CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 21: Calibration and Downstream Agent Utility

November 13, 2025

Lecturer: Nika Haghtalab *Readings: Kleinberg et al. [2023]*
Scribe: Lawrence Qian

1 From Calibration to Utility-Based Decision Making

Up to now we have mostly talked about calibration as a purely statistical notion. In the previous lecture, we started to see that calibration has impact on the downstream agent: Agents who act as if calibrated predictors at the true underlying distribution of events have no swap regret. Today, we dig deeper at the relationship between calibration and downstream agent utility. Roughly, the plan for today is:

- We set up a very simple decision problem (rain / no rain, boots / no boots).
- We define utilities and best responses to predictions.
- We introduce an agent regret notion.
- We relate calibration error, proper scoring rules, and agent regret.
- We see that calibration is sufficient for good downstream utility but not necessary.
- This motivates U-calibration as a notion directly about controlling agent utility.
- We start to discuss how we can bound U-calibration with rates significantly better than full calibration.

1.1 Utility, Actions, and Best Responses

Let us fix the basic setup. We have a binary outcome

$$a_t \in \{0, 1\}$$

for each time t; you can think of $a_t = 1$ as "it rains" and $a_t = 0$ as "it does not rain." The agent has a finite set of actions $\mathcal{S}$. In the boot example, $\mathcal{S}$ could just be $\{\text{boots}, \text{no-boots}\}$, but in general we do not restrict it. The agent gets utlity from taking actions, such as wearing boots, depending on what the outcome (the weather) actually is.

Definition 1.1 (Utility)**.** An agent utility function is a map

$$u : \mathcal{S} \times \{0, 1\} \to [-1, 1].$$

Throughout this lecture, we work with utility functions whose range is bounded in $[-1, 1]$.

So for example, if we wear boots on a day that ends up being sunny, we might decide that $u(\text{boots}, 0) = -\frac{1}{2}$ because we wore heavy boots for nothing. If we did not wear boots and it rained, that might also incur negative utility, etc.

Of course, we will also care about *expected* utility. If we believe that rain happens with probability p, then for every action s we can look at

$$u(s; p) \ := \ \mathop{\mathbb{E}}_{a \sim \text{Bern}(p)}[u(s, a)].$$

Now let us talk about the agent. Throughout this lecture we consider agents that *fully trust* the predictions they are given. That is, if the forecaster tells them that the probability of rain today is p_t, they act as if that is the true probability of rain.

Formally, if the forecast at time t is $p_t \in [0, 1]$, then the action s_t taken by the agent whose utility is u, is

$$s_t \ = \ \text{BR}_u(p_t),$$

where $\text{BR}_u(p_t)$ is a *best response* to the prediction. Concretely,

$$\text{BR}_u(p_t) \in \arg\max_{s \in \mathcal{S}} \mathop{\mathbb{E}}_{a \sim \text{Bern}(p_t)}[u(s, a)].$$

When the utility is clear from context, we simply refer to this as $\text{BR}(\cdot)$ as a function $[0, 1] \to \mathcal{S}$, and we will use this notation quite a bit: $\text{BR}(p)$, $\text{BR}(q)$, etc. to mean "the best response to prediction p" and "the best response to prediction q".

So these are the basic objects:

- a sequence of predictions $p_1, \ldots, p_T$,
- a sequence of outcomes $a_1, \ldots, a_t$,
- and an agent that plays $\text{BR}(p_t)$ at each time step.

We next recall calibration and then introduce a notion of regret that captures what the agent could have done instead.

2 Calibration Error

Let us remind ourselves how we defined calibration. Given $p_{1:T}$ and $a_{1:T}$, for each value p that we ever predict we define:

- $N(p)$ = number of time steps t with $p_t = p$,
- $\bar{a}(p)$ = empirical probability of rain of ones among those times:

$$\bar{a}(p) = \frac{1}{N(p)} \sum_{t:p_t=p} a_t.$$

We would like forecasts to be such that whenever we predict p, then among all the days on which we predicted p it really rains about p-fraction of the time. That is exactly what calibration is asking for.

Definition 2.1 (Calibration error)**.** The calibration error of $(p_{1:T}, a_{1:T})$ is

$$\mathrm{CalErr}(p_{1:T}, a_{1:T}) = \sum_{p\in[0,1]} \frac{N(p)}{T} \, |\bar{a}(p) - p|.$$

We have already seen this for a couple of lectures now, so this is just a reminder. Next we switch to the agent perspective.

