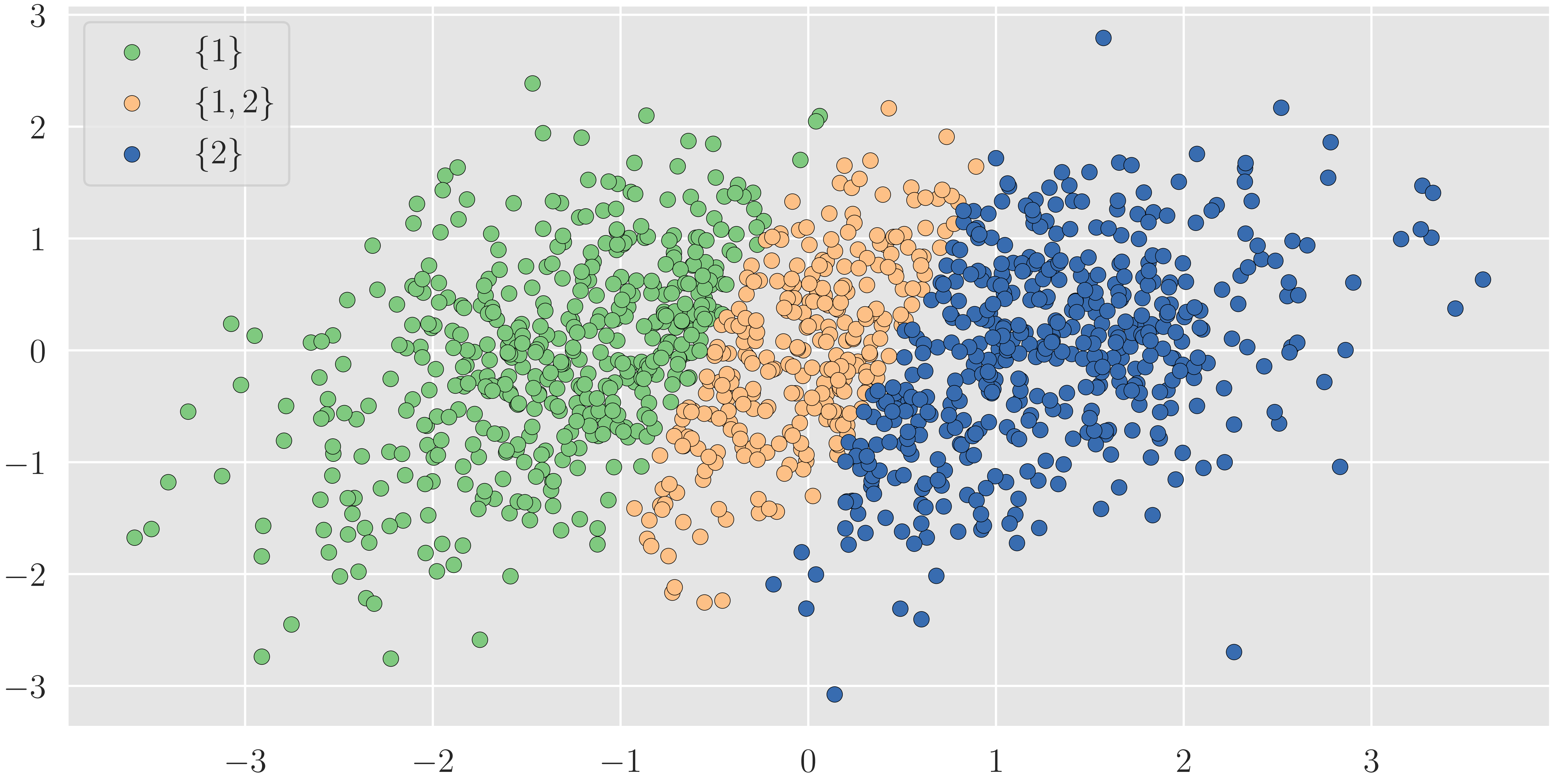
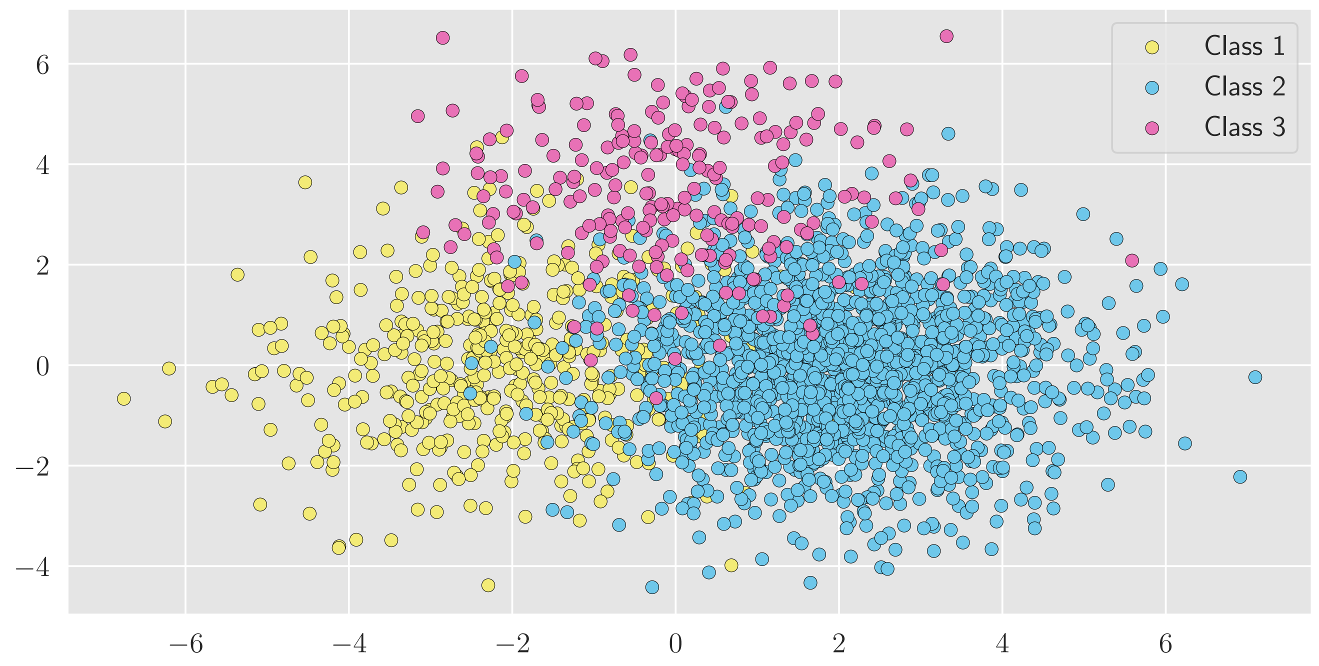
