3.6 Faster rates for calibration decision loss

Let's briefly recall the definitions we need. For each prediction p, let $S_p = \{t : p_t = p\}$ and let

$$\bar{y}_p := \frac{1}{|S_p|} \sum_{t \in S_p} y_t$$

be the empirical frequency of the outcome among the rounds in which we predicted p. Recalibrating the predictions amounts to replacing every prediction p by $\bar{y}_p$, and the ECE from Class 1 is

$$\mathrm{ECE}(\mathbf{p}; \mathbf{y}) = \frac{1}{T} \sum_{p} |S_p| \big| p - \bar{y}_p \big|.$$

Recall also that a decision task $G = (A, U)$ has a best-response function BR_G. For a fixed task, the decision loss from miscalibration is the improvement in average utility obtained by recalibrating before choosing an action:

$$\frac{1}{T} \sum_{t=1}^{T} \Big[U\big(\mathrm{BR}_G(\bar{y}_{p_t}), y_t\big) - U\big(\mathrm{BR}_G(p_t), y_t\big) \Big].$$

Taking the supremum over all decision tasks with utilities in $[0, 1]$ gives the calibration decision loss defined above.

The main result of Hu and Wu shows that we can obtain a substantially faster rate for this actionable notion of calibration than the classical rate for ECE.

Theorem 3.14 (Informal; [HW24]). *There is an efficient randomized sequential forecasting strategy such that, against every adversarial sequence of binary outcomes,*

$$\mathbb{E}[\mathrm{CDL}(\mathbf{p}; \mathbf{y})] \le O\left(\frac{\log T}{\sqrt{T}}\right).$$

The expectation is over the randomness of the forecasting strategy.

Thus CDL can decrease essentially as $1/\sqrt{T}$, up to a logarithmic factor. By comparison, the classical guarantee from Class 1 for ECE is $O(T^{-1/3})$. More precisely, there is an $\Omega(T^{-0.472})$ lower bound for expected ECE, so expected ECE does not even admit an $O(T^{-0.499})$ guarantee in the usual adversarial online setting [QV21]. The faster CDL rate shows that it is easier to produce predictions that are trustworthy for every downstream decision maker than it is to make every prediction bucket have small calibration bias.

The same algorithmic template with finer buckets. At a high level, the Hu–Wu algorithm looks much like an algorithm for controlling ECE: discretize the prediction interval into buckets and control the bias accumulated in each bucket. For ECE, we add the absolute errors from every bucket. Balancing the discretization error, which is about $1/m$ for m buckets, against the total bucket error, which is about $\sqrt{m/T}$, leads to $m \approx T^{1/3}$ and the familiar $T^{-1/3}$ rate. Hu and Wu instead use roughly $\sqrt{T}/\log T$ buckets. The important difference is mostly in the analysis: CDL does not need to pay for all of the bucket errors simultaneously, only those that can change the decision for the task under consideration. This intuition is perhaps cleanest in the offline setting, where a bucket containing

n_p observations has an unnormalized sampling error of order $\sqrt{n_p}$. ECE sums this error over every bucket, whereas the decision loss for a threshold rule only uses buckets whose correction crosses that threshold.

Why CDL can be smaller than ECE. The difference is that ECE charges us for the calibration error in every prediction bucket, whether or not that error changes a decision. CDL only charges us when miscalibration causes a decision maker to choose the wrong action. For example, suppose the decision is whether to take an umbrella, and the decision maker takes one exactly when the predicted probability of rain is at least $1/2$. Consider a predictor that outputs 0.4 on half the days and 0.6 on the other half, while the empirical frequencies of rain in those two buckets are 0.2 and 0.8, respectively. Its ECE is

$$\frac{1}{2}|0.4 - 0.2| + \frac{1}{2}|0.6 - 0.8| = 0.2.$$

Nevertheless, recalibration does not change the umbrella decision: both 0.4 and 0.2 say not to take an umbrella, while both 0.6 and 0.8 say to take one. The decision loss for this task is therefore zero. More generally, for any threshold decision rule, only buckets whose prediction and empirical frequency fall on opposite sides of the threshold contribute to decision loss. Hu and Wu's analysis turns this observation into a simultaneous guarantee for every bounded decision task: instead of summing the calibration errors from all buckets as ECE does, it controls the smaller collection of errors that can actually change a best response.

Acknowledgments and AI Use

This class was a guest lecture by Lunjia Hu, who prepared these notes. AI usage is unknown.

4 Class 4: Group Conditional Bias and Calibration

So far we have seen that calibrated predictions have interesting *decision theoretic* guarantees: we will always maximize our utility by treating the predicted probability $f(x) = p$ as if the probability of y given x is really p. Let's see what we can and cannot conclude from that.

Example 4.1. Consider a weather prediction task with outcome $y \in \{0, 1\}$, where $y = 1$ means that it rains. Each day has two features $x = (x_1, x_2)$: the humidity $x_1 \in \{\text{humid}, \text{dry}\}$ and the cloud cover $x_2 \in \{\text{cloudy}, \text{sunny}\}$. Suppose that the four possible feature vectors are equally likely and that

	cloudy	sunny
humid	3/4	1/2
dry	1/2	1/4

gives the probability of rain for each feature vector.
Consider a predictor that only uses the humidity:

$$f(x_1) = \begin{cases} 5/8 & \text{if } x_1 = \text{humid}, \\ 3/8 & \text{if } x_1 = \text{dry}. \end{cases}$$

This predictor is calibrated. Indeed, conditional on the prediction $f(x) = 5/8$, cloudy and sunny

days are equally likely, so the probability of rain is

$$\frac{1}{2}\left(\frac{3}{4}+\frac{1}{2}\right)=\frac{5}{8}.$$

Similarly, conditional on the prediction $f(x)=3/8$, the probability of rain is

$$\frac{1}{2}\left(\frac{1}{2}+\frac{1}{4}\right)=\frac{3}{8}.$$

Thus f is calibrated even though it ignores the cloud-cover feature.

