CS272 - Theoretical Foundations of Learning, Decisions, and Games

Lecture 22: Calibration and Incentives

November 18

Lecturer: Nika Haghtalab *Readings: Foster and Hart [2021], Haghtalab et al. [2024]*
Scribe: Mark Li, Nika Haghtalab

1 Introduction and Overview

This lecture covers incentives and strategic aspects of calibration and bullet predictions. As before, we explore forecasting algorithms in settings where outcomes may be adversarial. This lecture is divided to two parts:

First, we wwill contineu the previouse lecture material on U-calibration. Showing that there are algorithms for U-calibration that admit a $\sqrt{T}$ calibration rate. As you may recall, U-calibration is sufficeint for ensuring good utility for downstream agents. Given that this is a signifincat improvement over teh $T^{2/3}$ rate of full calibration, it shows that good agent usulity can be gurantees with significantly less effort adn sample complexity

The second part of hte lecture is about truthfulness of calibraiotn measures. Studying how agents to whom making claibrated forecasts is delegated, may have incentives not to reveal all infomration they possess.

2 U-Calibration (cont.)

In the last lecture, we defined

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

We also defined

Definition 2.1 (StepCal)**.** The Stepwise Calibration error is defined as:

$$\text{StepCalErr}(p, a) = \frac{1}{T} \max_{\alpha \in [0,1]} \left| \sum_{t=1}^{T} (a_t - p_t) \mathbb{I}(p_t < \alpha) \right| \tag{1}$$

where a_t is the outcome at time t, p_t is the prediction, and T is the time horizon.

Recall that in the last lecture we showed that step calibration is always upper bounded by the calibration error itself. More important for our purposes, however, is that step calibration is also a good upper bound on U-calibration. The following bound is due to Qiao and Zhao [2025].

Theorem 2.2. *For bounded utilities* $(u \in [-1, 1])$*:*

$$\text{U-Cal}(p_{1:T}, a_{1:T}) \;\leq\; 8\,\text{StepCalErr}(p_{1:T}, a_{1:T}).$$

In this lecture, we are going to analyze the algorithmic bounds for the StepCal. The main theorem that we are going to prove is:

Theorem 2.3. *There exists an algorithm such that*

$$\mathbb{E}[\,\text{StepCalErr}(p_{1:T}, a_{1:T})\,] \;\leq\; \frac{1}{\sqrt{T}} \tag{2}$$

even when outcomes are adversarial.

Stepwise Calibration serves as an upper bound on Agent Regret, which is the primary metric of interest in this context. It represents a form of calibration that accounts for the stepwise structure in decisions, specifically where a Best Response (BR) flips from one action to another. Consequently, if the focus is on this "stepiness" or threshold-based behavior, this framework provides the necessary tools.

2.1 Proving StepCalibration Rates

There are several direct algorithmic approaches for proving this bound, or even bounding U-calibration directly, such as Qiao and Zhao [2025], Kleinberg et al. [2023]. In this lecture, however, we will take an alternative approach: we will use Hart's minimax argument to reduce the problem to the "Bayesian setting," where the adversary announces their strategy first. Forecasting is subsequently conducted in a Bayesian setting, with the exception that we must now make predictions that lie on a discrete grid. See Lecture 19 for a review of this technique. This perspective will allow us to see clearly why step-calibration is significantly easier to control than full calibration from a statistical standpoint.

A note on bounding step-calibration To the best of Nika's knowledge, the use of Hart's minimax argument to bound the step-calibration error has not appeared before. The proof presented here is one that Nika produced for pedagogical purposes. There are two main reasons we take this minimax route instead of a direct algorithmic perspective:

1. We find the Hart-style minimax proof technique to be a great pedagogical and principled tool for calibration-like objectives. In comparison, existing algorithms for bounding step-calibration are either somewhat ad hoc or rely on tools more advanced than what you have encountered so far in this course.

2. We will revisit step-calibration and its extensions in the last week of the semester. There, we will introduce principled algorithm-design methods that can be used not only to bound step-calibration but also to address a broad range of related problems. So those of you interested in an algorithmic treatment will soon see one from a more systematic and principled perspective (under the umbrella of multi-objective learning).

The main theoretical result establishes that there exists an algorithm achieving low calibration error even against adversarial outcomes.

Theorem 2.4. *Given a known adversary's mixed strategy or distribution over outcomes $q_1, \ldots, q_T$, there is a forecasting algorithm makes forecasts $p_1, \ldots, p_T$ chosen from a discretized grid $D := \{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}$. The expected StepCal is bounded by:*

$$\mathbb{E}[\text{StepCalErr}(p, a)] \leq O\left(\frac{1}{N} + \sqrt{\frac{\log N}{T}}\right) \tag{3}$$

This implies that if we choose $N = \sqrt{T}$, the error decays at a rate of roughly $O(1/\sqrt{T})$.

