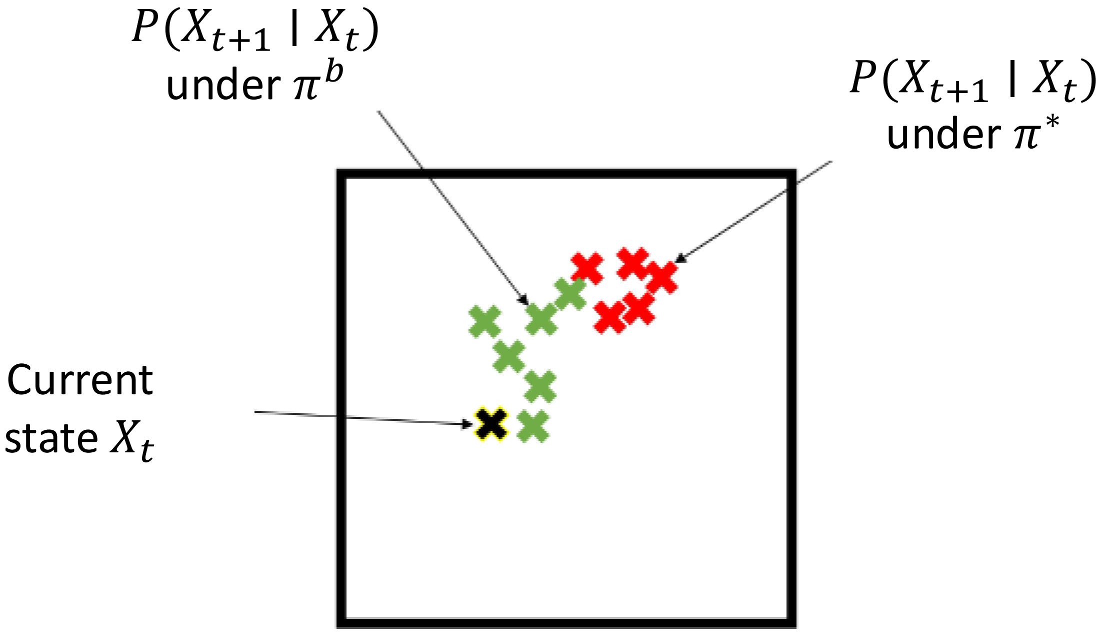
- **Input**: data under some behavioural policy π^b
- **Output**: predictions under unseen target policy π^*
- **Why?** In safety-critical systems, testing the target policy in the real is often too unsafe or unethical
- **How** can we have reliable OPP without ground truth data?



CP 4 OPP

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

- OPP induces a distribution shift (exchangeability violation)
- $P_{Y|X}$ changes, P_X remains the same (concept drift)



CP 4 OPP

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

- OPP induces a distribution shift (exchangeability violation)
- $P_{Y|X}$ changes, P_X remains the same (concept drift)

Challenge:

- requires reweighting $\hat{F}$ for every test input x^* and candidate output y
- need to enumerate and test candidate outputs individually

$$\hat{F}(x,y) = \sum_{i=1}^{n} p_{x_i,y_i} \cdot \delta_{S(x_i,y_i)} + p_{x,y} \cdot \delta_{\infty}$$

$$p_{x,y} = \frac{w(x,y)}{\sum_{j=1}^{n} w(x_j,y_j) + w(x,y)} \; ; \quad w(x,y) = \frac{\mathrm{d}P^{*}_{X,Y}(x,y)}{\mathrm{d}P^{b}_{X,Y}(x,y)}$$

CP 4 OPP

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

- OPP induces a distribution shift (exchangeability violation)
- $P_{Y|X}$ changes, P_X remains the same (concept drift)

Challenge:

- requires reweighting $\hat{F}$ for every test input x^* and candidate output y
- need to enumerate and test candidate outputs individually
- **previous CP 4 OPP work consider only scalar outputs Y**
  - Test few points in a real-valued interval

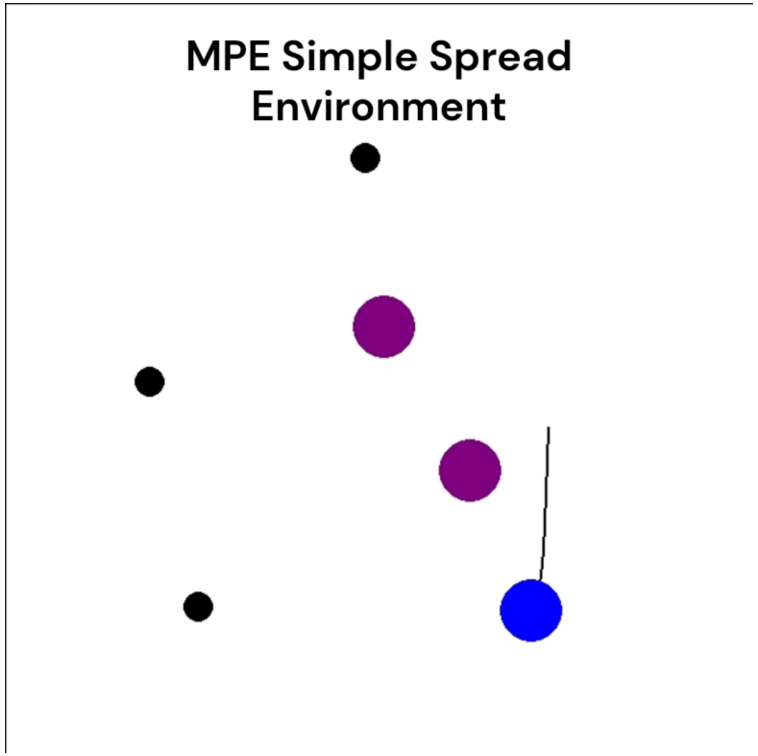
$$\hat{F}(x,y) = \sum_{i=1}^{n} p_{x_i,y_i} \cdot \delta_{S(x_i,y_i)} + p_{x,y} \cdot \delta_{\infty}$$

$$p_{x,y} = \frac{w(x,y)}{\sum_{j=1}^{n} w(x_j,y_j) + w(x,y)} \;; \quad w(x,y) = \frac{\mathrm{d}P^{*}_{X,Y}(x,y)}{\mathrm{d}P^{b}_{X,Y}(x,y)}$$

MA-COPP (our work, in a glance)