3 Agent Utility and Agent Regret

From the agent's viewpoint, a very basic benchmark is to ignore the day-to-day predictions and instead think about the *overall* frequency of rain. For example, suppose the agent somehow knows that over the entire year it rains on about 50 of the 365 days. Then they could simply pretend that every day is a "typical" day with rain probability equal to that overall frequency, and pick a single action that does best for that fixed probability.

Let us formalize that. We define

$$\beta = \frac{1}{T} \sum_{t=1}^{T} a_t,$$

the empirical frequency of rain in our sequence. If we ignore the predictions and only look at β, the agent could just always play the best response to β, i.e. $\mathrm{BR}(\beta)$.

What we want to measure is: *how much worse is it to follow* $\mathrm{BR}(p_t)$ *day-by-day than to always play* $\mathrm{BR}(\beta)$*?* This gives us a regret notion.

Definition 3.1 (Agent regret)**.** The (average) regret of an agent with utility u is

$$\mathrm{AgentReg}_u(p_{1:T}, a_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \big(u(\mathrm{BR}_u(\beta), a_t) - u(\mathrm{BR}_u(p_t), a_t)\big).$$

So if the regret is small, this says: even in hindsight, having seen the whole sequence and knowing that the base rate is β, we would not have done much better by playing the single best fixed action instead of reacting to the predictions. As stated before, we work with u is bounded in $[-1, 1]$, which is an important part of being able to provide any regret guarantees for agents.

4 Proper Scoring Rules and Utility-Induced Loss

You have already encountered proper scoring rules (or proper losses) in the homework. Let us very quickly recall what they are, and then we will connect them to utilities.

A scoring rule (or loss) is a function

$$\ell : [0,1] \times \{0,1\} \to \mathbb{R},$$

where $\ell(p, a)$ is the loss of predicting p when the realized outcome is a. We also define

$$\ell(p; q) = \mathop{\mathbb{E}}_{a \sim \mathrm{Bern}(q)}[\ell(p, a)]$$

to be the expected loss if the true distribution is Bernoulli(q) but we predict p.

Definition 4.1 (Proper scoring rule / proper loss)**.** A scoring rule ℓ is *proper* if for all $q \in [0,1]$,

$$\ell(q; q) \leq \ell(p; q) \quad \text{for all } p \in [0,1].$$

So if the forecaster knows that the true probability is q, the expected loss is minimized by predicting q. This is a kind of truthfulness condition.

Example: the Brier score

The Brier score is

$$\ell(p, a) = (a - p)^2.$$

If the true distribution is q, then

$$\begin{aligned} \ell(p; q) &= q(1-p)^2 + (1-q)p^2 \\ &= (p-q)^2 + q(1-q), \end{aligned}$$

and since $q(1-q)$ does not depend on p, we are really just minimizing $(p-q)^2$. So the unique minimizer is $p = q$, and the Brier score is proper.

4.1 Loss induced by a utility function

Now we want to connect utilities and scoring rules. Given a utility u and our best-response function $\mathrm{BR}(\cdot)$, we define a loss

$$\ell(p, a) = -u(\mathrm{BR}_u(p), a).$$

Intuitively, this is saying: if I forecast p, the agent plays $\mathrm{BR}(p)$, and we assess how bad that is by taking the *negative* of the utility the agent received.

Lemma 4.2 (Agents and proper scoring rules)**.** *The loss $\ell(p, a) = -u(\mathrm{BR}(p), a)$ is a bounded proper scoring rule. Moreover,* $\mathrm{Regret}_\ell = \mathrm{AgentReg}_u$.

Proof. For properness, fix a true distribution q. Then

$$\ell(p;q) = -\mathop{\mathbb{E}}_{a\sim\mathrm{Bern}(q)}[u(\mathrm{BR}(p),a)].$$

By definition, $\mathrm{BR}(q)$ is the best response to q, so

$$\mathop{\mathbb{E}}_{a\sim\mathrm{Bern}(q)}[u(\mathrm{BR}(q),a)] \geq \mathop{\mathbb{E}}_{a\sim\mathrm{Bern}(q)}[u(\mathrm{BR}(p),a)]$$

for every p. Multiplying both sides by -1 gives $\ell(q;q) \leq \ell(p;q)$, so ℓ is proper.