2.2 Proof Sketch: Hart's Minimax Argument

We utilize Hart's minimax argument to simplify the analysis. The original forecasting problem involves an adaptive adversary who selects outcomes sequentially. By applying the Minimax Theorem, we can effectively swap the order of play to consider a scenario where the adversary must first reveal their entire strategy (the distribution of outcomes $q_1, \ldots, q_T$). Consequently, it suffices to prove that a strategy exists which achieves the desired error bound against this fixed, known distribution. The theorem then guarantees that a strategy achieving the same expected bound exists for the original online setting.

2.2.1 Strategy Definition

Since the adversary's strategy $q_t = \mathbb{E}[a_t | \text{history}]$ is known to the forecaster, the forecaster employs a simple rounding strategy. For each time step t, the forecaster chooses p_t by rounding the true probability q_t to the nearest point in the grid D:

$$p_t = \text{argmin}_{p \in D} |q_t - p|$$

Since the grid points are spaced by $\frac{1}{N}$, the rounding error is bounded deterministically:

$$|q_t - p_t| \leq \frac{1}{N}$$

2.2.2 Decomposition of StepCal

We define the StepCal metric with respect to the grid thresholds $d \in D$. The objective is to bound the expected maximum deviation:

$$\mathbb{E}[\text{StepCalErr}] = \mathbb{E}\left[\frac{1}{T} \max_{d \in D} \left| \sum_{t=1}^{T} (a_t - p_t) \mathbb{I}(p_t < d) \right| \right]$$

We decompose the term inside the summation by adding and subtracting q_t:

$$\sum_{t=1}^{T}(a_t - p_t)\mathbb{I}(p_t < d) = \sum_{t=1}^{T}(a_t - q_t)\mathbb{I}(p_t < d) + \sum_{t=1}^{T}(q_t - p_t)\mathbb{I}(p_t < d)$$

Using the triangle inequality ($|A+B| \le |A| + |B|$), we separate the error into a **bias term** (due to rounding) and a **variance term** (due to randomness in outcomes):

$$\mathbb{E}[\text{StepCalErr}] \le \underbrace{\mathbb{E}\left[\frac{1}{T}\max_{d\in D}\left|\sum_{t=1}^{T}(q_t - p_t)\mathbb{I}(p_t < d)\right|\right]}_{\text{Bias Term}} + \underbrace{\mathbb{E}\left[\frac{1}{T}\max_{d\in D}\left|\sum_{t=1}^{T}(a_t - q_t)\mathbb{I}(p_t < d)\right|\right]}_{\text{Variance Term}}$$

2.2.3 Bounding the Bias Term

For the bias term, we use the property that our rounding strategy ensures $|q_t - p_t| \le \frac{1}{N}$ for all t.

$$\left|\sum_{t=1}^{T}(q_t - p_t)\mathbb{I}(p_t < d)\right| \le \sum_{t=1}^{T}|q_t - p_t| \cdot 1 \le \sum_{t=1}^{T}\frac{1}{N} = \frac{T}{N}$$

Dividing by T, the bias term is bounded by:

$$\text{Bias Term} \le \frac{1}{N}$$

2.2.4 Bounding the Variance Term (Concentration)

We now focus on the variance term: $\frac{1}{T}\mathbb{E}[\max_{d\in D}|S(d)|]$, where $S(d) = \sum_{t=1}^{T}(a_t - q_t)\mathbb{I}(p_t < d)$. Fix a specific threshold $d \in D$. Let $Z_t = (a_t - q_t)\mathbb{I}(p_t < d)$. Since $a_t \in \{0,1\}$ is drawn from a Bernoulli distribution with parameter q_t, we have $\mathbb{E}[a_t - q_t|\text{history}] = 0$. Thus, the partial sums form a martingale. Since the increments are bounded ($|Z_t| \le 1$), we can apply the **Azuma-Hoeffding inequality**:

$$\Pr[|S(d)| \ge \lambda] \le 2\exp\left(\frac{-\lambda^2}{2T}\right)$$

2.2.5 Union Bound and Expectation

We need to bound the maximum over all $d \in D$. Since the size of the grid is $|D| = N + 1 \approx N$, we apply a union bound:

$$\Pr\left[\max_{d\in D}|S(d)| \ge \lambda\right] \le \sum_{d\in D}\Pr[|S(d)| \ge \lambda] \le 2N\exp\left(\frac{-\lambda^2}{2T}\right)$$