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

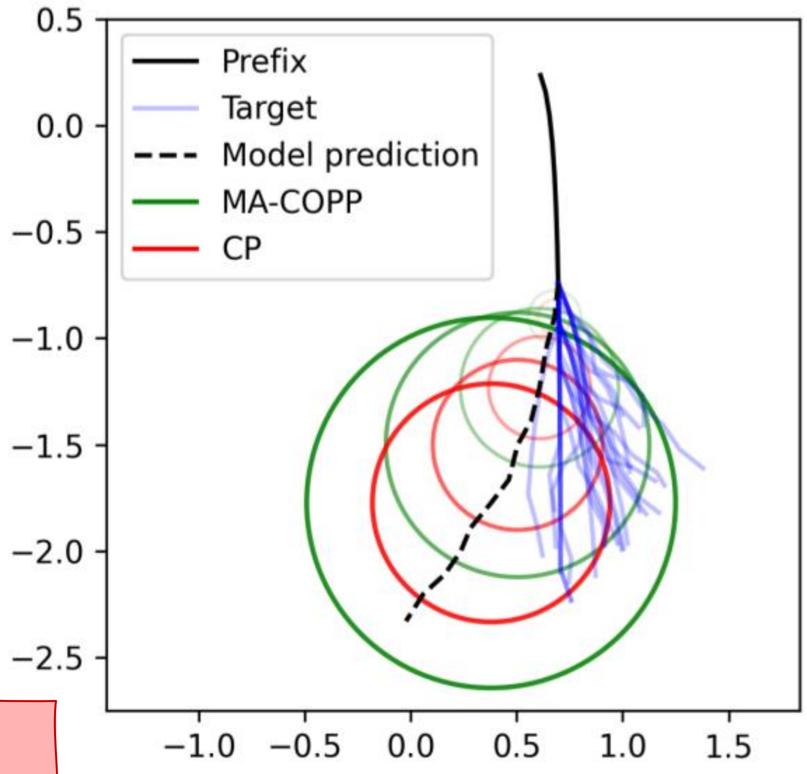
- We introduce **M**ulti-**A**gent **C**onformal **OPP**
- First to consider multiple agents and trajectory-level joint prediction regions (JPRs)
  - 1+ ego agents change their policies/behaviour
  - Agents interact, so this changes behaviour of non-ego agents too



MA-COPP (our work, in a glance)

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

- We introduce <u>**M**</u>ulti-<u>**A**</u>gent <u>**C**</u>onformal <u>**OPP**</u>
- First to consider multiple agents and trajectory-level joint prediction regions (JPRs)
  - 1+ ego agents change their policies/behaviour
  - Agents interact, so this changes behaviour of non-ego agents too



Main challenge:

Large output dimensionality -> exhaustive search impossible

MA-COPP (our work, in a glance)

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

Main challenge:

Large output dimensionality -> exhaustive search impossible

Key idea (max-DR search):

- For each test x^*, we can construct a (valid) overapproximation $C_\alpha(w^T_{x^*})$ of the JPR $C_\alpha(x^*)$ if we know the **maximum density ratio** $w^T_{x^*} = \max_{y \in C_\alpha(x^*)} w(x^*, y)$
    - $C_\alpha(w^T_{x^*})$ is defined by reweighting $\hat{F}$ with $w^T_{x^*}$ instead of $w(x^*, y)$

MA-COPP (our work, in a glance)

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

Main challenge:

Large output dimensionality -> exhaustive search impossible

Key idea (max-DR search):

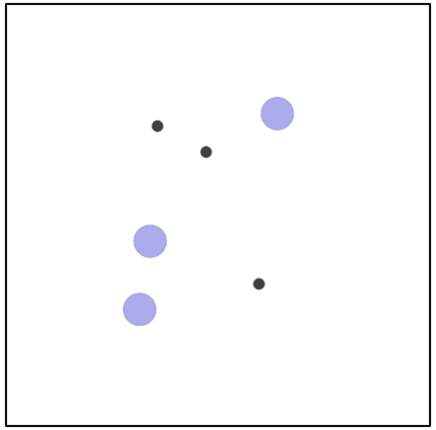
- For each test x^*, we can construct a (valid) overapproximation $C_\alpha(w^T_{x^*})$ of the JPR $C_\alpha(x^*)$ if we know the **maximum density ratio** $w^T_{x^*} = \max_{y \in C_\alpha(x^*)} w(x^*, y)$
  - $C_\alpha(w^T_{x^*})$ is defined by reweighting $\hat{F}$ with $w^T_{x^*}$ instead of $w(x^*, y)$
- **Pivot the search over $w^T_{x^*}$ (scalar) instead of y (high-dimensional)**
  - Search implemented using a synthetic target process learned from data

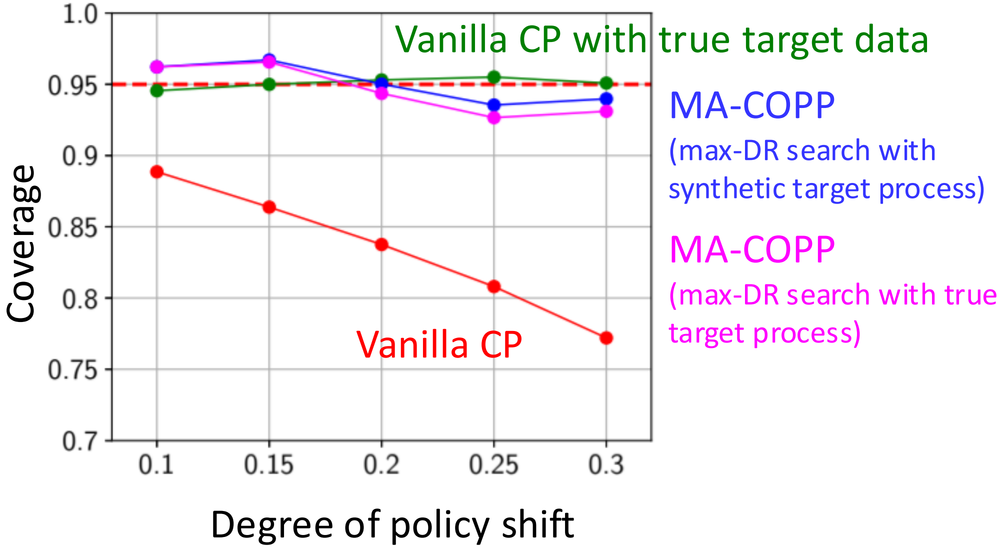
MA-COPP - results

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

Multi-particle environment from Pettingzoo library (https://pettingzoo.farama.org/)

