CS7180/7880 Lecture 3: Calibration and Decision Making

Guest Lecture by Lunjia Hu

Fall 2026

1 Decision Tasks and Best Response

Definition 1.1 (Decision Task)**.** *Let Y be the set of possible outcomes. A decision task $G = (A, U)$ consists of an action set A and a utility function $U : A \times Y \to \mathbb{R}$.*

In this lecture, we will focus on binary outcomes $Y = \{0, 1\}$.

If we know the true outcome $y \in \{0, 1\}$, it is not hard to pick the best action $a \in A$ that maximizes the utility $u(a, y)$. If we don't know y directly but instead have an accurate prediction $p \in [0, 1]$ about its expectation $\mathbb{E}[y]$, we can pick the best action $a \in A$ *in response to* the prediction p to maximize the *expected* utility:

Definition 1.2 (Best Response)**.** *Given a decision task $G = (A, U)$, we define the* best-response function $\mathsf{BR}_G : [0, 1] \to A$ *as*

$$\mathsf{BR}_G(p) := \arg\max_{a \in A} \mathbb{E}_{y \sim \mathsf{Ber}(p)}[U(a, y)] \quad \textit{for every } p \in [0, 1].$$

2 Proper Scoring Rules and the Revelation Principle

Given a prediction $p \in [0, 1]$ for an outcome $y \in \{0, 1\}$, a *scoring rule* $s : [0, 1] \times \{0, 1\} \to \mathbb{R}$ evaluates the quality $s(p, y)$ of the prediction p on the outcome y.

2.1 Scoring Rule Induced by Decision Task

According to Definition 1.2, $\mathsf{BR}_G(p)$ is the best action to take if our goal is to maximize our expected utility *assuming that* $\mathbb{E}[y] = p$. If this assumption does not hold and p is not a good predictor for $\mathbb{E}[y]$, then $\mathsf{BR}_G(p)$ may not maximize the expected utility. We can thus evaluate the quality of p using the true utility of $\mathsf{BR}_G(y)$.

Definition 2.1 (Scoring Rule Induced by Decision Task)**.** *Let $G = (A, U)$ be a decision task, and let $\mathsf{BR}_G : [0, 1] \to A$ be its corresponding best-response function. We define an induced scoring rule $s_G : [0, 1] \times \{0, 1\} \to \mathbb{R}$ by*

$$s_G(p, y) := U(\mathsf{BR}_G(p), y) \quad \textit{for every } p \in [0, 1], \textit{ and } y \in \{0, 1\}.$$

The design of scoring rules is a fundamental topic in economics theory. There may be many reasons to prefer some scoring rules over others. A specific family of scoring rules that people often like is the family of *proper* scoring rules. Proper scoring rules incentivize predictors to predict the true expectation $\mathbb{E}[y]$. In economics terms, proper scoring rules *elicit* the true expectation $\mathbb{E}[y]$.

Definition 2.2 (Proper Scoring Rule)**.** *We say a scoring rule* $s : [0,1] \times \{0,1\} \to \mathbb{R}$ *is* proper *if for every* $p, p' \in [0,1]$,

$$\mathbb{E}_{y \sim \mathsf{Ber}(p)}[s(p,y)] \geq \mathbb{E}_{y \sim \mathsf{Ber}(p)}[s(p',y)].$$

Example 2.1. *Let* $r \in [0,1]$.*Let* $G = (A,U)$ *be the decision task where* $A = \{0,1\}$ *and*

$$U(a,y) = \begin{cases} 1, & \text{if } a = y; \\ r, & \text{if } a = 0 \text{ and } y = 1; \\ 1-r, & \text{if } a = 1 \text{ and } y = 0. \end{cases}$$

The induced scoring rule is

$$s_G(p,y) = \begin{cases} 1, & \text{if } p < r \text{ and } y = 0; \\ 1, & \text{if } p \geq r \text{ and } y = 1; \\ r, & \text{if } p < r \text{ and } y = 1; \\ 1-r, & \text{if } p \geq r \text{ and } y = 0. \end{cases}$$

Theorem 2.1 (Revelation Principle)**.** *A scoring rule* $s : [0,1] \times \{0,1\} \to \mathbb{R}$ *is proper if and only if there exists a decision task* G *such that* $s = s_G$.

In particular, the scoring rule in Example 2.1 is proper. Here are some other examples of proper scoring rules:

$$s(p,y) = -(p-y)^2, \qquad \text{(Negation of Brier loss)}$$

$$s(p,y) = y \log p + (1-y)\log(1-p). \qquad \text{(Negation of cross-entropy loss)}$$

It is a good exercise to verify that these are indeed proper scoring rules.

3 External Regret and Swap Regret

3.1 Regret of Actions

Fix a decision task $G = (A, U)$.

1. Sequence of actions $a_1, \dots, a_T \in A$;
2. Sequence of outcomes $y_1, \dots, y_T \in Y$.

The external and swap regret of $a_1, \dots, a_T$ on $y_1, \dots, y_T$ are defined as follows.

Definition 3.1 (External Regret)**.**

$$\mathsf{Regret}_G(a_1, \dots, a_T; y_1, \dots, y_T) := \max_{a \in A} \frac{1}{T} \sum_{t=1}^{T} U(a, y_t) - \frac{1}{T} \sum_{t=1}^{T} U(a_t, y_t).$$

The external regret can be positive, zero, or negative.