To find the expectation $\mathbb{E}[\max_{d\in D} |S(d)|]$, we integrate the tail probability. Setting $\lambda \approx \sqrt{2T \log N}$, the tail probability becomes negligible. Formally, this integration yields a bound of order:

$$\mathbb{E}\left[\max_{d\in D} |S(d)|\right] \leq O(\sqrt{T \log N})$$

Dividing by T to normalize:

$$\text{Variance Term} \leq O\left(\frac{\sqrt{T \log N}}{T}\right) = O\left(\sqrt{\frac{\log N}{T}}\right)$$

2.2.6 Conclusion

Combining the bounds for the bias and variance terms:

$$\mathbb{E}[\text{StepCalErr}] \leq \frac{1}{N} + O\left(\sqrt{\frac{\log N}{T}}\right)$$

This confirms the theorem statement. □

3 Truthfulness of Calibration

Calibration measures are commonly used to evaluate the performance of forecasters, so it is important that their use encourages forecasters to incorporate the highest-quality information available to them about the next outcome. This property is formally referred to as *truthfulness*, and it requires that a calibration measure incentivizes forecasters to predict truthfully when the true distribution of the next outcome is known to them.

But is this the case for common calibration errors, such as the expected calibration error? As we will show below, the expected calibration error fails to satisfy this requirement. Let us begin by identifying two modes of failure.

3.1 Two failure modes of the expected calibration error

Recall that the expected calibration error (without scaling by $1/T$) of a sequence of predictions $p_1, \ldots, p_T \in [0, 1]$ and binary outcomes $x_1, \ldots, x_T \in \{0, 1\}$ is

$$\text{CalErr}(x_{1:T}, p_{1:T}) \;=\; \sum_{\alpha\in[0,1]} n_T(\alpha) \left|\hat{p}_T(\alpha) - \alpha\right|,$$

where $n_T(\alpha) = \#\{t : p_t = \alpha\}$ is the number of times the value α is predicted, and

$$\hat{p}_T(\alpha) \;=\; \frac{1}{n_T(\alpha)} \sum_{t: p_t = \alpha} x_t$$

is the empirical frequency of ones among those times. In a Bayesian setting, the *truthful* forecaster at time t predicts

$$p_t^\star \;=\; \Pr\bigl[x_t = 1 \mid x_1, \ldots, x_{t-1}\bigr].$$

A calibration measure is called *truthful* if this Bayesian forecaster nearly minimizes the expected penalty. We now see that the standard expected calibration error fails badly in this sense.

Failure mode 1: covering up past mistakes over time

Consider a time horizon T that is a multiple of 3, and partition it into $T/3$ consecutive blocks of length 3. For each block $j = 1, \ldots, T/3$, Nature samples a fresh bit

$$B_j \sim \mathrm{Bernoulli}\bigl(\tfrac{1}{2}\bigr)\,,$$

and then sets the three outcomes in that block to

$$x_{3j-2} = B_j, \qquad x_{3j-1} = 0, \qquad x_{3j} = 1.$$

The forecaster knows this data-generating process.

Truthful forecaster. The truthful Bayesian forecaster predicts, in each block,

$$p^\star_{3j-2} = \tfrac{1}{2}, \qquad p^\star_{3j-1} = 0, \qquad p^\star_{3j} = 1.$$

The predictions 0 and 1 are perfectly calibrated, because whenever the forecaster outputs 0 (resp. 1), the outcome is deterministically 0 (resp. 1). The only nontrivial part of the expected calibration error comes from the value $\alpha = \frac{1}{2}$.

There are exactly $T/3$ time steps at which $p_t^\star = 1/2$, namely the first step of each block. At those steps the outcomes are the i.i.d. bits $B_1, \ldots, B_{T/3}$. Thus

$$\hat{p}_T\bigl(\tfrac{1}{2}\bigr) = \frac{1}{T/3} \sum_{j=1}^{T/3} B_j,$$

and by standard concentration for sums of i.i.d. $\mathrm{Bernoulli}(1/2)$ variables, the deviation

$$\bigl|\hat{p}_T(1/2) - 1/2\bigr|$$

is typically on the order of $1/\sqrt{T}$, and in particular its expectation is $\Theta(1/\sqrt{T})$. Therefore the contribution of $\alpha = 1/2$ to the expected calibration error of the truthful forecaster satisfies

$$\mathbb{E}[\mathrm{CalErr}(x_{1:T}, p^\star_{1:T})] = \mathbb{E}\bigl[n_T\bigl(\tfrac{1}{2}\bigr)\,\bigl|\hat{p}_T(1/2) - 1/2\bigr|\bigr] = \Theta\Bigl(\sqrt{T}\Bigr)\,,$$

since $n_T(1/2) = T/3$ while the other bins contribute zero.