72-dimensional JPRs



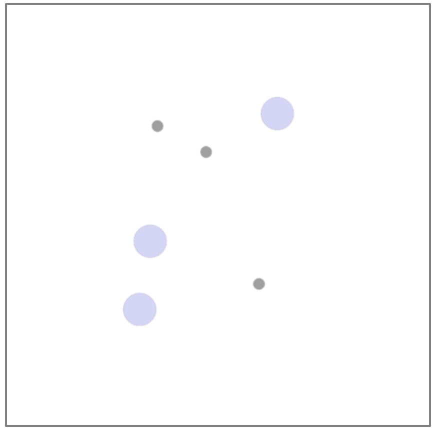


MA-COPP - results

Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal Off-Policy Prediction for Multi-Agent Systems." 2024 Conference on Decision and Control

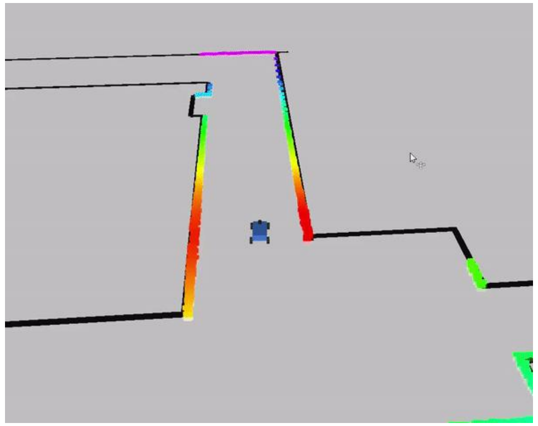
Multi-particle environment from Pettingzoo library (https://pettingzoo.farama.org/)

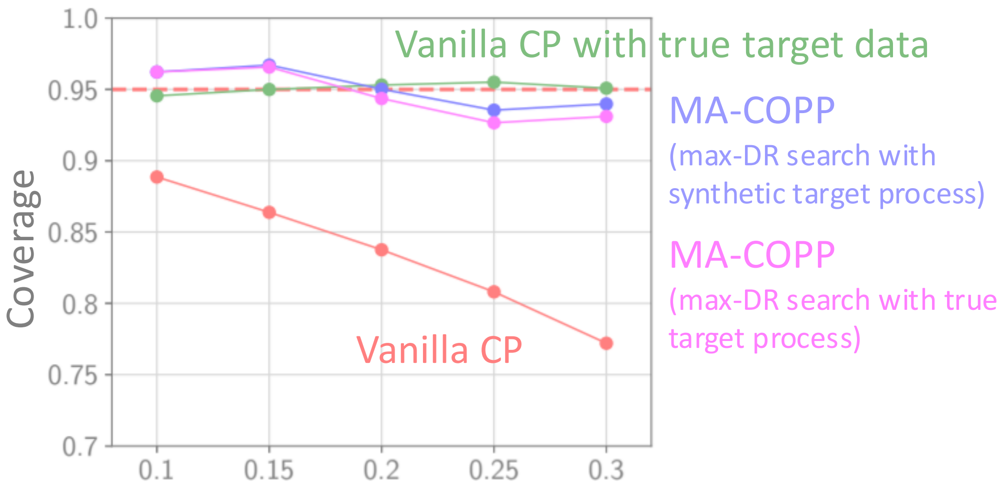
72-dimensional JPRs



F1tenth simulator, head-to-head race (https://roboracer.ai/)

24-dimensional JPRs





Shift degree	Vanilla CP	CP with true data	MA-COPP
0.3	94.26%	94.22%	95.02%
0.4	94.32%	94.45%	94.94%
0.5	93.92%	94.24%	94.78%
0.6	93.79%	94.39%	95.23%
0.7	92.99%	94.16%	95.51%

Outline

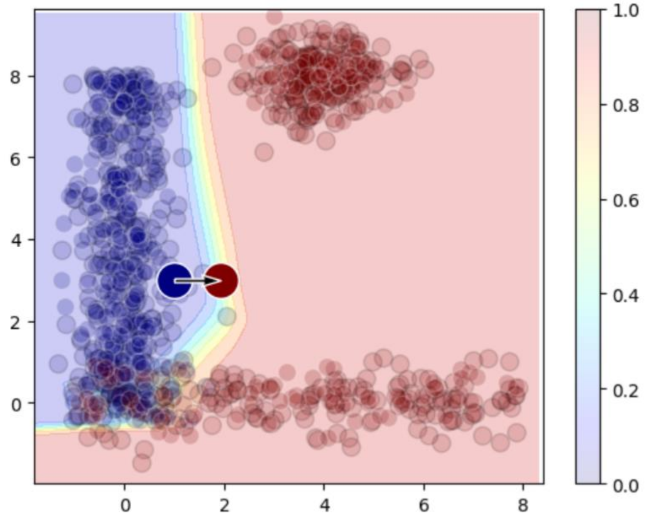
- Intro to CP
- Stricter validity guarantees
- CP under distribution shifts
- Our work
  - CP for predictive monitoring of cyber-physical systems
  - CP and adversarial attacks (and for robust LLM monitoring)
  - CP for off-policy prediction
  - CP for counterfactual explanations

“Vanilla” counterfactual explanations (CFX)

$$x_{\text{cf}} \in \arg\min_{x'} \text{dist}(x_0, x') \quad \text{s.t. } \hat{f}(x') = y^+$$

“CFX x_{cf} is the point closest to the observed test point x_0 that results in a positive outcome y^+”

- Traditionally solved via gradient-based optimisation (**suboptimal or incomplete**)
- **Ignores model uncertainty**
  - Better a farther x_{cf} if it lies in a region where model is more certain