Definition 3.2 (Swap Regret)**.**

$$\mathsf{SR}_G(a_1, \dots, a_T; y_1, \dots, y_T) := \max_{\sigma : A \to A} \frac{1}{T} \sum_{t=1}^{T} U(\sigma(a_t), y_t) - \frac{1}{T} \sum_{t=1}^{T} U(a_t, y_t).$$

The swap regret is always nonnegative. Also, the swap regret is always at least the external regret.

3.2 Regret of Predictions

Fix a scoring rule s.

1. Sequence of predictions $p_1, \ldots, p_T \in [0,1]$;
2. Sequence of outcomes $y_1, \ldots, y_T \in Y$.

The external and swap regret of $p_1, \ldots, p_T$ on $y_1, \ldots, y_T$ are defined as follows.

Definition 3.3 (External Regret)**.**

$$\mathsf{Regret}_s(p_1, \ldots, p_T; y_1, \ldots, y_T) := \max_{p \in [0,1]} \frac{1}{T} \sum_{t=1}^{T} s(p, y_t) - \frac{1}{T} \sum_{t=1}^{T} s(p_t, y_t).$$

The external regret can be positive, zero, or negative.

Definition 3.4 (Swap Regret)**.**

$$\mathsf{SR}_s(p_1, \ldots, p_T; y_1, \ldots, y_T) := \max_{\sigma:[0,1]\to[0,1]} \frac{1}{T} \sum_{t=1}^{T} s(\sigma(p_t), y_t) - \frac{1}{T} \sum_{t=1}^{T} s(p_t, y_t).$$

The swap regret is always nonnegative. Also, the swap regret is always at least the external regret.

3.3 Calibration Implies No Swap Regret

Theorem 3.1. *Let $p_1, \ldots, p_T \in [0,1]$ be a sequence of predictions and let $y_1, \ldots, y_T \in \{0,1\}$ the corresponding outcomes. The following three statements are equivalent:*

1. *$p_1, \ldots, p_T$ is calibrated on $y_1, \ldots, y_T$, i.e.,*

$$\sum_{t=1}^{T} y_t \mathbf{I}[p_t = p] = p \sum_{t=1}^{T} \mathbf{I}[p_t = p] \quad \textit{for every } p \in [0,1];$$

2. *for every proper scoring rule s,*

$$\mathsf{SR}_s(p_1, \ldots, p_T; y_1, \ldots, y_T) = 0;$$

3. *for every decision task G, letting $a_t := \mathsf{BR}_G(p_t)$ for $t = 1, \ldots, T$,*

$$\mathsf{SR}_G(a_1, \ldots, a_T; y_1, \ldots, y_T) = 0.$$

4 Calibration Decision Loss

Theorem 3.1 suggests that we can measure the calibration error of predictions using the swap regret on proper scoring rules or downstream decision tasks. This is exactly the idea of Calibration Decision Loss.

Definition 4.1 (Calibration Decision Loss [HW24])**.** *Let $p_1, \dots, p_T \in [0,1]$ be a sequence of predictions and let $y_1, \dots, y_T \in \{0,1\}$ be the corresponding outcomes. Let $\mathcal{S}$ be the family of proper scoring rules $s : [0,1] \times \{0,1\} \to [0,1]$ with a bounded co-domain $[0,1]$. Define the* Calibration Decision Loss (CDL) *of $p_1, \dots, p_T$ on $y_1, \dots, y_T$ as*

$$\mathsf{CDL}(p_1, \dots, p_T; y_1, \dots, y_T) := \sup_{s \in \mathcal{S}} \mathsf{SR}_s(p_1, \dots, p_T; y_1, \dots, y_T).$$

The above definition defines CDL as the maximum swap regret on bounded proper scoring rules. Similarly to the equivalence between Item 2 and Item 3 in Theorem 3.1, we can equivalently define CDL as the maximum swap regret on downstream decision tasks with bounded utility functions:

Lemma 4.1. *Let $p_1, \dots, p_T \in [0,1]$ be a sequence of predictions and let $y_1, \dots, y_T \in \{0,1\}$ be the corresponding outcomes. Let $\mathcal{G}$ be the family of decision tasks $G = (A, U)$ with arbitrary action set A and arbitrary utility function $U : A \times \{0,1\} \to [0,1]$ with a bounded co-domain $[0,1]$. Then*

$$\mathsf{CDL}(p_1, \dots, p_T; y_1, \dots, y_T) = \sup_{G \in \mathcal{G}} \mathsf{SR}_G(\mathsf{BR}_G(p_1), \dots, \mathsf{BR}_G(p_T); y_1, \dots, y_T).$$

This lemma shows that CDL provides an upper bound on the swap regret of downstream decision tasks with bounded utility functions, assuming that the decision maker applies the best-response function to the predictions. This property is called *actionability* [RSB+25, BHNU26].

References

[BHNU26] Konstantina Bairaktari, Lunjia Hu, Huy L Nguyen, and Jonathan Ullman. Testable and actionable calibration for full swap regret. *arXiv preprint arXiv:2605.17749*, 2026.

[HW24] Lunjia Hu and Yifan Wu. Calibration error for decision making. *arXiv preprint arXiv:2404.13503*, 2024.

[RSB+25] Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4937–4972. PMLR, 30 Jun–04 Jul 2025. URL: `https://proceedings.mlr.press/v291/rossellini25a.html`.