For the regret identity, we write

$$\begin{aligned} T\cdot\mathrm{Regret}_\ell &= \sum_{t=1}^{T}\big(\ell(p_t,a_t) - \ell(\beta,a_t)\big) \\ &= \sum_{t=1}^{T}\big(-u(\mathrm{BR}(p_t),a_t) + u(\mathrm{BR}(\beta),a_t)\big) \\ &= \sum_{t=1}^{T}\big(u(\mathrm{BR}(\beta),a_t) - u(\mathrm{BR}(p_t),a_t)\big) = T\cdot\mathrm{AgentReg}_u, \end{aligned}$$

which implies $\mathrm{Regret}_\ell = \mathrm{AgentReg}_u$. □

So we can think of agents and proper scoring rules as two equivalent ways of talking about the same thing.

4.2 A Lipschitz fact for proper scoring rules

For bounded proper scoring rules whose range is $[-1,1]$, we can also use the following fact, without proof:

Fact 4.3. *For any bounded proper scoring rule ℓ for all p,q,*

$$\ell(q;q) \leq \ell(p;q) \leq \ell(q;q) + 4|p-q|.$$

5 Main Theorem 1: Calibration Upper-Bounds Agent Regret

We are now ready to connect calibration error and agent regret. The theorem says: if we are calibrated, then every agent with bounded utility is happy. This statement is indeed weaker than what we had shown in the last lecture, which controlled the swap regret of the agent. Nevertheless, we will prove this independently.

Theorem 5.1 (Main Theorem 1)**.** *For any bounded utility u and any sequence $(p_{1:T}, a_{1:T})$,*

$$\mathrm{AgentReg}_u(p_{1:T},a_{1:T}) \leq 4\,\mathrm{CalErr}(p_{1:T},a_{1:T}).$$

Proof. By Lemma 4.2, it is enough to show that for the induced proper loss ℓ we have

$$\mathrm{Regret}_\ell(p_{1:T}, a_{1:T}) \le 4\,\mathrm{CalErr}(p_{1:T}, a_{1:T}).$$

We start from the definition:

$$T \cdot \mathrm{Regret}_\ell = \sum_{t=1}^{T} \big(\ell(p_t, a_t) - \ell(\beta, a_t)\big).$$

Let us group time steps according to the prediction $p_t = p$:

$$T \cdot \mathrm{Regret}_\ell = \sum_{p \in [0,1]} \sum_{t:p_t = p} \big(\ell(p, a_t) - \ell(\beta, a_t)\big).$$

Fix some p and look at all the t with $p_t = p$. Among those, the fraction of times $a_t = 1$ is $\bar{a}(p)$, and the fraction of times $a_t = 0$ is $1 - \bar{a}(p)$. So that inner sum can be rewritten as

$$\begin{aligned}\sum_{t:p_t=p} \big(\ell(p, a_t) - \ell(\beta, a_t)\big) &= N(p)\,\bar{a}(p)\,\ell(p,1) + N(p)\,(1-\bar{a}(p))\,\ell(p,0) \\ &\quad - N(p)\,\bar{a}(p)\,\ell(\beta,1) - N(p)\,(1-\bar{a}(p))\,\ell(\beta,0) \\ &= N(p)\,\big(\ell(p;\bar{a}(p)) - \ell(\beta;\bar{a}(p))\big).\end{aligned}$$

Therefore

$$T \cdot \mathrm{Regret}_\ell = \sum_p N(p)\,\big(\ell(p;\bar{a}(p)) - \ell(\beta;\bar{a}(p))\big).$$

Now we apply Fact 4.3 with $q = \bar{a}(p)$:

$$\ell(p;\bar{a}(p)) - \ell(\beta;\bar{a}(p)) \le 4\,|p - \bar{a}(p)|.$$

So

$$T \cdot \mathrm{Regret}_\ell \le 4 \sum_p N(p)\,|p - \bar{a}(p)| = 4T \cdot \mathrm{CalErr}(p_{1:T}, a_{1:T}).$$

Divide by T and use $\mathrm{Regret}_\ell = \mathrm{AgentReg}_u$ to finish. □

So calibration really is a very strong condition: if you have small calibration error, then for *every* bounded utility function u the agent regret is small.

Next we will see that the converse is not true at all — you can have very large calibration error while all agents are perfectly happy.

6 Main Theorem 2: Calibration Can Be Very Loose

We now construct a slightly awkward example showing that calibration can be a very loose upper bound for agent regret.