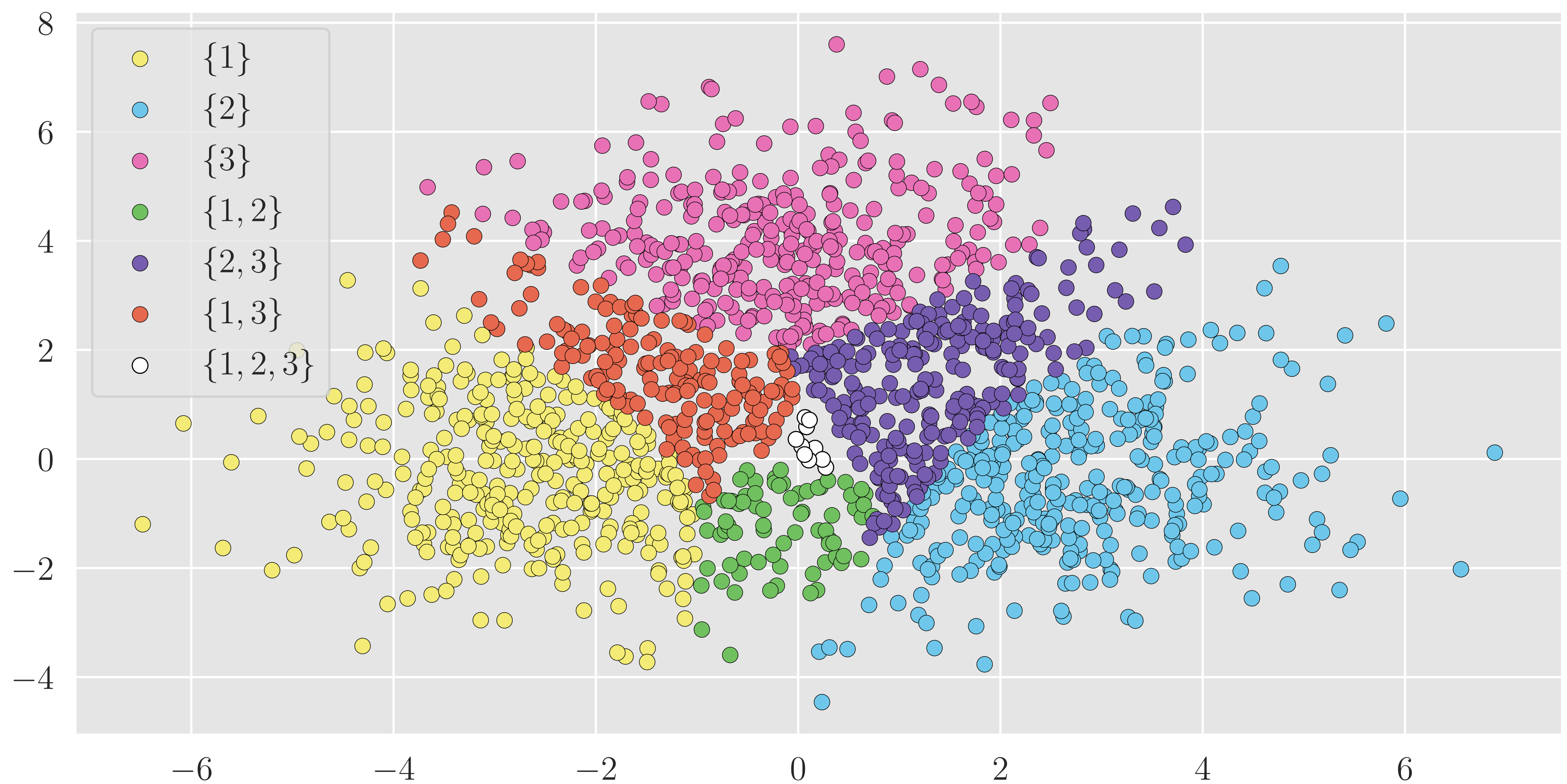
Post-hoc distribution-free calibration

