IPML

IMPRECISE PROBABILISTIC MACHINE LEARNING

Lecture 9: Conformal Prediction and Calibration

Krikamol Muandet

16 January 2026

Outline

1. Conformal Prediction

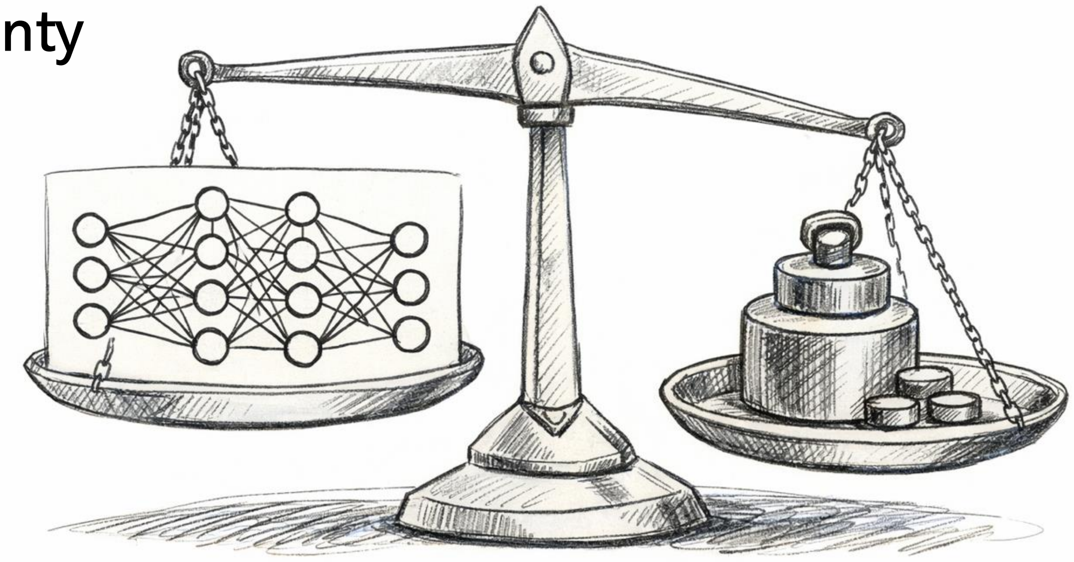
2. Probabilistic Calibration

3. Applications

Conformal Prediction

Conformal Prediction

- Conformal prediction is a *frequentist* approach to **distribution-free** uncertainty quantification that is:
  1. Agnostic to the model
  2. Agnostic to data distribution
  3. Valid in finite sample



Heuristic Uncertainty → **Conformal Prediction** → Rigorous Uncertainty

Conformal Coverage Guarantee

- Given a predictive model $\hat{f}: \mathcal{X} \to \Delta(\mathcal{Y})$ and the test point $(X_{\text{test}}, Y_{\text{test}})$, we seek to construct a **prediction set** $C(X_{\text{test}}) \subset \mathcal{Y}$ that is **valid**, meaning

$$P(Y_{\text{test}} \in C(X_{\text{test}})) \geq 1 - \alpha$$

where $\alpha \in [0,1]$ is a user-chosen error rate.

- **Marginal coverage**: *The probability that the prediction set contains the correct label is almost exactly* $1 - \alpha$.

Conformal Coverage Guarantee

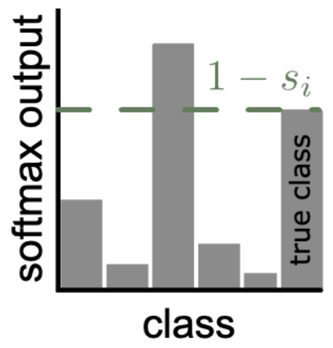
We want the prediction set to contains the correct label **with high probability.**
The prediction set $C(X_{\text{test}})$ captures the model's uncertainty on the data X_{test}.

Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Split Conformal Prediction

Given a **calibration set** $(X_1, Y_1), \dots, (X_n, Y_n)$:

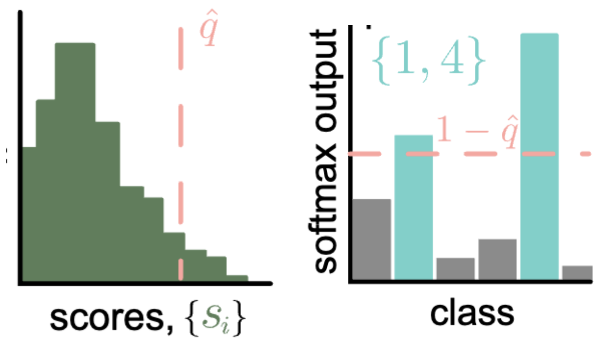
1. Identify a heuristic uncertainty using the pre-trained model
2. Define the **score function** $s(x, y) \in \mathbb{R}$
3. Compute $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the calibration scores



$$s_1, s_2, \dots, s_n := s(X_1, Y_1), s(X_2, Y_2), \dots, s(X_n, Y_n)$$

4. Form the prediction sets of the new example X_{test} as

$$C(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \le \hat{q}\}$$



Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Coverage Property

Suppose $(X_1, Y_1), \ldots, (X_\mathrm{n}, Y_\mathrm{n})$ and $(X_{\text{test}}, Y_{\text{test}})$ are i.i.d and we define $\hat{q}$ as

$$\hat{q} = \inf\left\{q : \frac{|\{i : s(X_i, Y_i) \leq q\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\}$$

The prediction set $C(X) = \{\mathrm{y} : s(X, y) \leq \hat{q}\}$ has a **valid marginal coverage**.

- WOLG, assume that the calibration scores are sorted: $s_1 < s_2 < \cdots < s_n$.
  1. If $\alpha < 1/(n+1)$, $\hat{q} = \infty$. Then, $C(X) = \mathcal{Y}$. **Valid!**
  2. If $\alpha \geq 1/(n+1)$, $\hat{q} = s_{\lceil (n+1)(1-\alpha) \rceil}$.
- Note that $\{Y_{\text{test}} \in C(X_{\text{test}})\} = \{s_{\text{test}} \leq \hat{q}\} = \{s_{\text{test}} \leq s_{\lceil (n+1)(1-\alpha) \rceil}\}$
- By exchangeability, $P(s_{\text{test}} \leq s_k) = k/(n+1)$.
- $P(s_{\text{test}} \leq s_{\lceil (n+1)(1-\alpha) \rceil}) = \lceil (n+1)(1-\alpha) \rceil/(n+1) \geq 1 - \alpha$.

Choice of Score Function

The **score function** $s(x, y) \in \mathbb{R}$ encodes *disagreement* between x and y. The larger the score $s(x, y)$, the worse the agreement between x and y.

While the prediction set may provide a valid marginal coverage, the set can be useless if the score function is uninformative.

