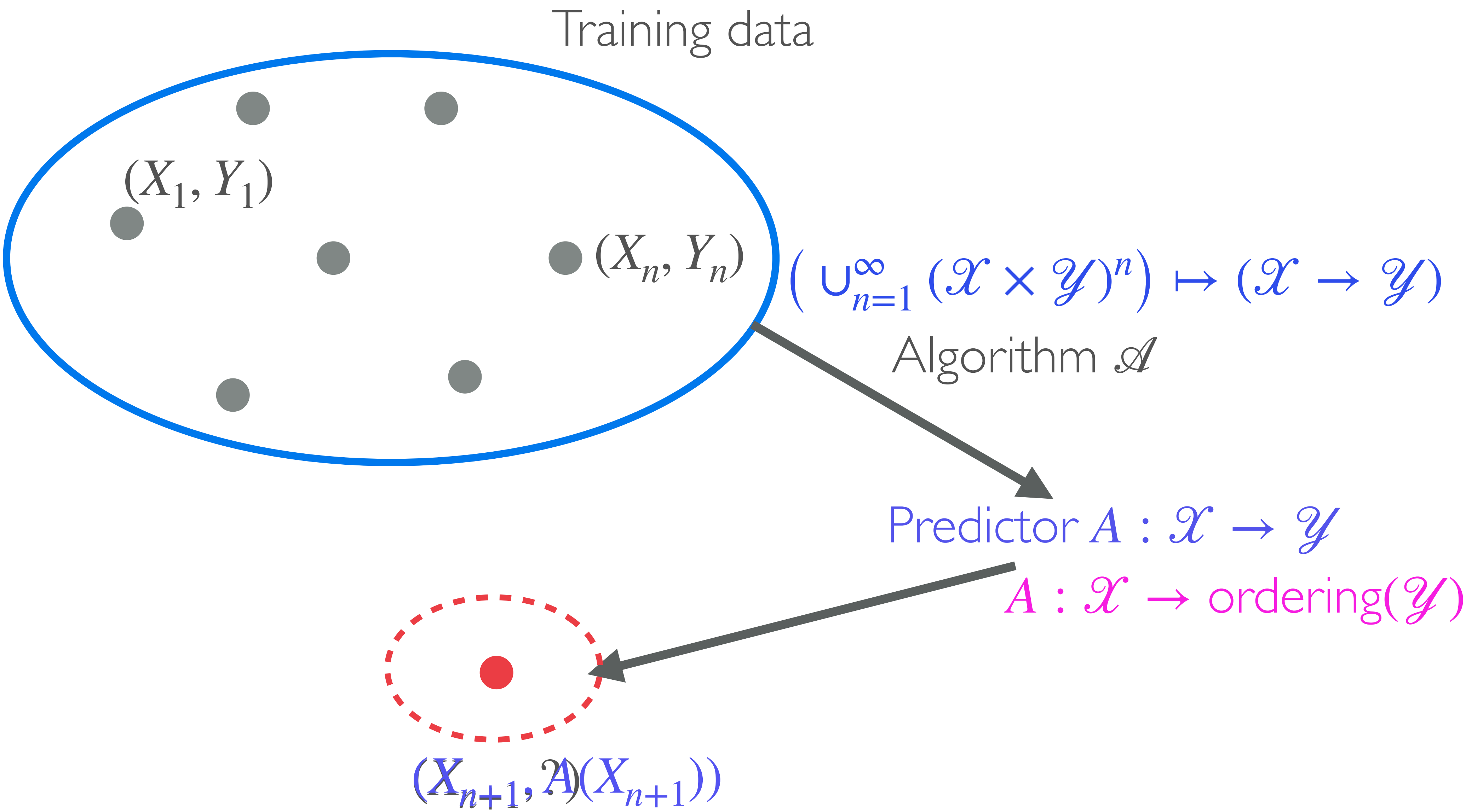
Uncertainty quantification for black-box classifiers without distributional assumptions