Theorem 6.1 (Main Theorem 2)**.** *There exist sequences* $(p_{1:T}, a_{1:T})$ *such that*

$$\mathrm{CalErr}(p_{1:T}, a_{1:T}) \in \Omega(T) \quad \textit{but} \quad \mathrm{AgentReg}_u(p_{1:T}, a_{1:T}) = 0 \textit{ for all bounded } u.$$

Proof sketch. We will first write down a very simple sequence where calibration error is zero, and then we will gently perturb it in a way that makes calibration error huge while not hurting regret.

Let T be even, and consider the outcome sequence

$$a_1 = \cdots = a_{T/2} = 0, \qquad a_{T/2+1} = \cdots = a_t = 1.$$

So exactly half the days it rains.

First, let us take the constant predictions $p_t = \frac{1}{2}$ for all t. Clearly, $\beta = \frac{1}{2}$ as well, and if we look at calibration, there is only one bucket: $p = \frac{1}{2}$, and among those days exactly half have $a_t = 1$. So

$$\mathrm{CalErr}(p_{1:T}, a_{1:T}) = 0.$$

By Main Theorem 1, we also know that $\mathrm{AgentReg}_u(p_{1:T}, a_{1:T}) = 0$ for all bounded u.

Now we slightly modify the predictions. Take numbers

$$z_1, \ldots, z_T \in (0, 0.001)$$

that are all distinct. Define new predictions

$$\hat{p}_t = \begin{cases} \frac{1}{2} - z_t, & t \le T/2, \\ \frac{1}{2} + z_t, & t > T/2. \end{cases}$$

So on days where the outcome is 0 we move the prediction a tiny bit towards 0, and on days where the outcome is 1 we move the prediction a tiny bit towards 1. Also, because all the z_t are distinct, each value $\hat{p}_t$ is unique, so each bucket has size $N(\hat{p}_t) = 1$.

What is the calibration error now? Well,

$$\mathrm{CalErr}(\hat{p}_{1:T}, a_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} |\hat{p}_t - a_t| \ge 0.49$$

where every term in this sum is at least 0.49, because $\hat{p}_t$ is very close to $\frac{1}{2}$ and $a_t \in \{0, 1\}$. So the calibration error is really bad, close to its maximum.

On the other hand, in terms of utility, we have moved every prediction *towards the correct label*. For proper losses (and hence for utilities through our construction) predicting closer to the

truth only helps. There is a monotonicity argument which says that the regret with respect to any bounded proper loss cannot increase under such a monotone move. So we still have

$$\mathrm{AgentReg}_u(\hat{p}_{1:T}, a_{1:T}) \leq 0$$

for all bounded u.

This proves that we can have extremely bad calibration and still have essentially perfect agent performance. □

So calibration is an upper bound on agent regret, but it can be a very loose upper bound. This motivates looking for a notion that is much closer to what agents actually care about, namely *U-calibration* and then *step calibration*.

7 U-Calibration and Step Calibration

Let us now define what we really want.

7.1 U-Calibration

Given a sequence $(p_{1:T}, a_{1:T})$, recall that for each bounded utility u we can compute its agent regret $\mathrm{AgentReg}_u(p_{1:T}, a_{1:T})$.

It is natural to say that a sequence is good if *every* agent is happy, i.e. every bounded u has small regret. We bundle this up under one quantity:

Definition 7.1 (U-calibration)**.** The U-calibration error of $(p_{1:T}, a_{1:T})$ is

$$\text{U-Cal}(p_{1:T}, a_{1:T}) = \sup_{u:\text{bounded}} \mathrm{AgentReg}_u(p_{1:T}, a_{1:T}),$$

where the supremum is over all bounded utilities u.

So saying $\mathrm{U\text{-}Cal}$ is small is saying: for any way an agent might care about the outcomes (subject to boundedness), the regret is small.

However, this is still not a convenient object to work with directly, because we are taking a supremum over a huge family of utilities. The next idea is that we can upper-bound $\mathrm{U\text{-}Cal}$ by a more tractable quantity that talks only about thresholds of the predictions. This is *step calibration*.