Adaptive Prediction Sets

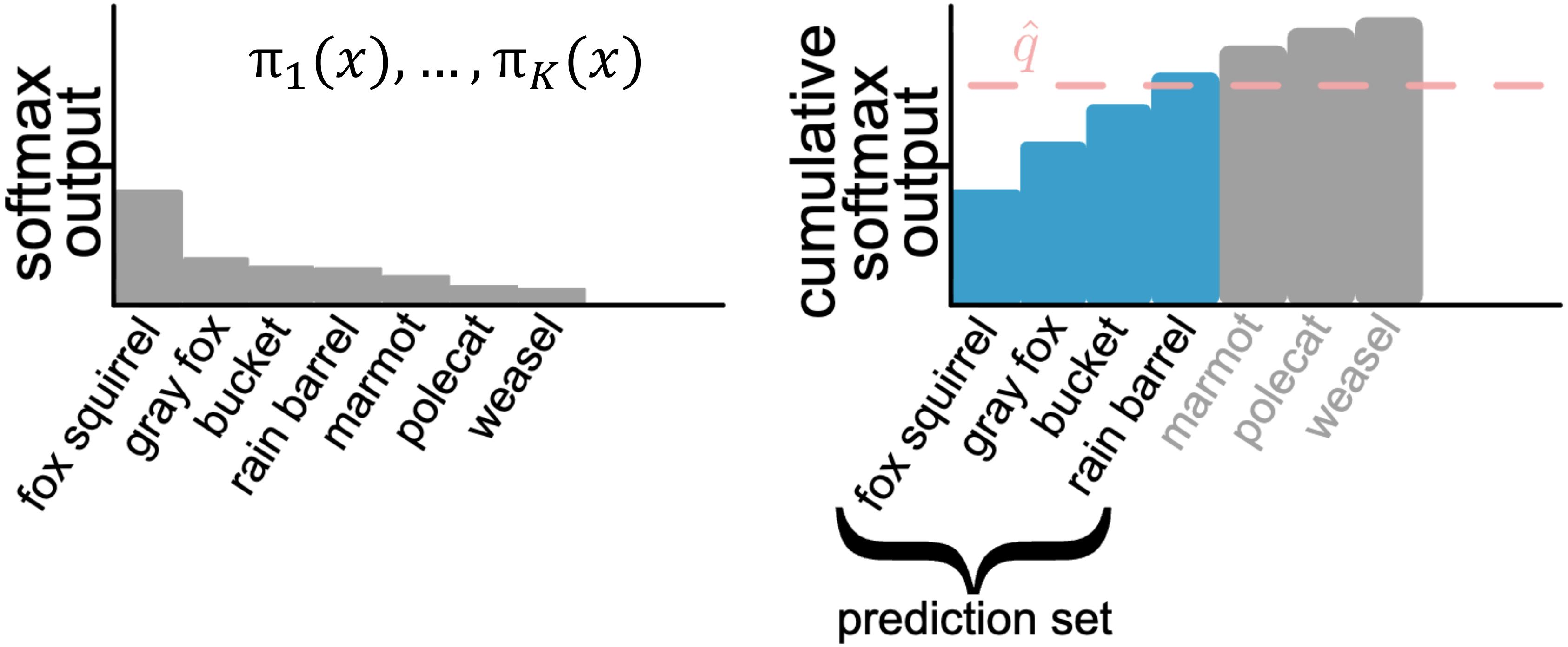
- Assume a predictive model $\hat{f}: \mathcal{X} \to \Delta(\mathcal{Y})$ where $\mathcal{Y} = \{1, \dots, K\}$.
- Let $\pi_1(x), \dots, \pi_K(x)$ be the permutation of $\{1, \dots, K\}$ that sorts $\hat{f}(x)$ from most likely to least likely.
- Then, we define a score function as

$$s(x, y) = \sum_{j=1}^{k} \hat{f}(x)_{\pi_j(x)}, \qquad \text{with} \quad y = \pi_k(x)$$

- Calculate the quantile $\hat{q} = \mathrm{quantile}\left(s_1, \dots, s_n; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$ and form the set

$$C(x) = \{\pi_1(x), \dots, \pi_k(x)\}, \qquad k = \sup\left\{k' : \sum_{j=1}^{k'} \hat{f}(x)_{\pi_j(x)} < \hat{q}\right\} + 1$$

Adaptive Prediction Sets



Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

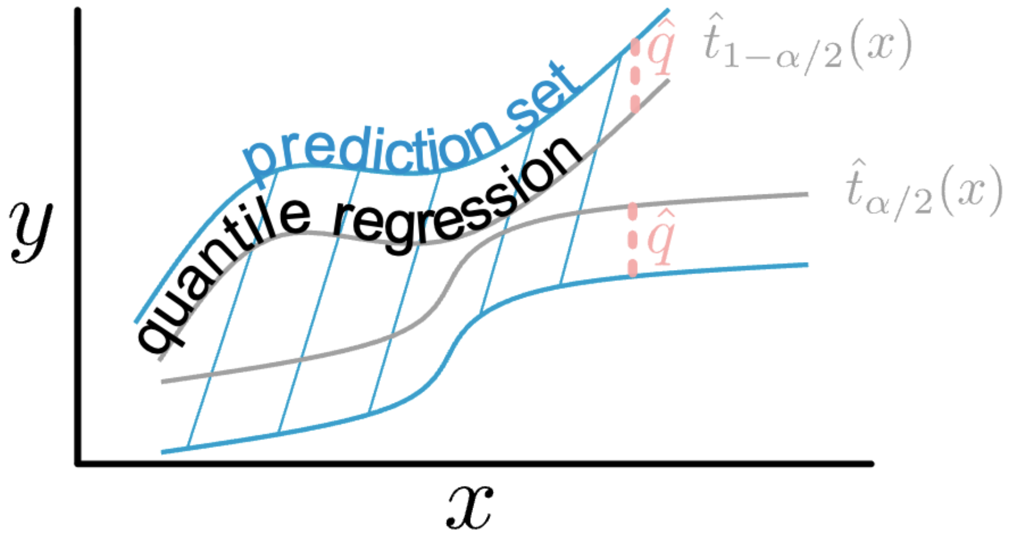
Conformalized Quantile Regression

- **Quantile regression:** learn the γ quantile t_γ of $Y_{\text{text}} \mid X_{\text{text}} = x$
- The quantile $[\hat{t}_{0.05}(x), \hat{t}_{0.95}(x)]$ has approximately 90% coverage.