(Mixture of two Gaussians in two dimensions)



(Mixture of three Gaussians in two dimensions)





Part 2: Calibrated probabilities

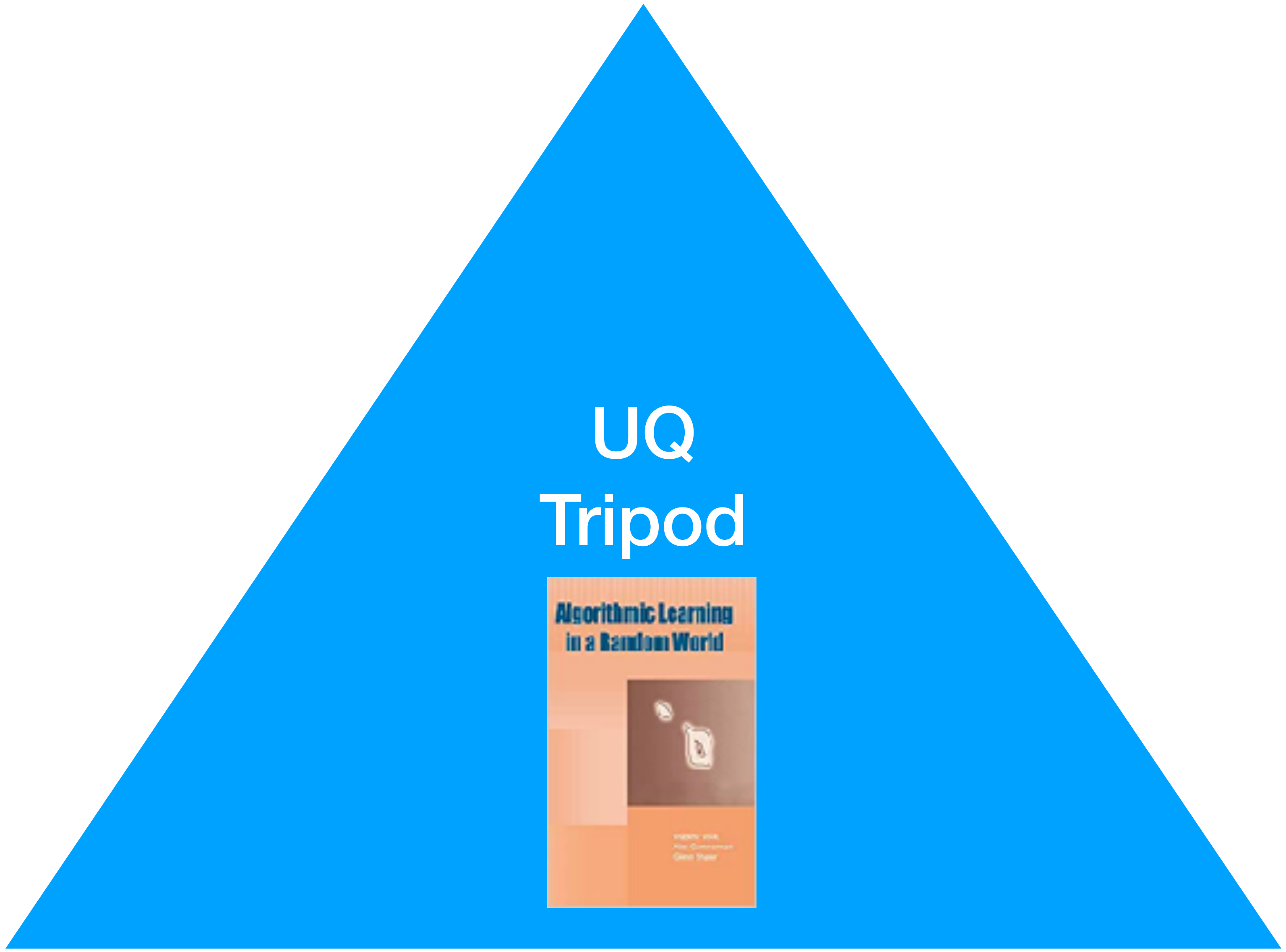
- **Top-label calibration**
  C. Gupta, A. Ramdas `ICLR, 2022` arxiv
- **Distribution-free calibration guarantees for histogram binning without sample splitting**
  C. Gupta, A. Ramdas `ICML, 2021` arxiv proc
- **Distribution-free uncertainty quantification for classification under label shift**
  A. Podkopaev, A. Ramdas `UAI, 2021` arxiv
- **Distribution-free binary classification: prediction sets, confidence intervals and calibration**
  C. Gupta, A. Podkopaev, A. Ramdas `NeurIPS, 2020` arxiv proc talk

First: binary, later: multiclass

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

Uninformative (binary)

Prediction sets



Confidence intervals

$$\Pr(\mathbb{E}[Y_{n+1} \,|\, X_{n+1}] \in C(X_{n+1})) \geq 1 - \alpha$$

Trivial (binary)

Calibration

$$\mathbb{E}[Y_{n+1} \,|\, f(X_{n+1})] \approx f(X_{n+1})$$

???