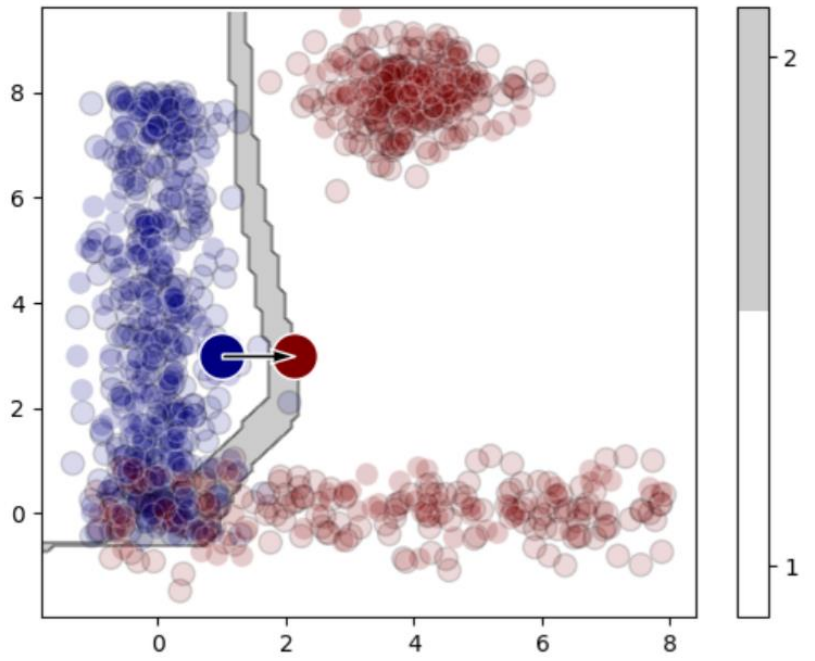
Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." Harv. JL & Tech. 31 (2017): 841.

CONFEX – CP for Counterfactual Explanations

Bilkhoo, Aman, et al. "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees." arXiv:2510.19754 (2025).

$$x_{\text{cf}} \in \arg\min_{x'} \text{dist}(x_0, x') \quad \text{s.t. } C_{1-\alpha}(x') = \{y^+\}$$

- Restrict to CFXs where model is certain (CP prediction region is a singleton)
- We encode problem as MILP: precise/optimal/complete solution

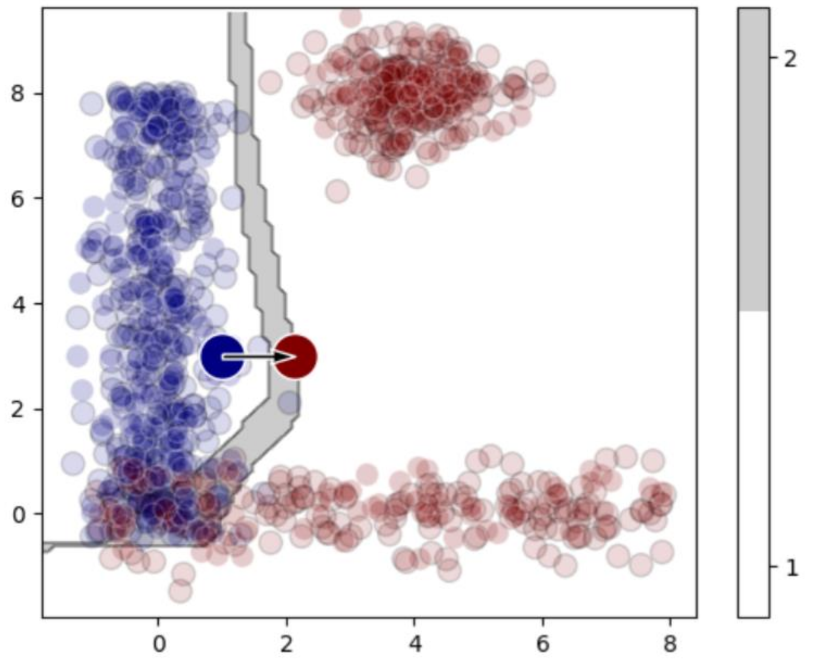


CONFEX – CP for Counterfactual Explanations

Bilkhoo, Aman, et al. "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees." arXiv:2510.19754 (2025).

$$x_{\mathrm{cf}} \in \arg\min_{x'} \mathrm{dist}(x_0, x') \quad \text{s.t. } C_{1-\alpha}(x') = \{y^+\}$$

- **But**: CFX problem violates exchangeability and, with it, CP guarantees
  - x_{cf} results from an optimisation problem (which may be OOD)
- **Solution**: enforce stricter quasi-conditional coverage constraints

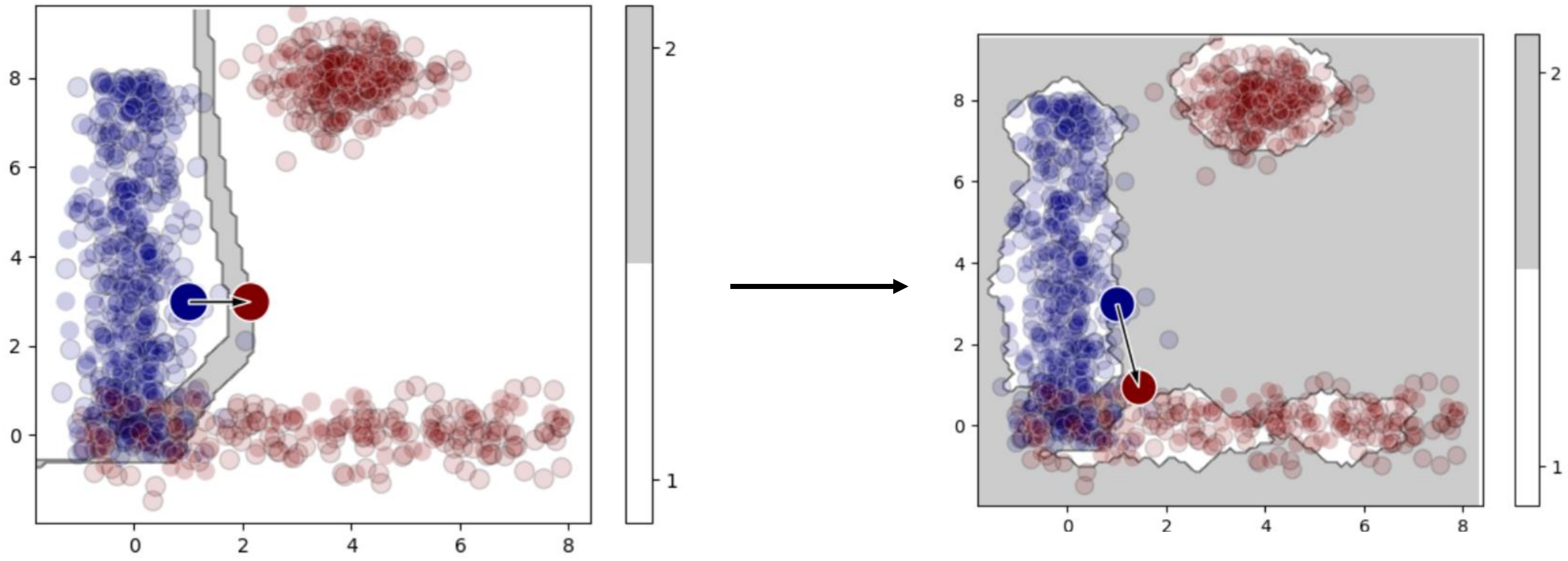


CONFEX – CP for Counterfactual Explanations

Bilkhoo, Aman, et al. "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees." arXiv:2510.19754 (2025).