$$s(x,y) = \max\{\hat{t}_{\alpha/2}(x) - y, y - \hat{t}_{1-\alpha/2}(x)\}$$

$$C(x) = [\hat{t}_{\alpha/2}(x) - \hat{q}, \hat{t}_{1-\alpha/2}(x) + \hat{q}]$$

$$\hat{q} = \text{ Quantile}\left(s_1, \dots, s_n; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$$



Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Conformalized Uncertainty Estimates

Given some **point prediction** $\hat{f}(x)$ and some **uncertainty scalar** $u(x)$, we can provide a conformal guarantee:

- **Standard deviation** $\widehat{\sigma}(x)$: We can assume that

$$Y_{\text{test}} \mid X_{\text{test}} = x \sim N(\mu(x), \sigma(x))$$

  with models $\hat{f}(x)$ and $\widehat{\sigma}(x)$. Conformal prediction gives $\hat{f}(x) \pm \hat{q}\widehat{\sigma}(x)$.

- **Magnitude of the residual** $\hat{r}(x)$: After fitting a model $\hat{f}$, we fit a second model $\hat{r}$ that predicts $\left|y - \hat{f}(x)\right|$.

$$s(x, y) = \frac{\left|y - \hat{f}(x)\right|}{u(x)}, \qquad C(x) = \left[\hat{f}(x) - u(x)\hat{q}, \hat{f}(x) + u(x)\hat{q}\right]$$

Example

Given a trained probabilistic classifier and a calibration set, compute a prediction set $C(x) \subseteq \{A, B, C\}$ for a new input x_{test}, using split conformal in the multiclass classification setting.

- For a new x_{test}, the classifier outputs:
  - $\hat{p}(A \mid x_{\text{test}}) = 0.50$
  - $\hat{p}(B \mid x_{\text{test}}) = 0.35$
  - $\hat{p}(C \mid x_{\text{test}}) = 0.15$

- We want 90% marginal coverage, so $\alpha = 0.1$.

i	**True** y_i	$\hat{p}(y_i \mid x_i)$
1	A	0.80
2	B	0.55
3	C	0.60
4	A	0.40
5	B	0.70
6	C	0.30
7	A	0.90
8	B	0.20
9	C	0.45
10	A	0.65

Example

- Define the non-conformity score as

$$s(x,y) = 1 - \hat{p}(y \mid x)$$

- Let $n = 10$. Then, compute the index

$$k = \lceil (n+1)(1-\alpha) \rceil = 10$$

- Sort $s_1, \dots, s_{10}$ increasingly and set the threshold $\tau = s_{(k)}$ (the *k*-th smallest) [$\tau = 0.8$]

- For each label $y \in \{A, B, C\}$, compute $s(x_{\text{test}}, y)$ and include y in the set if $s(x_{\text{test}}, y) \leq \tau$.

i	True y_i	$\hat{p}(y_i \mid x_i)$	s_i
1	A	0.80	0.20
2	B	0.55	0.45
3	C	0.60	0.40
4	A	0.40	0.60
5	B	0.70	0.30
6	C	0.30	0.70
7	A	0.90	0.10
8	B	0.20	0.80
9	C	0.45	0.55
10	A	0.65	0.35

Sol: $s(x_{\text{test}}, A) = 0.5, s(x_{\text{test}}, B) = 0.65, s(x_{\text{test}}, C) = 0.85 \implies C, (x_{\text{test}}) = \{A, B\}$

Adaptivity

The procedure should return large sets for harder inputs and smaller sets for easier inputs.