Barber'20, Vovk et al'05

Calibration in the binary setting

Gupta*, Podkopaev*, Ramdas'20

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

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

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.

Reality: exact calibration is impossible with a finite data of size n.
Typically, trained classifiers are not automatically calibrated.

"Post-hoc calibration" method $\mathscr{A} : (g, D_n) \mapsto f_n$

(Take in a classifier and some "calibration dataset",
output approximately calibrated classifier)

We say that $\mathscr{A}$ is distribution-free (ϵ_n, α)-calibrated if

$$\forall P \text{ over } \mathcal{X} \times \{0,1\},\ P(\,|\,\mathbb{E}[Y_{n+1} \,|\, f_n(X_{n+1})] - f_n(X_{n+1})\,| > \epsilon_n) \le \alpha,$$

("probably approximately calibrated")

Calibration in the binary setting

Gupta*, Podkopaev*, Ramdas'20

Theorem (informal):

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

(The output of the calibrated classifier has to be "discrete")

Asymptotic distribution-free calibration is impossible if $\mathscr{A}(\,\cdot\,, D_n)$ is an injective map.

(Many input classifiers must be mapped to the same output classifier)

Split Binning

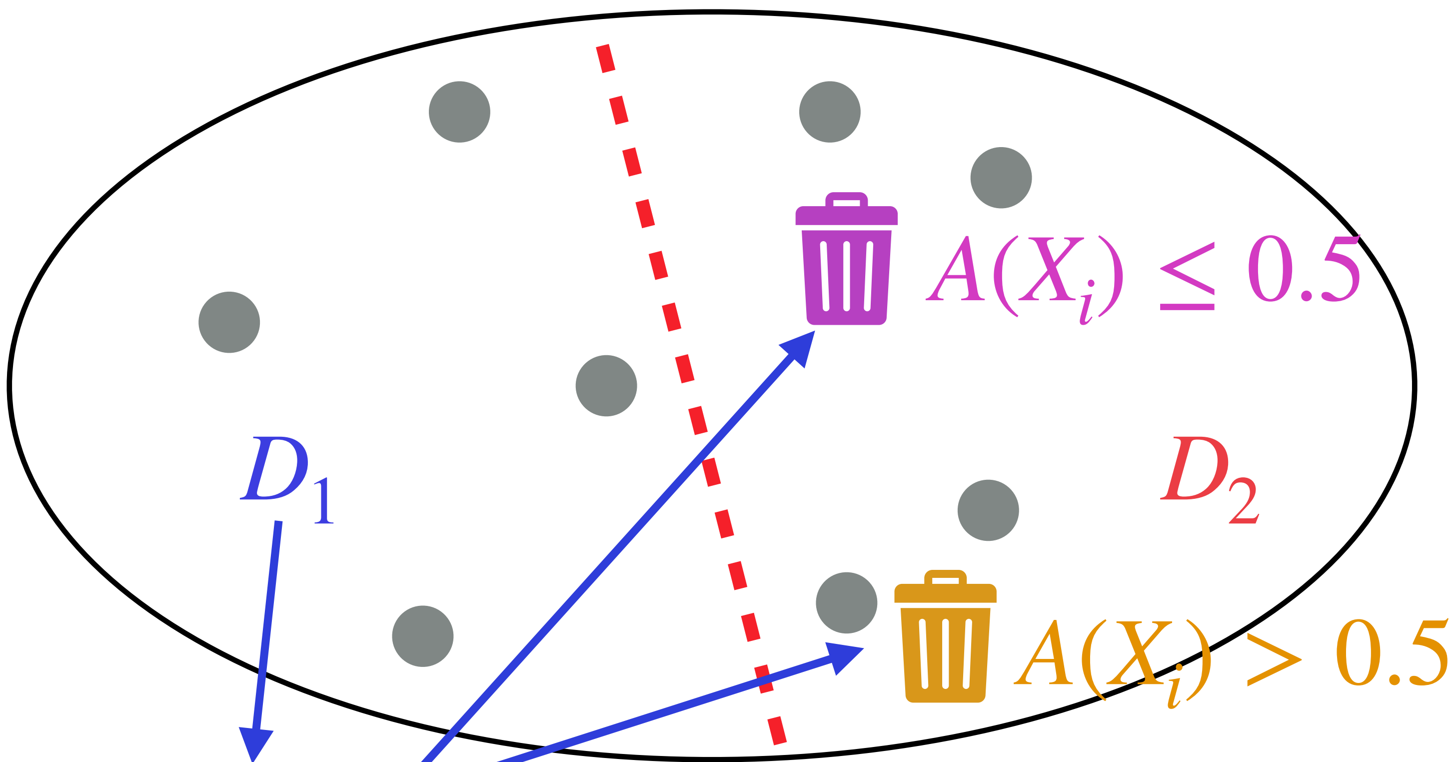
Zadrozny, Elkan'99

Gupta*, Podkopaev*, Ramdas'20

Gupta, Ramdas'21

Training data

Random split



Improve to "uniform mass binning"

→ smaller half of $A(X_i)$

→ larger half of $A(X_i)$

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

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

(generalize to any number of bins)

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

no assumptions on distribution P_{XY} or algorithm A

Sharpness?

One cannot guarantee sharpness without distributional assumptions.