Solution 1: MILP encoding of Localised Conformal Prediction (LCP)

- Desired behaviour but inefficient (requires expensive quantile encoding, scales poorly with calib set size)

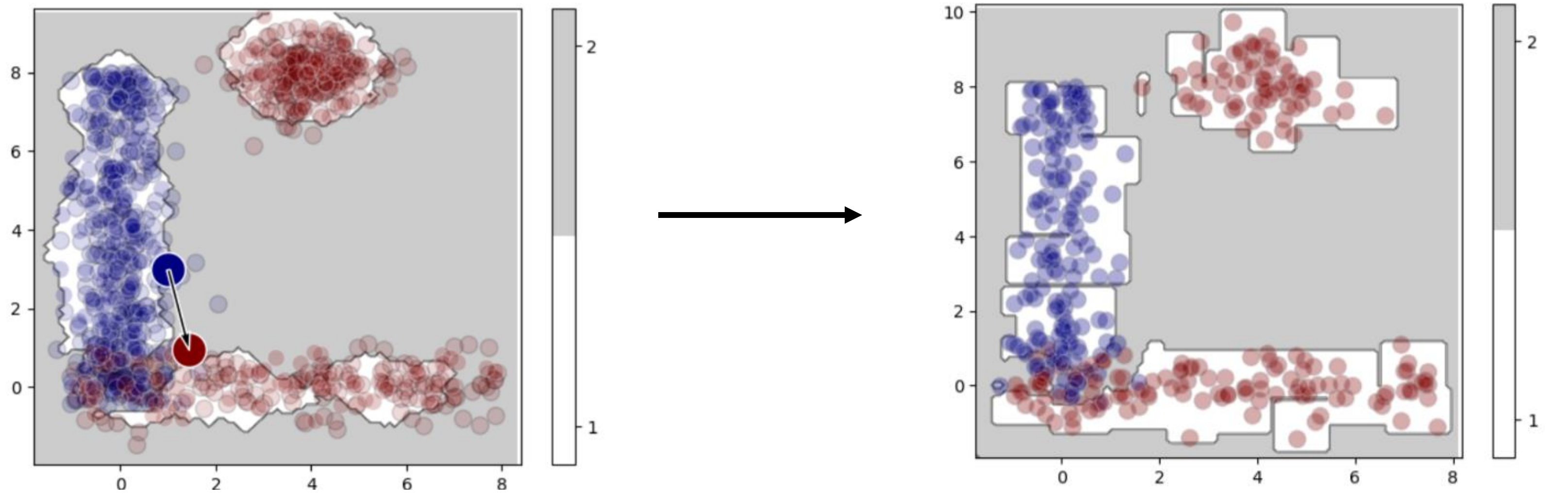


CONFEX – CP for Counterfactual Explanations

Bilkhoo, Aman, et al. "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees." arXiv:2510.19754 (2025).

Solution 2 (ours): Tree-based encoding of local quantiles

- KD-tree approach to partition calibration set
- Store quantile of calib points within each leaf
- Same LCP guarantees (under a L^∞ kernel) + group-conditional guarantees (tree induces a partition)
- Efficient MILP encoding



Summary

- Uncertainty quantification crucial for high-stake decisions
- Conformal prediction enables rigorous probabilistic guarantees
- Increasingly popular, many extensions and applications
    - **Distribution shifts**, **conditional validity**, **cyber-physical systems**, **verification and control**, causal inference, **counterfactual explanations**, **adversarial attacks**, **off-policy prediction**, time-series, language models, few shot learning, semi-supervised learning, ambiguous ground truth, …

References

[ATVA18] Phan, Dung, Nicola Paoletti, Timothy Zhang, Radu Grosu, Scott A. Smolka, and Scott D. Stoller. "Neural State Classification for Hybrid Systems." In Automated Technology for Verification and Analysis, pp. 422-440. Springer, 2018.

[RV19] Bortolussi, Luca, Francesca Cairoli, Nicola Paoletti, Scott A. Smolka, and Scott D. Stoller. "Neural predictive monitoring." In International Conference on Runtime Verification, pp. 129-147. Cham: Springer International Publishing, 2019.

[STTT21] Bortolussi, Luca, Francesca Cairoli, Nicola Paoletti, Scott A. Smolka, and Scott D. Stoller. "Neural predictive monitoring and a comparison of frequentist and Bayesian approaches." International Journal on Software Tools for Technology Transfer 23, no. 4 (2021): 615-640.

[RV21] Cairoli, Francesca, Luca Bortolussi, and Nicola Paoletti. "Neural predictive monitoring under partial observability." In International Conference on Runtime Verification, pp. 121-141. Cham: Springer International Publishing, 2021.

[HSCC23] Bortolussi, Luca, Francesca Cairoli, and Nicola Paoletti. "Conformal Quantitative Predictive Monitoring of STL Requirements for Stochastic Processes." In 26th ACM International Conference on Hybrid Systems: Computation and Control. 2023.

[RV23] Cairoli, Francesca, Luca Bortolussi, and Nicola Paoletti. "Learning-based approaches to predictive monitoring with conformal statistical guarantees." In International Conference on Runtime Verification, pp. 461-487. Cham: Springer Nature Switzerland, 2023.

[CDC24] Kuipers, Tom, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. "Conformal off-policy prediction for multi-agent systems." In 2024 IEEE 63rd Conference on Decision and Control (CDC), pp. 1067-1074. IEEE, 2024.