{ fox squirrel 0.99 }

{ fox squirrel 0.82, gray fox 0.03, bucket 0.02, rain barrel 0.02 }

{ marmot 0.30, fox squirrel 0.22, mink 0.18, weasel 0.16, beaver 0.03, polecat 0.01 }

Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

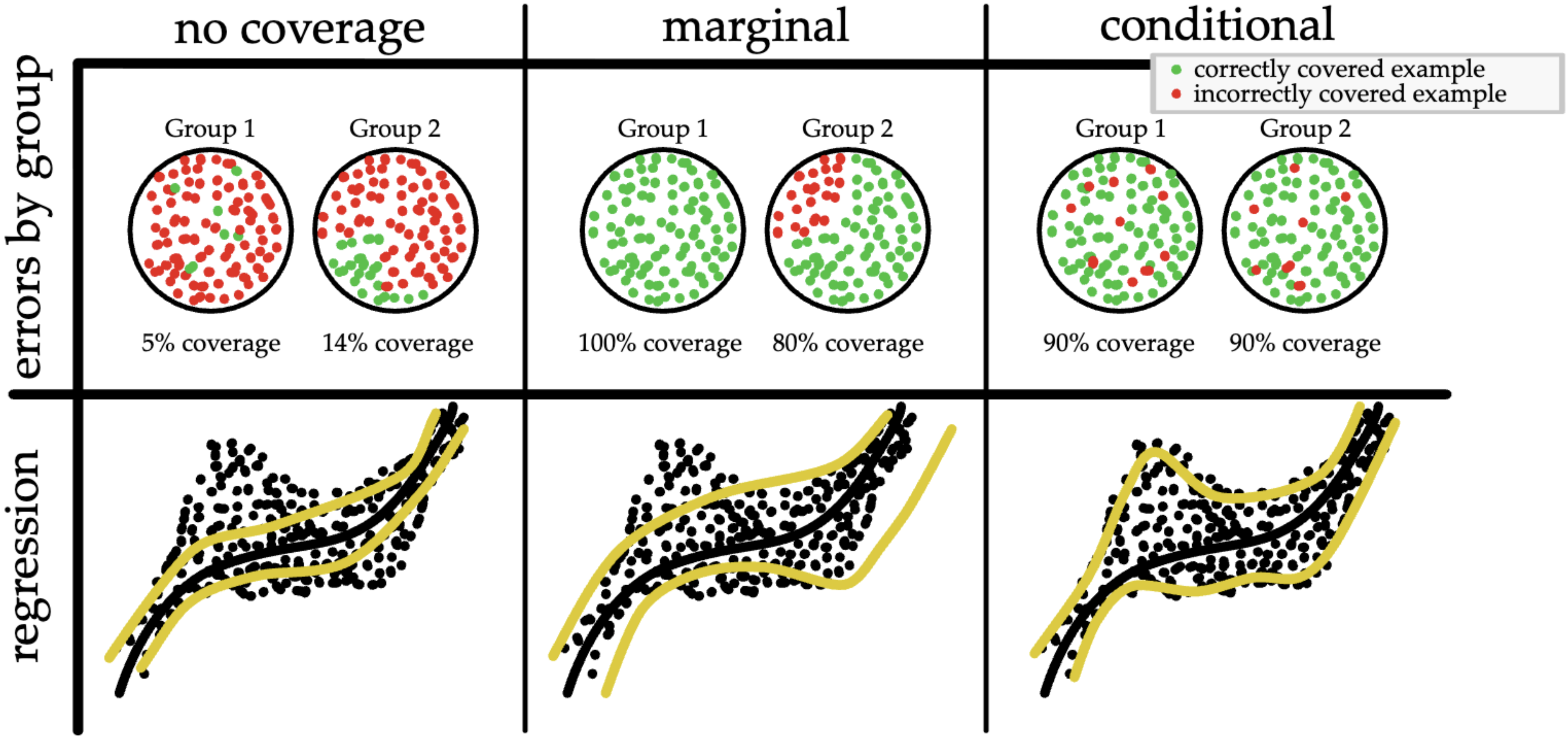
Conditional Coverage

- The **conditional coverage** property:

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid X_{\text{test}}) \geq 1 - \alpha$$

- **Interpretation**: *for every value of the input* X_{test}, *we seek to return prediction sets with* $1 - \alpha$ *coverage*.
  - If we have two groups, A and B, comprising 90% and 10% of the population, conditional coverage requires that the prediction sets cover Y with at least 90% probability within each group.
- This is a stronger property than the *marginal coverage* property and is *impossible* to achieve in general.