$|\mathcal{Y}| = 2$

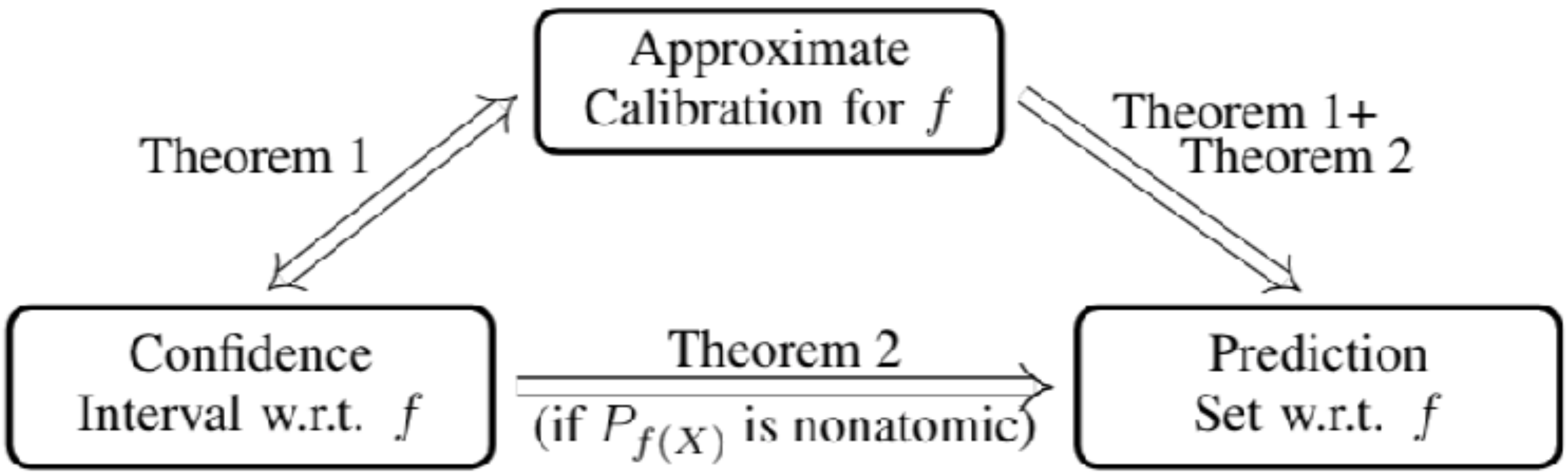
Number of bins, properties of P_{XY} and quality of original classifier, all together determine sharpness, but not calibration.

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.

(Halfway)

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

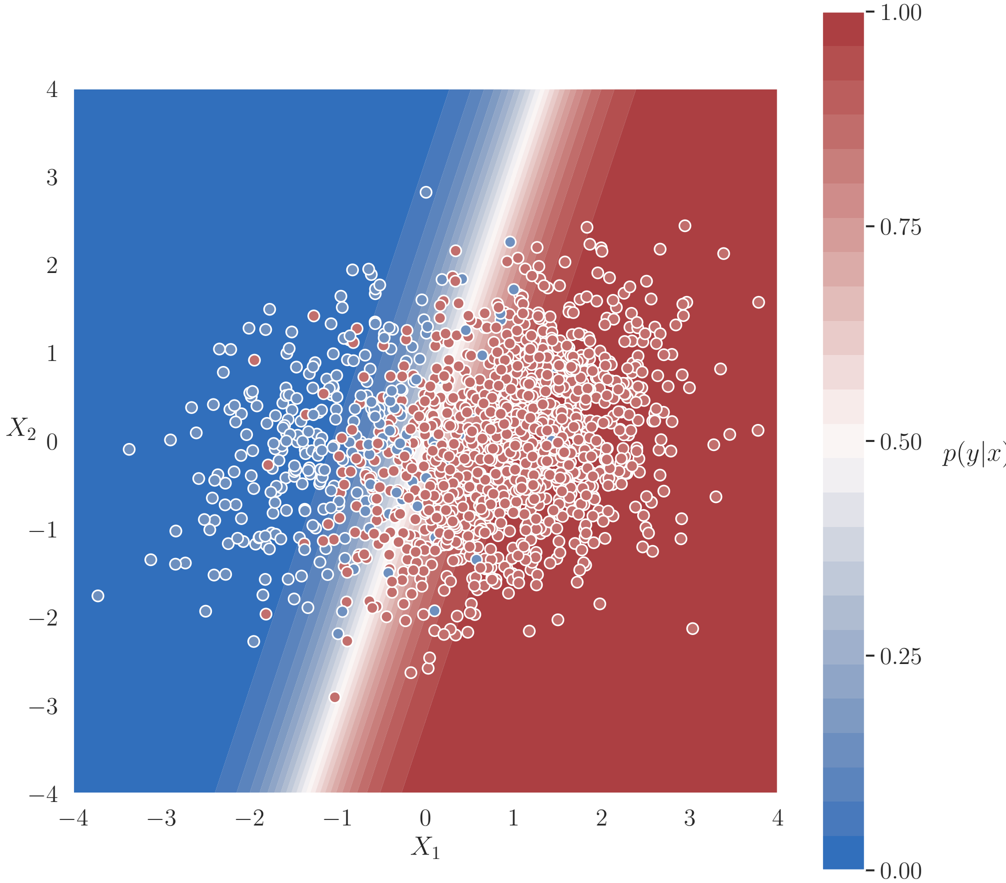
Chirag Gupta*[1], Aleksandr Podkopaev*[1,2], Aaditya Ramdas[1,2]