7.2 Step Calibration

Think about how we actually make decisions from probabilities. It is very reasonable that a decision rule should have a threshold behavior: maybe when the probability of rain is below 30% we never wear boots, and once it is above 70% we always wear boots, and in between we might do something else, but we do not expect the decision to switch back and forth endlessly as p wiggles around. In particular, no best responses can match the behavior of wearing boots at 31% chance of

rain and then switching again to not wearing boots at 32% chance of rain. This is quite similar to how we describe convexity of the best response regions for Stackelberg games. In a binary decision problems, this leads us to thinking about threshold functions as a proxy for an agents risk tolerance and lead to the next notion of calibration that only controls error at the level of step-functions, not at every point on the level-set of predictions.

Definition 7.2 (Step calibration)**.** For a threshold $\alpha \in [0, 1]$, the step calibration at level α is

$$\mathrm{StepCalErr}_\alpha(p_{1:T}, a_{1:T}) = \frac{1}{T} \sum_{t:p_t \le \alpha} (a_t - p_t).$$

The step calibration error is

$$\mathrm{StepCalErr}(p_{1:T}, a_{1:T}) = \max_{\alpha \in [0,1]} \left| \mathrm{StepCalErr}_\alpha(p_{1:T}, a_{1:T}) \right|.$$

So for each α we gather together all time steps whose prediction is at most α, sum their signed errors $a_t - p_t$, normalize by T, and then take the worst absolute value over all thresholds. It turns out that Step Calibration provides a good upper bound for U-calibration because of the intuitive explanation above about how agent's best response depends on the prediction.

Theorem 7.3 (Qiao and Zhao [2025])**.** *For any sequences* $(p_{1:T}, a_{1:T})$

$$\text{U-Cal}(p_{1:T}, a_{1:T}) \le 8\, \mathrm{StepCalErr}(p_{1:T}, a_{1:T})$$

Sanity check: StepCal is weaker than CalErr

We should at least check that we have not defined something *stronger* than calibration. In fact, step calibration is always at most calibration error.

Fix $\alpha \in [0, 1]$. Then

$$\mathrm{StepCalErr}_\alpha(p_{1:T}, a_{1:T}) = \frac{1}{T} \sum_{t:p_t \le \alpha} (a_t - p_t).$$

We can rewrite the indicator $\mathbf{1}\{p_t \le \alpha\}$ as a sum over all prediction values p' with $p' \le \alpha$:

$$\mathbf{1}\{p_t \le \alpha\} = \sum_{p' \le \alpha} \mathbf{1}\{p_t = p'\}.$$

So

$$\sum_{t:p_t \le \alpha} (a_t - p_t) = \sum_{p' \le \alpha} \sum_{t:p_t = p'} (a_t - p_t).$$

By the triangle inequality,

$$\left| \sum_{t:p_t \le \alpha} (a_t - p_t) \right| \le \sum_{p' \le \alpha} \left| \sum_{t:p_t = p'} (a_t - p_t) \right|.$$

But for each p',

$$\sum_{t:p_t=p'} (a_t - p_t) = N(p')\,(\bar{a}(p') - p'),$$

so

$$\left|\sum_{t:p_t\le\alpha} (a_t - p_t)\right| \le \sum_{p'\le\alpha} N(p')\,|\bar{a}(p') - p'|.$$

Dividing by T and then maximizing over α gives

$$\mathrm{StepCalErr}(p_{1:T}, a_{1:T}) \le \frac{1}{T}\sum_{p'} N(p')\,|\bar{a}(p') - p'| = \mathrm{CalErr}(p_{1:T}, a_{1:T}).$$

So step calibration is indeed a weaker requirement than full calibration: there are sequences where calibration error is large but step calibration is small. On the other hand, step calibration turns out to be tightly connected to Ucalibration and can be achieved with good (roughly $\sqrt{T}$) rates. We will show this next time.

8 Historical Notes

The notion of U-calibration was introduced by Kleinberg et al. [2023]. In their paper, the authors also define a related concept called V-calibration (corresponding to utility functions that are V-shaped), which serves as a simplification that facilitates algorithm design.

In contrast, we take a different route: rather than working directly with V-shaped utilities, we connect U-calibration to threshold functions and to a form of calibration that focuses solely on such step functions. This perspective was developed in the paper by Qiao and Zhao [2025].

We find Step Calibration to be both an interesting and novel research tool and an excellent pedagogical device. As we will see in the next lecture—and again toward the end of the semester—it allows us to design principled algorithms for minimizing this notion of calibration.

References

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5143–5145. PMLR, 2023.

Mingda Qiao and Eric Zhao. Truthfulness of decision-theoretic calibration measures. *arXiv preprint arXiv:2503.02384*, 2025.