Conditional Coverage



Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Extensions

- Group-balanced conformal prediction:

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid X_{\text{test},1} = g) \geq 1 - \alpha, \qquad \forall g \in \{1, \dots, G\}$$

- Class-conditional conformal prediction:

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} = y) \geq 1 - \alpha, \qquad \forall y \in \{1, \dots, K\}$$

- Conformal risk control:

$$\mathbb{E}[\ell(C(X_{\text{test}}), Y_{\text{test}})] \leq \alpha, \qquad \forall y \in \{1, \dots, K\}$$

- Conformal prediction under distribution shift

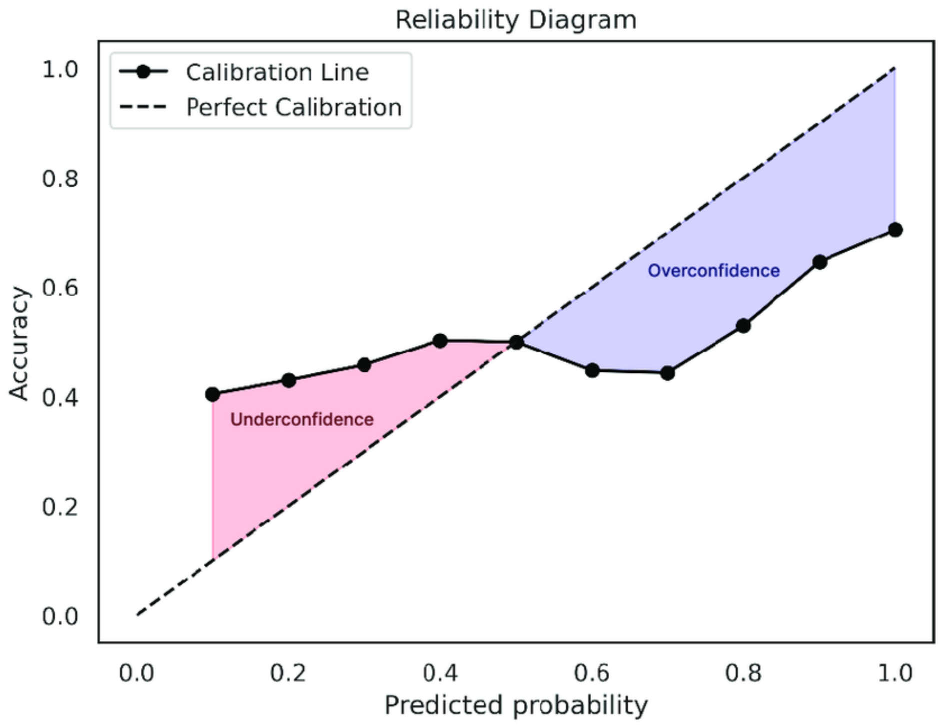
Probabilistic Calibration

Probabilistic Calibration

- Calibration ensures **honesty** in probability estimates overall.
- A probabilistic predictor $\hat{f}$ is *calibrated* if predicted probabilities match empirical frequencies:

$$P\left[Y = 1 \mid \hat{f}(x) = \rho\right] = \rho, \quad \forall \rho \in [0,1]$$

- **Interpretation**: *Among all instances predicted with probability ρ, about ρ fraction of them should be positive.*
- Violations indicate **systematic misrepresentation** of uncertainty.

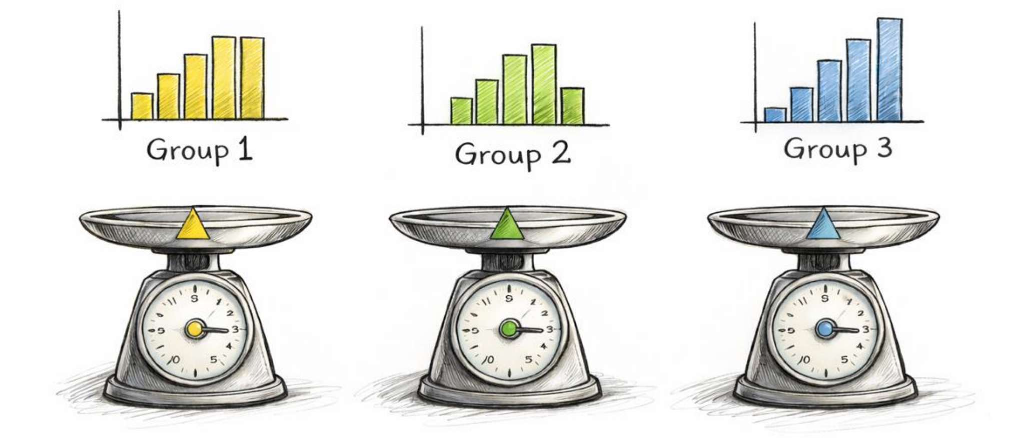


Multi-Calibration

- A predictor is *multi-calibrated* if it is calibrated **simultaneously** across many subpopulations:

$$P\left[Y = 1 \mid \hat{f}(x) = \rho, G = g\right] = \rho, \qquad \forall \rho \in [0,1], \forall g \in \{1, \dots, K\}$$

- Calibration holds not only overall, but for *every relevant group*.



Applications

Multilabel Classification

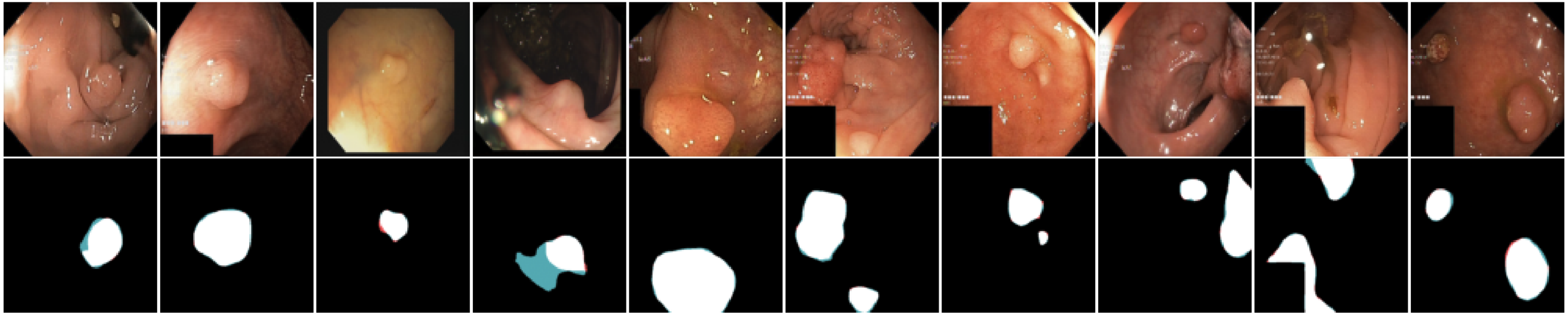
The model's output is thresholded to get the subset of K classes, $C_\lambda(x) = \{y : \hat{f}(x) \geq \lambda\}$. The conformal risk control is used to pick the threshold λ certifying a low false negative rate (FNR).

Red = false negatives, Blue = false positives, Black = true positives

Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Tumor Segmentation

The model's output is thresholded to get the predicted binary mask, $C_\lambda(x) = \{(i,j) : \hat{f}(x)_{(i,j)} \geq \lambda\}$. The conformal risk control is used to pick the threshold λ certifying a low false negative rate (FNR).