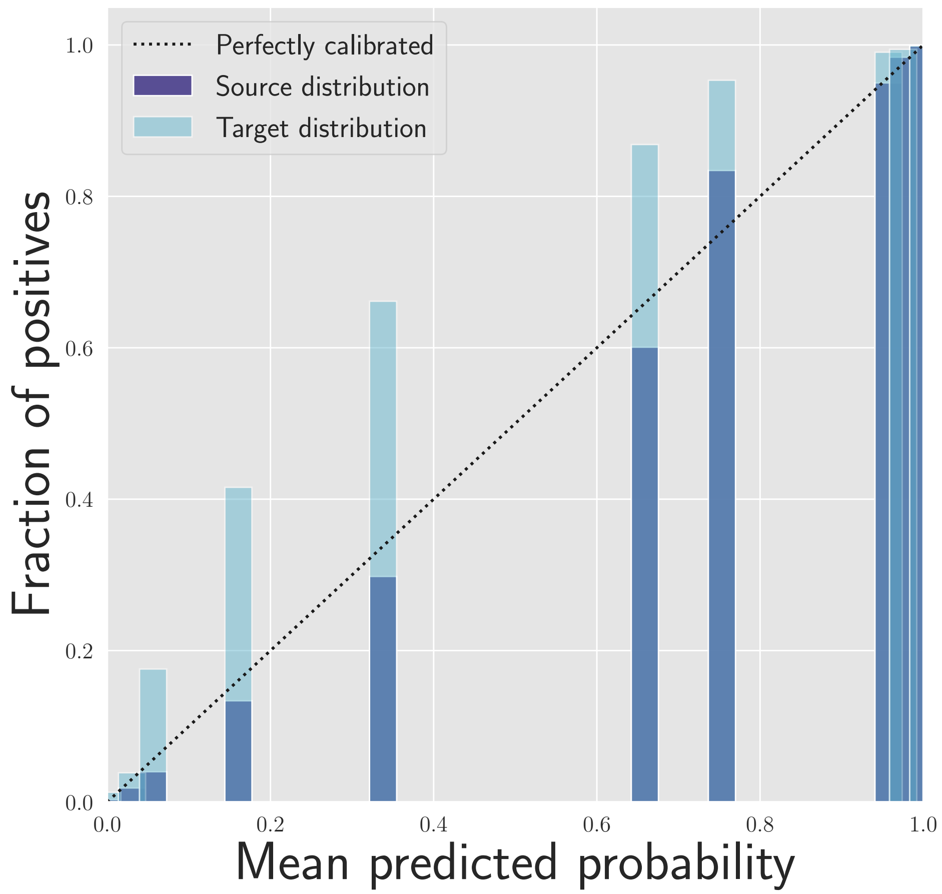


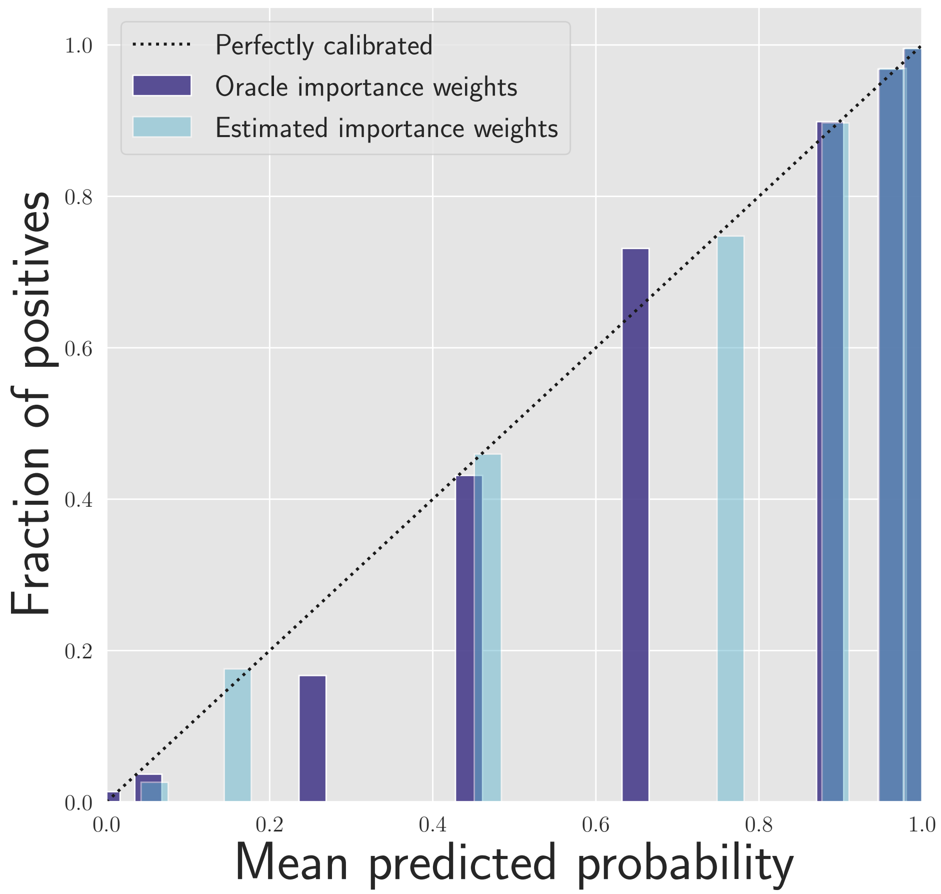
4.3 Distribution-free calibration in the online setting

Use "confidence sequences" (or anytime-valid confidence intervals) for estimating the bias (the fraction of ones) in each bin.