Aaditya Ramdas

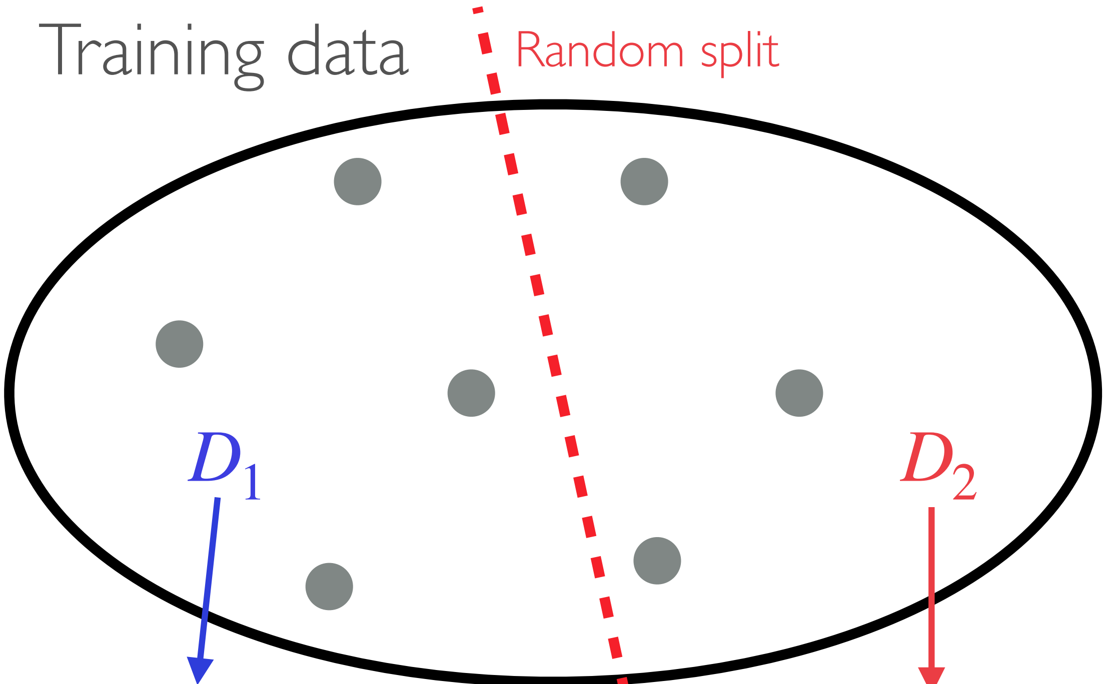
Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

Prediction vs “Predictive Inference”



$$\Pr(Y_{n+1} \in C_n(X_{n+1})) \geq 1 - \alpha \,.$$

Split Conformal Prediction for classification



$R_i :=$ rank(Y_i) in sort$(A(X_i))$

Let $\bar{R} = \{R_i\}_{i \in D_2}$

Classifier $A = \mathscr{A}(D_1)$

$A : \mathscr{X} \to \mathbb{R}^{|\mathscr{Y}|}$

$q_{1-\alpha}(\bar{R})$ $:= \lceil (1-\alpha)(n_2+1) \rceil$ ranked element of $\bar{R}$

$C(X_{n+1}) :=$ {sort$(A(X_{n+1}))$ and pick top $q_{1-\alpha}(\bar{R})$ elements}