A strategic forecaster that "repairs" past mistakes. Now we construct an alternative forecaster that uses its knowledge of the process to game the expected calibration error . In each block it proceeds as follows.

1. On the first step of block j, it predicts $p_{3j-2} = 1/2$ (same as the truthful forecaster) and observes $x_{3j-2} = B_j$.
2. If $B_j = 1$, then on the second step it *pretends not to know* that the outcome will be 0 and predicts $p_{3j-1} = 1/2$, while on the third step it predicts $p_{3j} = 1$.
3. If $B_j = 0$, then on the second step it predicts $p_{3j-1} = 0$ and on the third step it deliberately misreports $p_{3j} = 1/2$, despite knowing for sure that $x_{3j} = 1$.

In either case, in block j there are *two* time steps with prediction $1/2$. If $B_j = 1$ their outcomes are $(1, 0)$; if $B_j = 0$ their outcomes are $(0, 1)$. Hence in every block, among predictions equal to $1/2$ we see exactly one 1 and one 0, so the empirical frequency of ones conditioned on predicting $1/2$ is

$$\hat{p}_T(1/2) = \frac{\#\{\text{ones with } p_t = 1/2\}}{\#\{t : p_t = 1/2\}} = \frac{(T/3)}{2T/3} = \tfrac{1}{2}.$$

For $\alpha = 0$ and $\alpha = 1$, the forecaster only predicts these values on steps where the outcome is deterministically 0 or 1, so those bins are also perfectly calibrated. Consequently

$$\mathrm{CalErr}(x_{1:T}, p_{1:T}^{\mathrm{cover}}) = 0$$

for every realization of the process, where $p_{1:T}^{\mathrm{cover}}$ denotes this "cover up past mistakes" strategy.

Interpretation. On this simple distribution,

$$\mathbb{E}\big[\mathrm{CalErr}(x_{1:T}, p_{1:T}^{\mathrm{cover}})\big] = 0 \quad \text{but} \quad \mathbb{E}\big[\mathrm{CalErr}(x_{1:T}, p_{1:T}^{\star})\big] = \Theta\Big(\sqrt{T}\Big)\,.$$

Thus a forecaster that *deliberately lies* in later rounds, using future predictions to cancel random fluctuations from earlier rounds, is judged strictly better (in fact, perfect) by the expected calibration error than the truthful Bayesian forecaster.

The underlying problem is that the expected calibration error is evaluated as a single average over all time steps. A strategic forecaster can track, for each probability level α, the current empirical bias $\hat{p}_T(\alpha) - \alpha$ and then manipulate future predictions to drive this bias toward zero, even if doing so requires knowingly mispredicting the next outcome.

Failure mode 2: sharp probability buckets and discontinuity

The second issue is that the expected calibration error is extremely sensitive to the *exact* values of the predictions. Because it groups together time steps with the same numerical probability α (via the indicators $1\{p_t = \alpha\}$), arbitrarily small changes in the predictions can change which samples share a bucket and can dramatically alter the measured error.

A distribution with many slightly different probabilities. Fix a small constant $\delta > 0$, say $\delta = 0.01$. For each time step $t = 1, \dots, T$ Nature independently draws a probability

$$p_t^\star \sim \mathrm{Unif}\big[\, \tfrac{1}{2} - \delta,\ \tfrac{1}{2} + \delta \,\big],$$

and then generates the outcome

$$x_t \mid p_t^\star \ \sim\ \mathrm{Bernoulli}(p_t^\star),$$

independently across t. The forecaster knows $p_t^\star$ at time t, so the truthful forecast is $p_t = p_t^\star$.

Because each $p_t^\star$ is drawn from a continuous distribution, with probability 1 all the values $p_1^\star, \dots, p_T^\star$ are distinct, so each one forms its own "bucket" in the definition of expected calibration error .

expected calibration error of the truthful forecaster. Under the truthful forecaster $p_t = p_t^\star$, every bucket $\alpha = p_t^\star$ contains exactly one time step, and the empirical frequency inside that bucket is simply the realized outcome x_t. Thus the expected calibration error simplifies to

$$\mathrm{CalErr}(x_{1:T}, p_{1:T}^\star) \ =\ \sum_{t=1}^{T} \big| x_t - p_t^\star \big| \ge 0.49T,$$

much like an example we saw in the previous lecture.