Covariate shift or Label shift





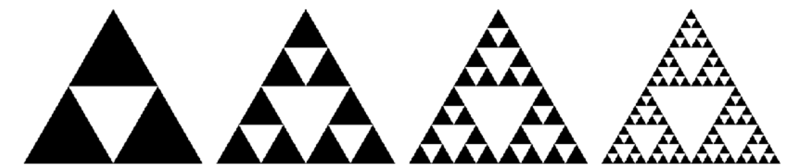


Binning for multiclass calibration

(Gupta, Ramdas arXiv'21)

The classifier A is assumed to output an uncalibrated probability vector.

Size of bin
determines sharpness



$|\mathscr{Y}| = 3$

"Sierpinski binning"

For held-out calibration data:

$A(X_i) = (0.9,\ 0.05,\ 0.05)$ gets mapped to top triangle bin,
$A(X_i) = (0.4,\ 0.3,\ 0.3)$ gets mapped to middle triangle bin.

For test data:

Use $A(X_{n+1})$ to identify the bin.
Report empirical distribution of labels in the bin.

Results in distribution-free multi-class calibration.

What if there are too many classes?

(Gupta, Ramdas arXiv'21)

The classifier A is assumed to output an uncalibrated score vector.

Number of bins grows exponentially large in $|\mathcal{Y}|$.

"Full-vector/joint calibration" or "marginal calibration" is not practical.
For any given X_{n+1}, we care little about nearly impossible labels.

We define a new notion called "top-label calibration".
Intuitively, if $A(X_{n+1}) = (0.1,\ 0.8,\ 0.05,\ 0.01,\ 0.01,\ 0.01, \cdots)$, then we care most about whether the 0.8 is not over/under-confident.

But the top label is not a fixed label, it is a data-dependent label, so one has to be careful with analysis.

Can achieve top-label calibration without distributional assumptions.

Top-label calibration

(Gupta, Ramdas arXiv'21)

"Post-hoc calibration" method $\mathscr{A} : (f, D) \mapsto (c, h)$

(Take in a classifier and a calibration dataset,
output top-label and its predicted probability)

We say that $\mathscr{A}$ is distribution-free (ϵ, α) top-label-calibrated if

$$\forall P, \; P[Y = \ell \mid c(X) = \ell, h(X) = r] \approx_{\epsilon,\alpha} r$$

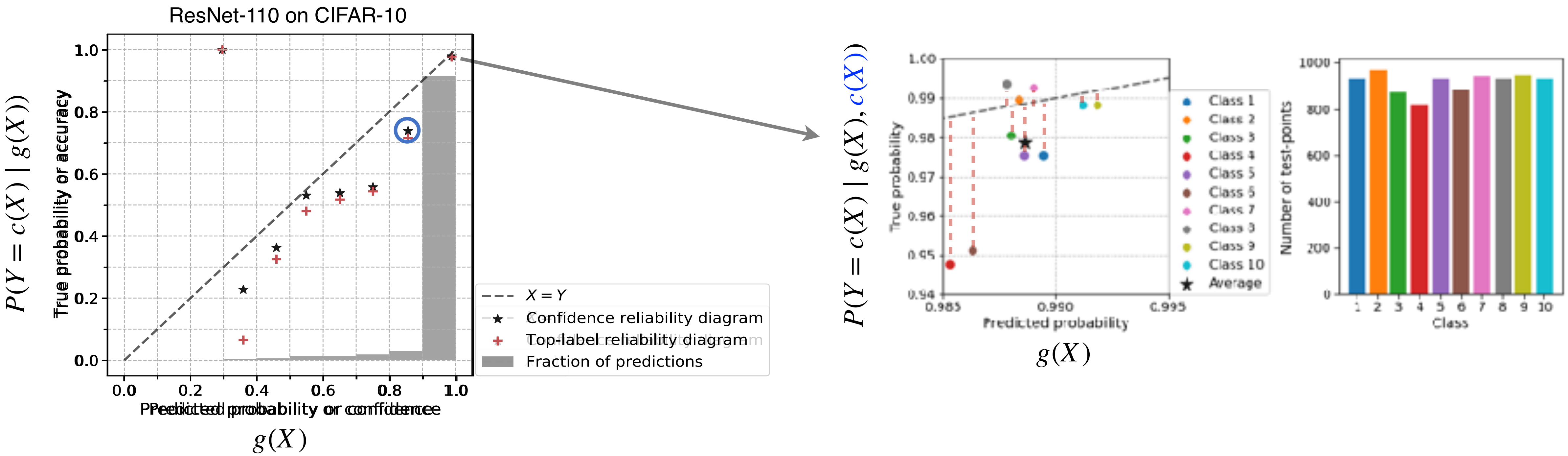
"Confidence calibration" (Guo et al.'17) only conditions on h(X).
Not intuitive (we always output c,h), and empirically fails to calibrate.