$\Pr(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha\,.$

no assumptions on distribution P_{XY} or algorithm A

size(C) depends on distribution P_{XY} and algorithm A

Better residual for probabilistic classifiers

$A : \mathcal{X} \to \Delta^{|\mathcal{Y}|}$

$$R_i := \sum_{\ell \neq Y_i : [A(X_i)]_\ell > [A(X_i)]_{Y_i}} [A(X_i)]_\ell$$

= total prob. of more likely, wrong labels

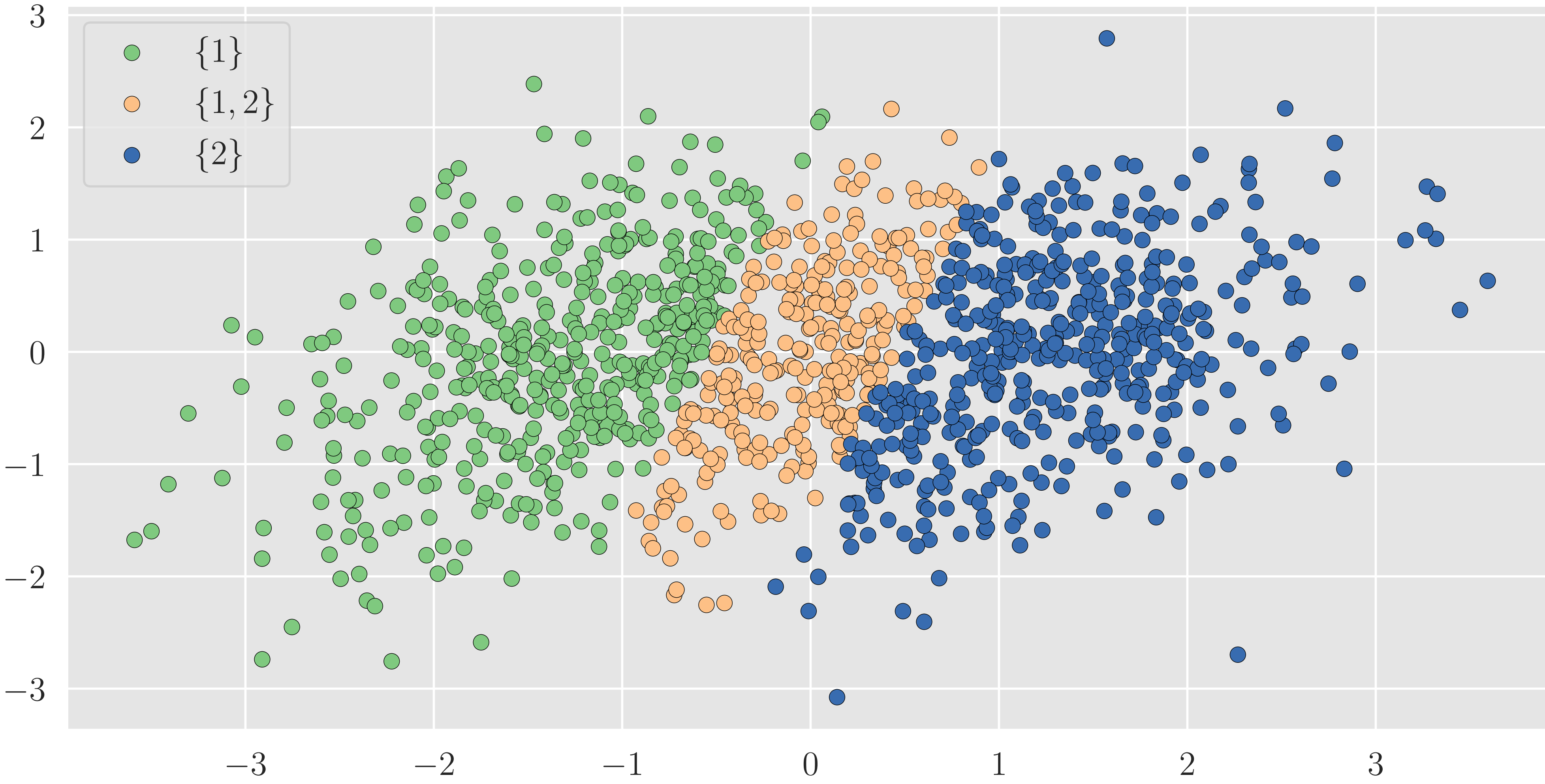
Let $\bar{R} = \{R_i\}_{i \in D_2}$

$C(X_{n+1}) :=$ {least number of labels whose total mass $\geq q_{1-\alpha}(\bar{R})$}