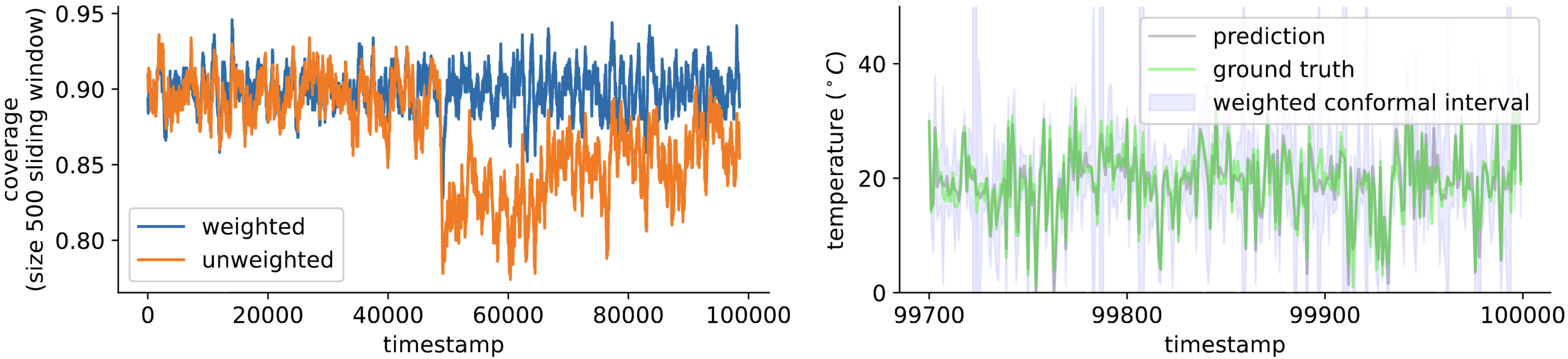


Red = false negatives, Blue = false positives, Black = true positives

Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Weather Prediction with Time-Series

We seek to predict the temperature of different locations on Earth given covariates such as the latitude, longitude, altitude, atmospheric pressure, and so on.



Exchangeability is violated!

Anastasios N. Angelopoulos and Stephen Bates. **A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification**. arXiv:2107.07511 (2022).

Recommended Reading

- Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511 (2022).
- Glenn Shafer, Vladimir Vovk. A Tutorial on Conformal Prediction. Journal of Machine Learning Research, 9(12):371−421, 2008.

Exercise I

You are given a trained probabilistic classifier over 4 classes $\{A, B, C, D\}$ and a calibration set of $n = 12$ examples. Your goal is to compute a split conformal prediction set $C(x_{test}) \subseteq \{A, B, C, D\}$ for a new input x_{test}, targeting 80% marginal coverage.

For the new input x_{test}, the classifier outputs:

- $\hat{p}(A \mid x_{test}) = 0.34$
- $\hat{p}(B \mid x_{test}) = 0.27$
- $\hat{p}(C \mid x_{test}) = 0.22$
- $\hat{p}(D \mid x_{test}) = 0.17$

i	True y_i	$\hat{p}(y_i \mid \boldsymbol{x}_i)$
1	A	0.88
2	B	0.62
3	C	0.51
4	D	0.44
5	A	0.73
6	B	0.39
7	C	0.81
8	D	0.57
9	A	0.29
10	B	0.76
11	C	0.35
12	D	0.68

Exercise II

You are given a trained probabilistic classifier over 4 classes $\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ and a calibration set of $n = 10$ examples. Your goal is to compute a split conformal prediction set $C(\mathrm{x}_{\mathrm{test}}) \subseteq \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ for a new input $\mathrm{x}_{\mathrm{test}}$, targeting 80% marginal coverage.

For the new input x_{test}, the classifier outputs:

- $\hat{p}(\mathrm{A} \mid x_{\mathrm{test}}) = 0.41$
- $\hat{p}(\mathrm{B} \mid x_{\mathrm{test}}) = 0.33$
- $\hat{p}(\mathrm{C} \mid x_{\mathrm{test}}) = 0.16$
- $\hat{p}(\mathrm{D} \mid x_{\mathrm{test}}) = 0.10$

i	True y_i	$\hat{\mathrm{p}}(\mathrm{y}_i \mid \boldsymbol{x}_i)$
1	A	0.91
2	B	0.78
3	C	0.66
4	A	0.59
5	D	0.84
6	B	0.72
7	C	0.63
8	D	0.55
9	A	0.47
10	B	0.40

Exercise III

You are given a trained probabilistic classifier over 4 classes $\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ and a calibration set of $n = 10$ examples. Your goal is to compute a split conformal prediction set $C(\mathrm{x}_{\text{test}}) \subseteq \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ for a new input x_{test}, targeting 80% marginal coverage.

For the new input x_{test}, the classifier outputs:

- $\hat{p}(\mathrm{A} \mid x_{\text{test}}) = 0.38$
- $\hat{p}(\mathrm{B} \mid x_{\text{test}}) = 0.26$
- $\hat{p}(\mathrm{C} \mid x_{\text{test}}) = 0.21$
- $\hat{p}(\mathrm{D} \mid x_{\text{test}}) = 0.15$

i	True y_i	$\hat{\mathrm{p}}(\mathrm{y}_i \vert \boldsymbol{x}_i)$
1	A	0.93
2	B	0.81
3	C	0.74
4	D	0.70
5	A	0.65
6	B	0.61
7	C	0.58
8	D	0.54
9	A	0.49
10	B	0.42