Class-wise calibration asks that $\forall \ell \in [L], P(Y = \ell \mid h_\ell(X)) \approx h_\ell(X)$

Full-vector/joint calibration asks that $P(Y = \ell \mid h(X)) \approx h_\ell(X)$

$$\text{TopLabel ECE} := \mathbb{E}\,|\,\text{LHS} - \text{RHS}\,|$$

(Extra slides)


ResNet-110 on CIFAR-10
$P(Y = c(X) \mid g(X))$
True probability or accuracy
Predicted probability or confidence
$g(X)$
$X = Y$
Confidence reliability diagram
Top-label reliability diagram
Fraction of predictions
$P(Y = c(X) \mid g(X), c(X))$
True probability
Predicted probability
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
Class 10
Average
Number of test-points
Class

Reducing top-label calibration to binary calibration

1. For each class l, create a separate dataset:

$$D_l = \{(X_i, \mathbf{1}\{Y_i = l\}) : c(X_i) = l\}$$

$$D = \{\underbrace{(X_1,2), (X_2,1)}_{c(\cdot)=1}, \ldots\}$$

$$D_1 = \{(X_1, 0)), (X_2, 1)\}$$

2. For each l, learn a different calibrated predictor $h_l : \mathcal{X} \to [0,1]$:

$$h_l = \text{binary-calibrator}(D_l, g)$$

3. Merge the h_l's to form a single $h : \mathcal{X} \to [0,1]$, given by

$$h(x) = h_{c(x)}(x)$$

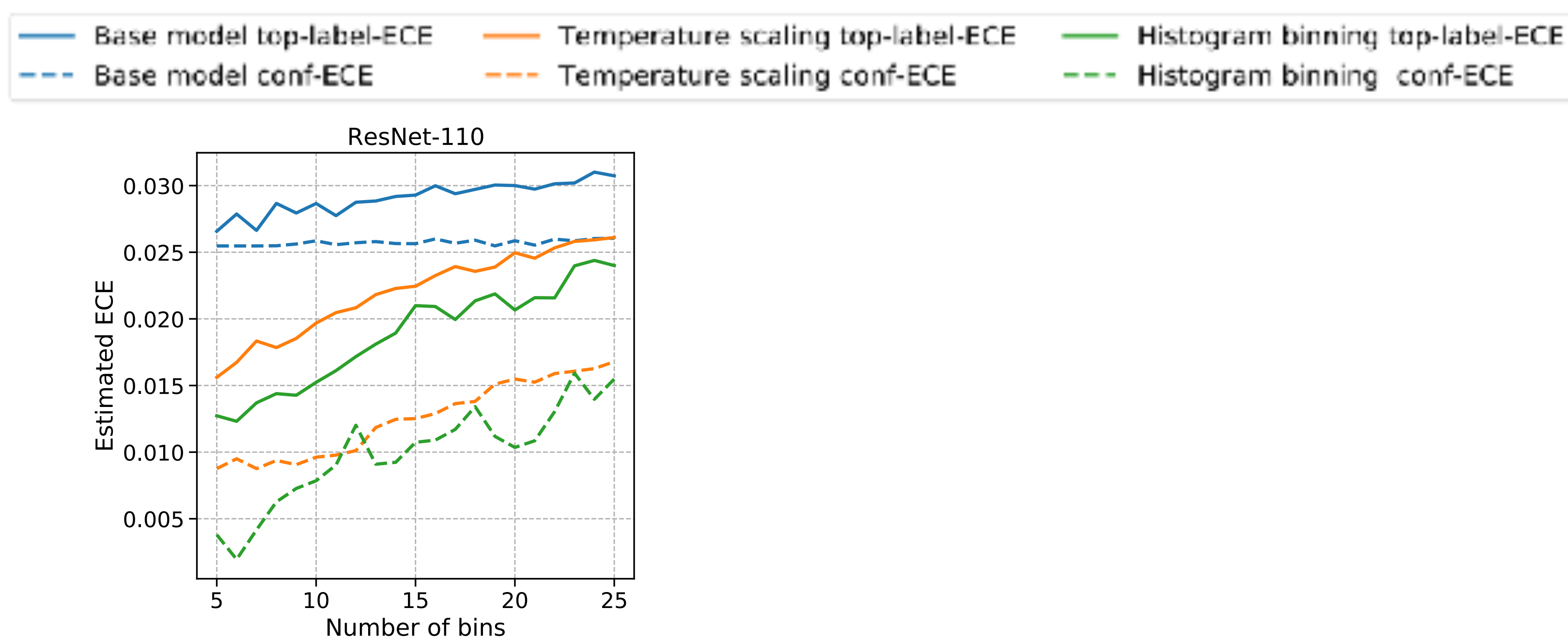
Theorem (informal): Fix number of points per bin k (say 50). For any distribution, $\text{TopLabel ECE} \leq \sqrt{1/2k}\ (= 0.1)$.

(Also high probability probably-approximately-calibrated bounds)

Top-label histogram binning performs better than temperature scaling on CIFAR-10

[Guo et al. 2017]

—— Solid lines: Top-label ECE



- The improvement in performance is higher for top-label maximum calibration error (MCE)
- We also propose a class-wise version of histogram binning that performs better than temperature scaling on CIFAR-10 and CIFAR-100