A "smoothed" forecaster that looks better by expected calibration error . Consider instead a forecaster that *ignores* the precise probabilities $p_t^\star$ and simply predicts

$$p_t = \tfrac{1}{2} \qquad \text{for every } t.$$

All T time steps now fall into the single bucket $\alpha = 1/2$, so

$$\hat{p}_T(1/2) = \frac{1}{T}\sum_{t=1}^{T} x_t, \qquad \mathrm{CalErr}(x_{1:T}, p_{1:T}) = T\, \big|\hat{p}_T(1/2) - \tfrac{1}{2}\big|\,.$$

The random variables x_t are independent and bounded in $[0, 1]$, so their sum concentrates. That gives us

$$\mathbb{E}[\mathrm{CalErr}(x_{1:T}, p_{1:T})] \ \le\ O\Big(\sqrt{T}\Big)\,.$$

Hence, on this distribution the uninformative "always predict $1/2$" forecaster achieves *asymptotically much smaller* expected expected calibration error than the truthful forecaster that outputs the correct per-round probabilities.

Interpretation. Together, the two failure modes show that low expected calibration error does not guarantee that a forecaster is using their knowledge of the data-generating process truthfully: future predictions can be chosen to repair past calibration mistakes, and small changes in the reported probabilities can exploit the discontinuities created by the sharp calibration buckets.

4 Proposed Solution: Subsampled Smooth Calibration

To mitigate these failure modes—specifically the ability to game the history and the fragility of sharp boundaries—we propose a robust metric. The proposed solution involves two key modifications:

1. **Smoothing:** Replacing the sharp indicator functions of buckets with Lipschitz continuous functions (f).

2. **Subsampling:** Preventing the agent from fully knowing when they are evaluated.

The metric is defined by evaluating on a random subset S (subsampled uniformly or with size constraints like $\sqrt{T}$). The definition given is:

$$SSCE(p_{1:T}, a_{1:T}) = \mathbb{E}_S \left[\sup_{f \in \text{Lip}} \left| \sum_{t \in S} (a_t - p_t) f(p_t) \right| \right] \tag{4}$$

By introducing this variance and smoothing, the metric becomes harder to game via historical correction and robust to boundary issues. Generally, an agent aiming to minimize a calibration measure (CM) satisfies a bound of the form:

$$\mathbb{E}[CM(\text{Alg}, a_{1:T})] \leq \alpha \cdot OPT_{CM}(D) + \beta$$

where larger α, β indicate a departure from strict truthfulness.

Theorem 4.1. *The subsampled smooth calibration error is nearly truthful. In particular,* SSCE *is truthful with parameters $\alpha = O(1)$ and $\beta = 0$.*

On the other hand, we have the following negative result:

Theorem 4.2. *As shown above, the expected calibration error has a large truthfulness gap. In particular, there exists a sequence of outcomes and a data-generating process under which the truthful forecaster incurs calibration error $\Omega(T)$, while a strategic forecaster can achieve calibration error 0.*

Similarly, all other well-known calibration measures suffer from a large truthfulness gap. For example, in the homework you will show that the smooth calibration error incurs a $\sqrt{T}$ gap—i.e., the truthful strategy suffers loss $\Omega(\sqrt{T})$ while a strategic forecaster can obtain error 0.

4.1 What about Proper Scoring Rules?

By definition, proper scoring rules are *perfectly truthful*: a forecaster minimizes the expected loss by reporting their true belief. However, these scoring rules aren't really calibration measures! In particular, proper scoring rules are *not* "complete" as a measure of calibration.

Completeness here means the following: if a forecaster reports the correct probability at every time step, then the cumulative penalty should be $o(T)$. Proper scoring rules fail this requirement.

Even when a forecaster predicts the true probability at all T steps, the cumulative proper score (e.g., the Brier score $\sum_t (x_t - p_t)^2$) can be $\Omega(T)$. In contrast, every calibration measure must satisfy completeness: truthful predictions should incur cumulative $o(T)$ error.

In this sense, calibration measures and proper scoring rules differ fundamentally. Proper scoring rules ensure truthful reporting but do *not* isolate calibration error like calibration measures do. Can we find a calibration measure that has no gap at all in truthfulness? This remains an open problem!

5 Acknowledgment

Mark Li would like to thank his classmates Sam Huang and Jivat N. Kaur for their help in lending him their notes.

References

Dean P Foster and Sergiu Hart. Forecast hedging and calibration. *Journal of Political Economy*, 129(12):3447–3490, 2021.

Nika Haghtalab, Mingda Qiao, Kunhe Yang, and Eric Zhao. Truthfulness of calibration measures. *Advances in Neural Information Processing Systems*, 37:117237–117290, 2024.

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.