Verified Uncertainty Calibration

Ananya Kumar Percy Liang Tengyu Ma

Why Uncertainty Calibration Matters

- Important for testicular cancer (Calster & Vickers), bipolar disorder (Lindhiem et al), criminal recidivism (Fazel et al)

Tumor size
Teratoma
...
→ Random Forest → 20% chance of cancer

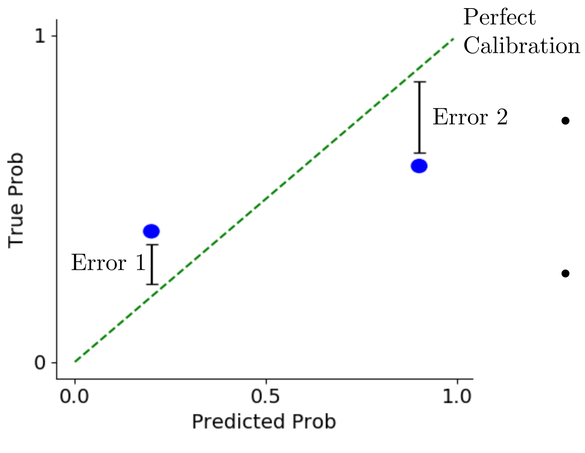
Reality: 40% such people have cancer (!)

Implication: Wrong Treatment

- Resnet on CIFAR-100 has poor uncertainties (Guo et al)

Model's perceived accuracy	90%
Actual accuracy	70%

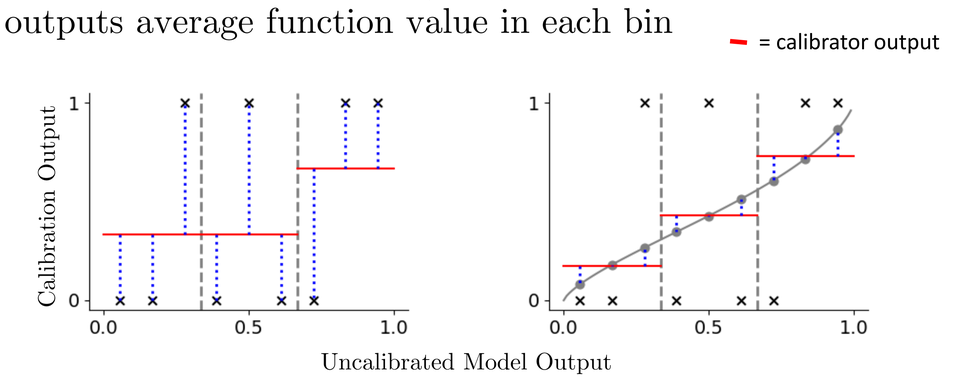
- Calibration error: average difference between models predicted probability and true probability



- $CE = \sqrt{E[(m-p)^2]}$
  - m is predicted prob
  - p is true prob i.e. $E[Y \mid m]$
- Used in meteorology, NLP, medicine, fairness, ML

Scaling-Binning Calibrator

- Histogram binning outputs average label value in each bin
- Scaling-binning calibrator (ours) fits a function to data, and outputs average function value in each bin



(a) Histogram (b) Scaling-binning (ours)

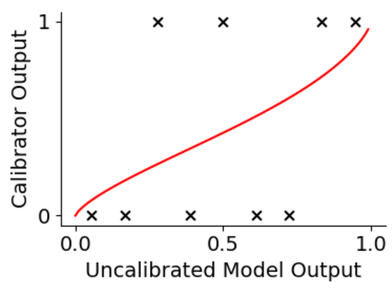
- Can calibrate with few samples + check that we are calibrated

Recalibration Method	*Samples Needed*	*Can Estimate Calibration?*
Platt Scaling	**Few**: $O\left(\frac{1}{\varepsilon^2}\right)$	✗
Histogram Binning	More: $O\left(\frac{B}{\varepsilon^2}\right)$	✓
Scaling-Binning (Ours)	**Few**: $O\left(\frac{1}{\varepsilon^2}+B\right)$	✓

- Validate these theoretically + experiments on CIFAR, ImageNet

Is Platt, Temperature Scaling Calibrated?

- Scale probabilities to make them better, scaling learned from labeled data



✕ = labels

- Widely used, but do they produce calibrated probabilities?

Deep Model → Platt Scaling → Model'

Is resulting Model' calibrated?

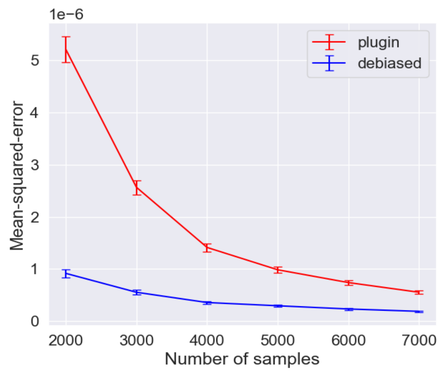
- Prior work would report calibration error = 2%
- We show that the calibration error is greater than 4%

Impossible to measure calibration error of scaling, can only underestimate

Measuring Calibration Error

- Given model, estimate calibration error from finite samples
- Plugin estimator: standard, everyone uses it (just average)
- Debiased: more sophisticated method from meteorology

Estimation Method	*Samples Needed*
Plugin	More: $O\left(\frac{B}{\varepsilon^2}\right)$
Debiased	**Fewer**: $\boldsymbol{O}\left(\frac{\sqrt{B}}{\varepsilon^2}\right)$



- Requires fewer samples to measure calibration error on CIFAR-10

Practical Takeaways

Train Model → Recalibrate probabilities → Measure calibration error

Use scaling-binning calibrator

Use debiased calibration estimator

- For scaling methods: can only lower bound calibration error
- Still use debiased estimator, estimates lower bound with fewer samples

- Measure accuracy **and calibration** with our library
- For multiclass measure calibration per-class

```
pip install calibration
```

```
import calibration as cal
ce = cal.get_calibration_error(logits, labels)
```