Our theory of calibrated decision loss says that if we only base our decisions on the predictions, then we should treat these predictions as if they are ground truth probabilities. But should we only base our decisions on the predictions? If it's cloudy or sunny, and incorporate that information into our decsion making. If we do that, then on a day that is humid and cloudy, the prediction is $p=5/8$ but we know that the true probability is actually $3/4$, and can do better than simply trusting the prediction. Thus the predictions can be biased after we condition on some of the observable features that factor into our decision.

4.1 Group conditional bias

To this end, we can revisit the definition of average bias and add in the notion of groups that capture the information we are using to make a decision.

As before, let's start with the concept of average bias. Recall that the average bias of a predictor is the expected difference between its prediction and the outcome.

Definition 4.2 (Average bias)**.** For a distribution $\mathcal{D}$ and a predictor f, the *average bias* is

$$\beta_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{\mathcal{D}}[f(X)-Y].$$

The predictor is *unbiased on average* if $\beta_{\mathcal{D}}(f)=0$.

As we saw in Class 1, we can remove the average bias without increasing the mean squared error. In fact, the improvement in MSE is exactly the squared bias.

Lemma 4.3 (Unbiasing reduces MSE)**.** *Let $f_{\beta\to 0}$ be the unbiasing of f, defined by*

$$f_{\beta\to 0}(x) := f(x)-\beta_{\mathcal{D}}(f).$$

Then $f_{\beta\to 0}$ is unbiased on average and

$$\mathrm{MSE}_{\mathcal{D}}(f_{\beta\to 0}) = \mathrm{MSE}_{\mathcal{D}}(f)-\beta_{\mathcal{D}}(f)^2.$$

Proof. First,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\beta\to 0}(X) - Y] = \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y] - \beta_{\mathcal{D}}(f) = 0,$$

so $f_{\beta\to 0}$ is unbiased on average. Moreover,

$$\begin{aligned}\mathrm{MSE}_{\mathcal{D}}(f_{\beta\to 0}) &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[(f(X) - Y - \beta_{\mathcal{D}}(f))^2\right] \\ &= \mathrm{MSE}_{\mathcal{D}}(f) - 2\beta_{\mathcal{D}}(f)\mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y] + \beta_{\mathcal{D}}(f)^2 \\ &= \mathrm{MSE}_{\mathcal{D}}(f) - \beta_{\mathcal{D}}(f)^2.\end{aligned}$$

□

Now, we want to capture the idea that the bias is small *conditioned on some property of the feature vector*. We will eventually combine this with calibration, where we condition on the prediction itself as well, but for now what we do will be orthogonal to calibration.

We can define a *group* $G \subseteq \mathcal{X}$ to represent feature vectors with some property. The group is represented by a function $g\colon \mathcal{X} \to \{0, 1\}$, where $g(x) = 1$ means that x is in the group and $g(x) = 0$ means x is not in the group. For example, if G is the group of cloudy days then

$$G = \{(x_1, x_2) : x_2 = \text{cloudy}\}$$

and

$$g(x_1, x_2) = \begin{cases} 1 & \text{if } x_2 = \text{cloudy}, \\ 0 & \text{if } x_2 = \text{sunny}. \end{cases}$$

We can ask whether a predictor is biased when we restrict attention to a particular group. Let $\mathcal{G}$ be a collection of group indicators $g\colon \mathcal{X} \to \{0, 1\}$, and let all expectations below be over $(X, Y) \sim \mathcal{D}$. Define the probability mass of a group by

$$\mu_{\mathcal{D}}(g) := \mathop{\mathbb{P}}_{\mathcal{D}}[g(X) = 1].$$

Group conditional bias is the average bias from Class 1, computed under the distribution restricted to the group. This definition is especailly interesting when we consider a large family $\mathcal{G} \subseteq 2^{\mathcal{X}}$ of groups simultaneously.

Definition 4.4 (Group conditional bias)**.** For a predictor $f\colon \mathcal{X} \to \mathbb{R}$ and a group g with $\mu_{\mathcal{D}}(g) > 0$, the *group conditional bias* is

$$\beta_{\mathcal{D}}(f, g) := \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y \mid g(X) = 1].$$

The predictor is *unbiased conditional on* g if $\beta_{\mathcal{D}}(f, g) = 0$. Given a collection of groups $\mathcal{G}$ and a parameter $\alpha \geq 0$, we require the following group conditional bias bound for every $g \in \mathcal{G}$ with $\mu_{\mathcal{D}}(g) > 0$:

$$\beta_{\mathcal{D}}(f, g)^2 \leq \frac{\alpha^2}{\mu_{\mathcal{D}}(g)}.$$

When $\alpha = 0$, this requires zero group conditional bias on every group of positive mass.