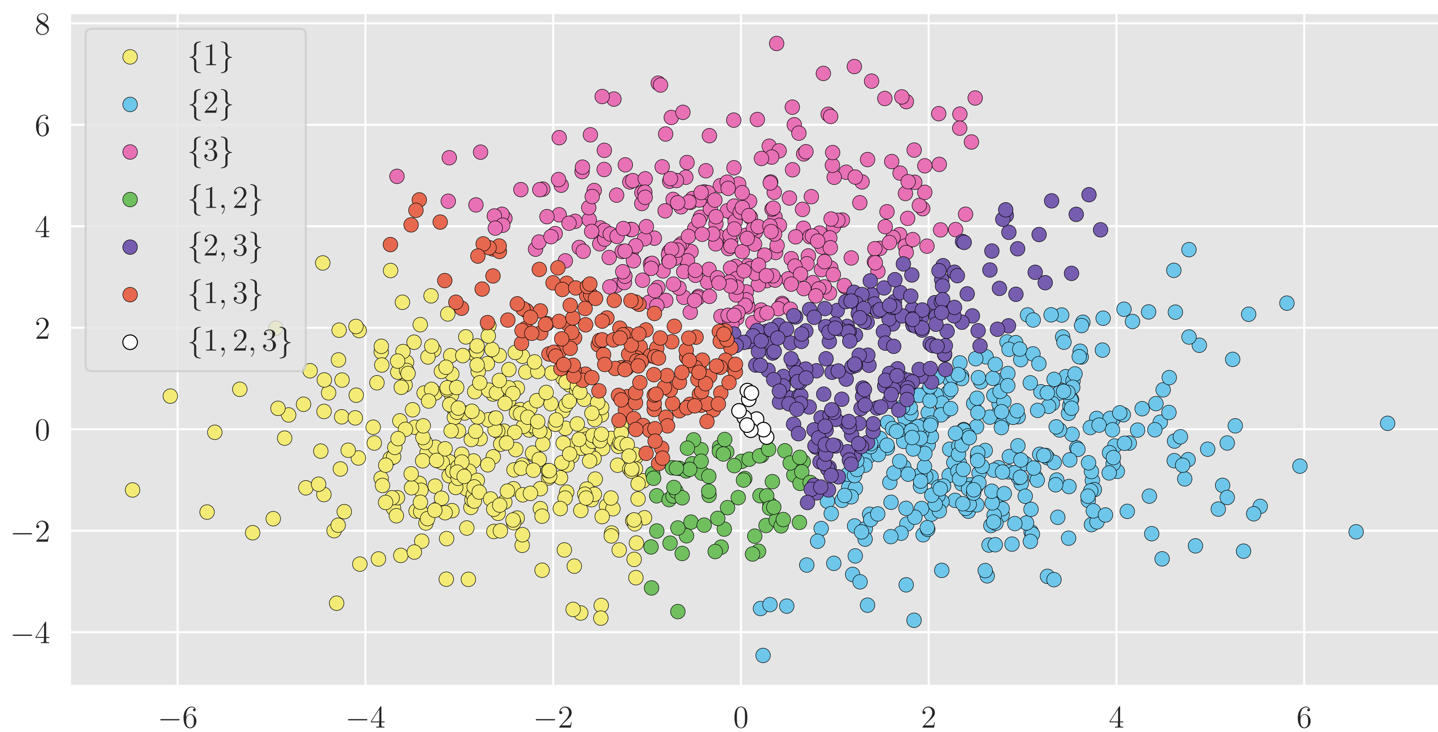
= {first k labels in sort$(A(X_{n+1}))$ with cumulative prob. $\geq q_{1-\alpha}(\bar{R})$}

