Reliable Decisions with Threshold Calibration
Roshni Sahoo¹, Shengjia Zhao¹, Alyssa Chen², Stefano Ermon¹
¹rsahoo, sjzhao, ermon @ cs.stanford.edu, ²alyssa.chen@utsouthwestern.edu

Example: Hospital Scheduling Decisions

Forecaster predicts patient length-of-stay in the hospital.
Hospital decides whether they have capacity to admit new patients based on threshold decision on model's predictions of current patients' length of stay.
Incorrect decisions from based on the model's predictions cause the hospitals to accrue costs.

Question: What notion of calibration is necessary and sufficient to guarantee that a forecaster (ML model) enables decision makers to predict their decision loss prior to deployment under threshold decision rules?

Threshold Calibration: Necessary and Sufficient Condition for Zero Reliability Gap.

Experiments: Threshold calibration outperforms other calibration methods in minimizing the reliability gap across different threshold loss functions without compromising on the accuracy of the decisions.

We provide a recalibration algorithm that takes an uncalibrated forecaster as input and provably outputs a threshold calibrated forecaster.

Procedure converges after at most $O\left(\frac{1}{\epsilon^4}\right)$ iterations and outputs a forecaster with threshold calibration error at most $\epsilon$.

Inspired by previous work on multicalibration.

Experiments: Threshold calibration outperforms other calibration methods in minimizing the reliability gap across different threshold loss functions without compromising on the accuracy of the decisions.