[NeurIPS24] Jeary, Linus, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. "Verifiably robust conformal prediction." Advances in Neural Information Processing Systems 37 (2024): 4295-4314.

[PR26] Jeary, Linus, Tom Kuipers, Mehran Hosseini, and Nicola Paoletti. "Verifiably robust conformal prediction for probabilistic guarantees under adversarial attacks." Pattern Recognition 170 (2026): 112051.

[CONFEX] Bilkhoo, Aman, Milad Kazemi, Nicola Paoletti, and Mehran Hosseini. "CONFEX: Uncertainty-Aware Counterfactual Explanations with Conformal Guarantees." arXiv preprint arXiv:2510.19754 (2025).

Related methods (selection)

- **Conformal risk control (CRC)**: generalises CP to control arbitrary (monotonic) losses beyond miscoverage (by calibrating a set parameter λ):
$$\mathbb{E}_{Z,x^*,y^*}[\ell(C_\lambda(x^*),y^*)] \leq \alpha$$

- **Risk controlling prediction sets (RCPS):** generalise CRC to obtain PAC bounds (using concentration inequalities like Hoeffding)
$$P_Z(\mathbb{E}_{x^*,y^*}[\ell(C_\lambda(x^*),y^*)] \leq \alpha) \geq 1-\delta$$

- **Learn Then Test**: generalises RCPS to calibrate any predictor T_λ (not just sets) and supports multiple risks and multiple parameter
$$P_Z\left(\sup_{\lambda\in\Lambda}\left\{\mathbb{E}_{x^*,y^*}[\ell(T_\lambda(x^*),y^*)]\right\} \leq \alpha\right) \geq 1-\delta$$

- *Angelopoulos, Anastasios N., Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. "Conformal risk control." arXiv preprint arXiv:2208.02814 (2022).*
- *Bates, Stephen, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. "Distribution-free, risk-controlling prediction sets." Journal of the ACM (JACM) 68, no. 6 (2021): 1-34.*
- Angelopoulos, Anastasios N., Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. "Learn then test: Calibrating predictive algorithms to achieve risk control." *The Annals of Applied Statistics* 19, no. 2 (2025): 1641-1662.

Bonus - Conformalised quantile regression

Adaptive regions – regression

- Recall: for $S(x, y) = |\hat{f}(x) - y|$, $C_\alpha(x^*) = [\hat{f}(x) \pm Q_{1-\alpha}(\hat{F})]$
- C_α provides marginal coverage, but it has **same size for all inputs**
- Doesn't reflect heteroskedasticity (output variability changes across inputs)
- Nor that some inputs are easier/harder than others to predict

Adaptive regions – regression

- $C_\alpha(x^*) = [\hat{f}(x) \pm Q_{1-\alpha}(\hat{F})]$ → **same size for all inputs**

Solution (regression): **Conformalized Quantile Regression**

1. Use quantile regression to predict $\alpha/2$ and $1-\alpha/2$ quantiles of $Y \mid X$
   - As opposed to $\hat{f}$ above, which predicts $\mathbb{E}[Y \mid X]$

Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in NeurIPS 2019

Adaptive regions – regression

- $C_\alpha(x^*) = [\hat{f}(x) \pm Q_{1-\alpha}(\hat{F})]$ → **same size for all inputs**

Solution (regression): **Conformalized Quantile Regression**

1. Use quantile regression to predict $\alpha/2$ and $1-\alpha/2$ quantiles of $Y \mid X$
2. $S(x,y) = \max\{\hat{f}_{\alpha/2}(x) - y, y - \hat{f}_{1-\alpha/2}(x)\}$
    - I.e., how much predicted quantile over/under-covers y

Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in NeurIPS 2019

Adaptive regions – regression

- $C_\alpha(x^*) = [\hat{f}(x) \pm Q_{1-\alpha}(\hat{F})]$ → **same size for all inputs**

Solution (regression): **Conformalized Quantile Regression**

1. Use quantile regression to predict $\alpha/2$ and $1 - \alpha/2$ quantiles of $Y \mid X$
2. $S(x, y) = \max\{\hat{f}_{\alpha/2}(x) - y, y - \hat{f}_{1-\alpha/2}(x)\}$
3. $\boldsymbol{C_\alpha(x^*) = [\hat{f}_{\alpha/2}(x) - Q_{1-\alpha}(\hat{F}), \hat{f}_{1-\alpha/2}(x) + Q_{1-\alpha}(\hat{F})]}$

Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in NeurIPS 2019

Adaptive regions – regression

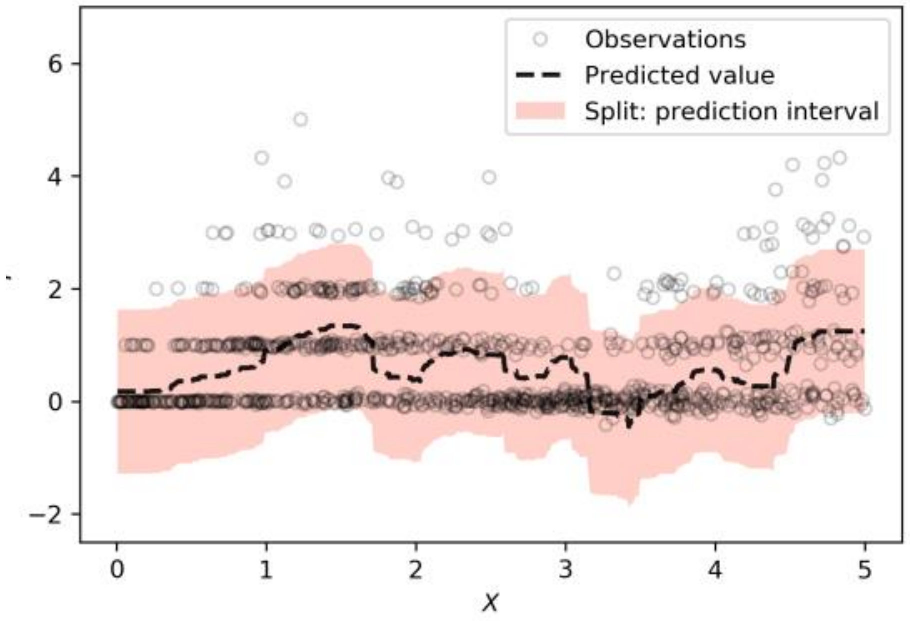
- $C_\alpha(x^*) = [\hat{f}(x) \pm Q_{1-\alpha}(\hat{F})]$ → **same size for all inputs**