(Can be "smoothed" out with randomization)

(Mixture of two Gaussians in two dimensions)



(Mixture of three Gaussians in two dimensions)



Part 2: Calibrated probabilities

Calibration in the binary setting

Gupta*, Podkopaev*, Ramdas'20

A function $f : \mathcal{X} \to [0,1]$ returns calibrated probabilities if

$$\mathbb{E}[Y_{n+1} \mid f(X_{n+1})] = f(X_{n+1})$$

Eg: Suppose we predict $f(X_{n+1}) \approx 0.3$ for 100 points,
then ≈ 30 of those will have label one, and the rest label zero.

Fact: if f is calibrated, then $f(X) = \mathbb{E}[Y \mid g(X)]$ for some g.
Reality: exact calibration is impossible with a finite data of size n.

We say that $f_n : \mathcal{X} \to [0,1]$ is distribution-free (ϵ_n, α)-calibrated if

$$\forall P_{XY}, \ \Pr(|\mathbb{E}[Y_{n+1} \mid f_n(X_{n+1})] - f_n(X_{n+1})| > \epsilon_n) \leq \alpha,$$

and asymptotically calibrated if $\epsilon_n \to 0$.

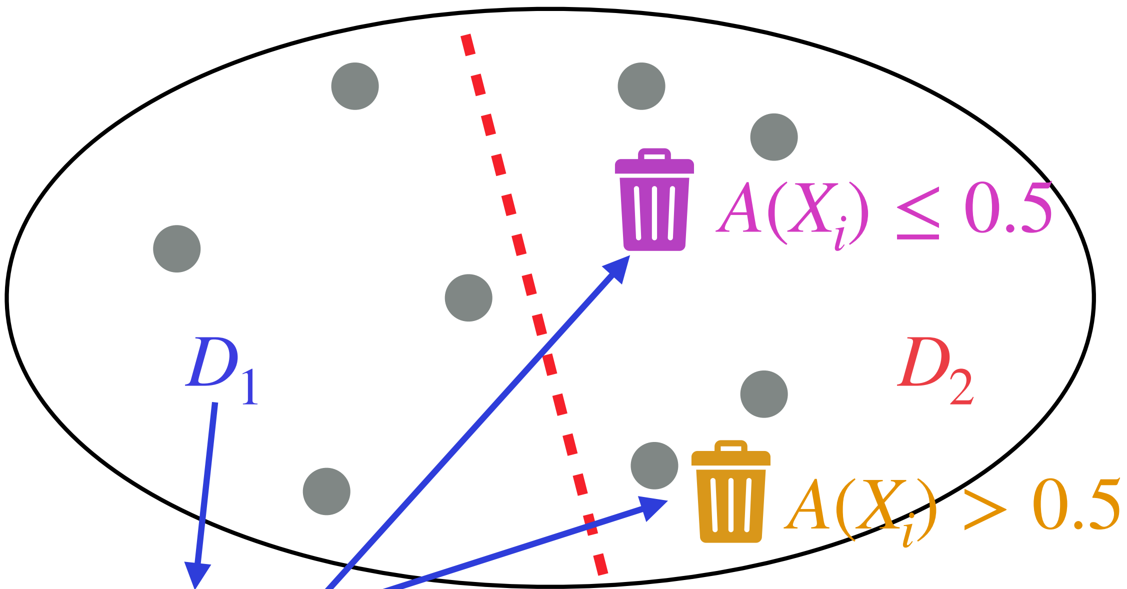
Theorem: Asymptotic distribution-free calibration is impossible
if $\lim_{n\to\infty} \text{Range}(f_n)$ is uncountable.

Split Binning

Gupta*, Podkopaev*, Ramdas'20

Training data

Random split



Improve to "uniform mass binning"

$\rightarrow$ smaller half of $A(X_i)$

$\rightarrow$ larger half of $A(X_i)$

$A = \mathscr{A}(D_1) : \mathscr{X} \rightarrow [0,1]$

$A(X_{n+1}) = $ 🗑 $\qquad f(X_{n+1}) = \frac{\sum_{i \in 🗑} Y_i}{|🗑|}$

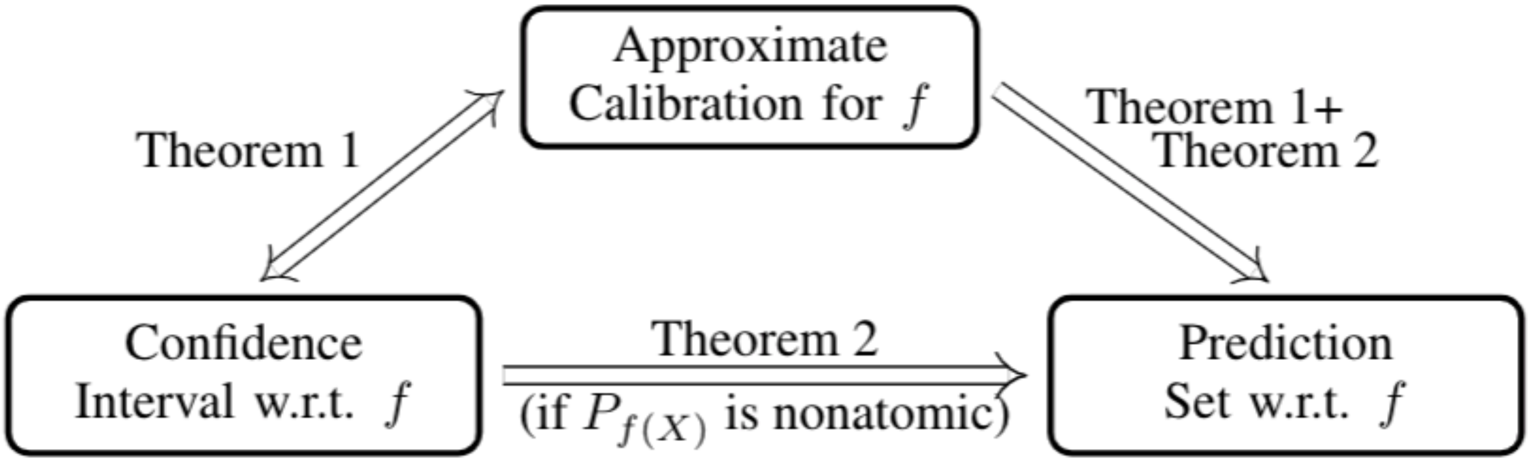
(generalize to any number of bins)

$$\Pr\left(|\mathbb{E}[Y_{n+1} \,|\, f(X_{n+1})] - f(X_{n+1})| \leq c\hat{\sigma}\sqrt{\frac{\ln(1/\alpha)}{n}} \right) \geq 1 - \alpha \,.$$

no assumptions on distribution P_{XY} or algorithm A

Distribution-free binary classification: prediction sets, confidence intervals and calibration

Chirag Gupta[*1], Aleksandr Podkopaev[*1,2], Aaditya Ramdas[1,2]



Theorem 3. *Let $\alpha \in (0, 0.5)$ be a fixed threshold. If a sequence of scoring functions $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level α for every distribution P then*

$$\limsup_{n\to\infty} |\mathcal{X}^{(f_n)}| \leqslant \aleph_0.$$

4.3 Distribution-free calibration in the online setting

4.4 Calibration under covariate shift

Sharpness?

One cannot guarantee sharpness without distributional assumptions.

$|\mathscr{Y}| = 2$

Number of bins, properties of P_{XY} and quality of original classifier, all together determine sharpness, but not calibration.

Eg: consider the setting where $P(Y = 1 \,|\, X) = 0.5$, i.e. $Y \perp X$.
No classifier can be sharp, and not all classifiers are calibrated.