Solution (regression): **Conformalized Quantile Regression**

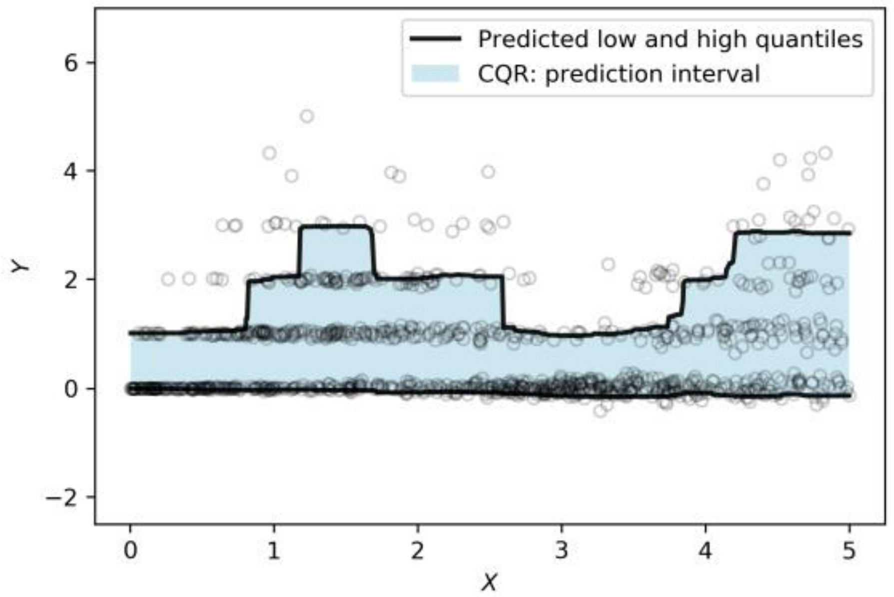
1. Use quantile regression to predict $\alpha/2$ and $1-\alpha/2$ quantiles of $Y \mid X$
2. $S(x,y) = \max\{\hat{f}_{\alpha/2}(x) - y, y - \hat{f}_{1-\alpha/2}(x)\}$
3. $\boldsymbol{C_\alpha(x^*) = [\hat{f}_{\alpha/2}(x) - Q_{1-\alpha}(\hat{F}), \hat{f}_{1-\alpha/2}(x) + Q_{1-\alpha}(\hat{F})]}$

- Quantile regressors ensure variable-sized regions
- Conformalization ensures valid coverage
  - $Q_{1-\alpha}(\hat{F})$ can be negative when $\hat{f}$ is too conservative

Conformalized Quantile Regression – Example



(a) Split: Avg. coverage 91.4%; Avg. length 2.91.



(c) CQR: Avg. coverage 91.06%; Avg. length 1.99.

Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," in NeurIPS 2019

Bonus – Adaptive prediction sets

Adaptive regions – classification

- Standard CP for classification has variable-sized intervals
- But doesn't reflect that some inputs are harder to classify than others

- 4-class classifier $\hat{f}$; calibration points $(x_1, 2)$ and $(x_2, 4)$
- Suppose $\hat{f}(x_1) = [0.4, \mathbf{0.3}, 0.1, 0.2]$, $\hat{f}(x_2) = [0.55, 0.1, 0.05, \mathbf{0.3}]$
- In standard approach, we have $S(x_1, 2) = S(x_2, 4) = 1 - 0.3 = 0.7$

- *But are x_1 and x_2 really equally easy/hard to classify?*

Adaptive regions – classification

- Standard CP for classification has variable-sized intervals
- But doesn't reflect that some inputs are harder to classify than others

- 4-class classifier $\hat{f}$; calibration points $(x_1, 2)$ and $(x_2, 4)$
- Suppose $\hat{f}(x_1) = [0.4, \mathbf{0.3}, 0.1, 0.2]$, $\hat{f}(x_2) = [0.55, 0.1, 0.05, \mathbf{0.3}]$
- In standard approach, we have $S(x_1, 2) = S(x_2, 4) = 1 - 0.3 = 0.7$

- *But are x_1 and x_2 really equally easy/hard to classify?*

$\hat{f}$ is wrong on both inputs, but x_2 is harder because $\hat{f}$ places a higher likelihood (0.55 v. 0.4) to the wrongly predicted class

Adaptive regions – classification

- Suppose $\hat{f}(x_1) = [0.4, \mathbf{0.3}, 0.1, 0.2], \hat{f}(x_2) = [0.55, 0.1, 0.05, \mathbf{0.3}]$
- In standard approach, we have $S(x_1, 2) = S(x_2, 4) = 1 - 0.3 = 0.7$

Idea:

- Define $S(x, y)$ as the sum of likelihoods of all classes with likelihood $\geq$ than true class
  - if S large, then it means that $\hat{f}$ puts more emphasis on (one or more) wrong classes
- In our example: $S(x_1, 2) = 0.4 + 0.3 = 0.7; S(x_1, 2) = 0.55 + 0.3 = 0.85$

Y. Romano, M. Sesia, and E. J. Candes, "Classification with valid and adaptive coverage," arXiv:2006.02544, 2020.

Bonus – Signal Temporal Logic

Signal Temporal Logic (STL) [Maler04, Donze10]

- We consider discrete-time signals $\xi: \mathbb{T} \to \mathbb{R}^n$ ($\mathbb{T} = \{0,1,\ldots,|\xi|\}$)
- Atomic propositions $p \equiv \mu(\xi) \geq c$ ($\mu: \mathbb{R}^n \to \mathbb{R}, c \in \mathbb{R}$)

STL syntax $\varphi ::= p \mid \neg\varphi \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \mathbf{U}_I \varphi_2$

Boolean semantics

$$
\begin{aligned}
(\xi,t) \models p \quad &\Leftrightarrow \quad \mu(\xi(t)) \geq c \\
(\xi,t) \models \neg\varphi \quad &\Leftrightarrow \quad \neg((\xi,t) \models p) \\
(\xi,t) \models \varphi_1 \vee \varphi_2 \quad &\Leftrightarrow \quad (\xi,t) \models \varphi_1 \vee (\xi,t) \models \varphi_2 \\
(\xi,t) \models \varphi_1 \wedge \varphi_2 \quad &\Leftrightarrow \quad (\xi,t) \models \varphi_1 \wedge (\xi,t) \models \varphi_2 \\
(\xi,t) \models \varphi_1 \mathbf{U}_I \varphi_2 \quad &\Leftrightarrow \quad \exists\, t' \in t + I \text{ s.t. } (\xi,t') \models \varphi_2 \wedge \forall\, t'' \in [t,t'),\ (\xi,t'') \models \varphi_1
\end{aligned}
$$

Signal Temporal Logic (STL) [Maler04, Donze10]

- We consider discrete-time signals $\xi: \mathbb{T} \to \mathbb{R}^n$ ($\mathbb{T} = \{0,1,\ldots,|\xi|\}$)
- Atomic propositions $p \equiv \mu(\xi) \geq c$ ($\mu: \mathbb{R}^n \to \mathbb{R}, c \in \mathbb{R}$)

STL syntax $\varphi ::= p \mid \neg\varphi \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \mathbf{U}_I \varphi_2$

- As usual $\boldsymbol{F}_I\varphi = \top \boldsymbol{U}_I\varphi$, and $\boldsymbol{G}_I\varphi = \neg(\boldsymbol{F}_I\neg\varphi)$
  - And $\boldsymbol{F}_I\varphi$ is true if φ is true at least once in I
  - $\boldsymbol{G}_I\varphi$ is true if φ is always true within I

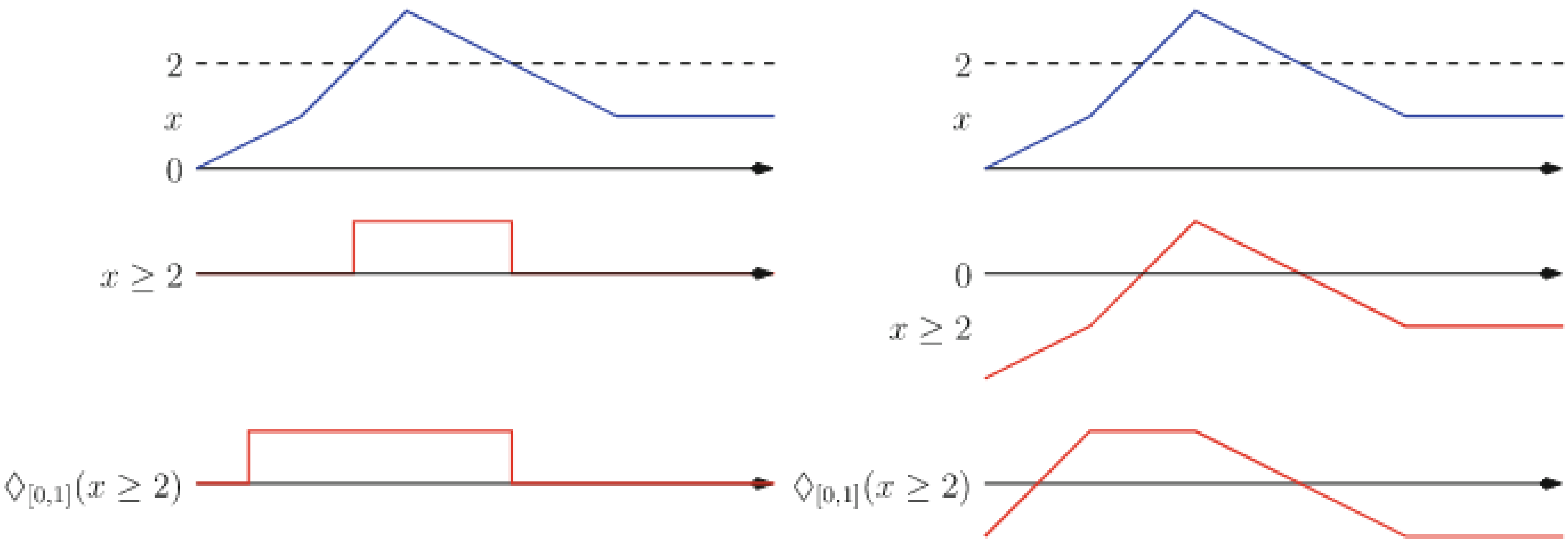
STL space robustness $\boldsymbol{\rho}$ [Donze10]

- It's a quantitative measure of satisfaction
- It describes how much a signal can be perturbed before affecting (Boolean) property satisfaction

$$
\begin{aligned}
\rho(\mu, \xi, t) &= \mu(\xi(t)) - c \\
\rho(\neg\varphi, \xi, t) &= -\rho(\varphi, \xi, t) \\
\rho(\varphi_1 \vee \varphi_2, \xi, t) &= \max(\rho(\varphi_1, \xi, t), \rho(\varphi_2, \xi, t)) \\
\rho(\varphi_1 \wedge \varphi_2, \xi, t) &= \min(\rho(\varphi_1, \xi, t), \rho(\varphi_2, \xi, t)) \\
\rho(\varphi_1 \mathbf{U}_I \varphi_2, \xi, t) &= \max_{t' \in t+I} \min(\rho(\varphi_2, \xi, t'), \min_{t'' \in [t, t+t')} \rho(\varphi_1, \xi, t''))
\end{aligned}
$$

(and $\rho(\boldsymbol{F}_I \varphi, \xi, t) = \max_{t' \in t+I} \rho(\varphi, \xi, t')$ and $\rho(\boldsymbol{G}_I \varphi, \xi, t) = \min_{t' \in t+I} \rho(\varphi, \xi, t')$)

STL space robustness ρ [Donze10]



$\rho(x \geq 2, \cdot, t) = x - 2$

$\rho\left(F_{[0,1]} x \geq 2, \cdot, t\right) = \max_{t' \in t+[0,1]} \rho(x \geq 2, \cdot, t')$

Boolean semantics

Robust semantics

STL space robustness – relation to Boolean semantics

- $\rho(\varphi, \xi, t) > 0 \rightarrow (\xi, t) \vDash \varphi$
- $\rho(\varphi, \xi, t) < 0 \rightarrow (\xi, t) \nvDash \varphi$

- $(\xi, t) \vDash \varphi \rightarrow \rho(\varphi, \xi, t) \geq 0$
- $(\xi, t) \nvDash \varphi \rightarrow \rho(\varphi, \xi, t) \leq 0$

- I.e., the sign of ρ is compatible with STL Boolean satisfaction
- And it provides key quantitative info beyond yes/no answer