CS7180/7880 Lecture 2: Calibration Measures

Guest Lecture by Lunjia Hu

Fall 2026

1 The Challenge: Measuring Calibration Error Meaningfully

Basic setup:

- Feature domain $\mathcal{X}$;
- Ground-truth distribution $\mathcal{D}$ on data points $(x, y) \in \mathcal{X} \times \{0, 1\}$;
- Predictor $f : \mathcal{X} \to [0, 1]$.

Definition 1.1 (Perfect calibration)**.** *We say predictor f is (perfectly) calibrated on $\mathcal{D}$ if for every $p \in [0, 1]$ in the range of f with $\Pr_{\mathcal{D}}[f(x) = p] > 0$,*

$$\mathbb{E}_{\mathcal{D}}[y|f(x) = p] = p.$$

In practice, a predictor is rarely perfectly calibrated. Can we quantify its level of miscalibration in a meaningful way? This is the question of designing *calibration measures*.

Formally, a calibration measure assigns a value from $[0, 1]$ to every predictor-distribution pair $(f, \mathcal{D})$ to indicate the level of miscalibration of f on $\mathcal{D}$. We discuss some important calibration measures in the following sections.

2 ECE

In the previous lecture, we have seen a basic calibration measure: ECE.

Definition 2.1. *Given predictor f and distribution $\mathcal{D}$, define $\bar{y}_p := \mathbb{E}_{\mathcal{D}}[y|f(x) = p]$ for every $p \in [0, 1]$. We define the ECE of f on $\mathcal{D}$ as*

$$\mathsf{ECE}(f, \mathcal{D}) = \mathbb{E}[|p - \bar{y}_p|],$$

where the expectation is over $p = f(x)$ for $(x, y) \sim \mathcal{D}$.

Example 2.1. *Let $\mathcal{D}$ be the uniform distribution on $\{(x_0, 0), (x_1, 1)\}$. Let f_α be the predictor such that $f(x_0) = \alpha, f(x_1) = 1 - \alpha$. Then*

$$\mathsf{ECE}(f, \mathcal{D}) = \begin{cases} \alpha, & \text{if } \alpha \in [0, 1/2) \cup (1/2, 1]; \\ 0, & \text{if } \alpha = 1/2. \end{cases}$$

The example above shows that ECE is not a continuous function of the predictor f. Also, it is not always possible to estimate ECE up to a small error, say $1/4$, given a finite set of data points. That is, ECE is not *testable* (see Section 6). In general, given a finite dataset of arbitrary size, it is not always possible to distinguish $\mathsf{ECE}(f, \mathcal{D}) = 0$ from $\mathsf{ECE}(f, \mathcal{D}) > 1/4$ with success probability above 50%. The reason, roughly speaking, is that the conditional expectation $\bar{y}_p$ cannot be estimated within small error unless at least two data points x, x' receive the same prediction $f(x) = f(x') = p$. The probability of seeing such a collision in a finite data set can be zero.

3 Binned ECE

Definition 3.1. *Let $B_1, \ldots, B_k$ partition $[0, 1]$ into k bins. Given predictor f and distribution $\mathcal{D}$, define $\bar{y}_i := \mathbb{E}_{\mathcal{D}}[y | f(x) \in B_i]$ and $\bar{p}_i := \mathbb{E}_{\mathcal{D}}[f(x) | f(x) \in B_i]$. We define the binned ECE of f on $\mathcal{D}$ as*

$$\mathsf{bECE}(f, \mathcal{D}) := \sum_{i=1}^{k} \Pr_{\mathcal{D}}[f(x) \in B_i] \cdot |\bar{y}_i - \bar{p}_i|.$$

Binned ECE is still not continuous, but it can be estimated within error $O(\sqrt{k/n})$ with success probability at least 99% using n data points.

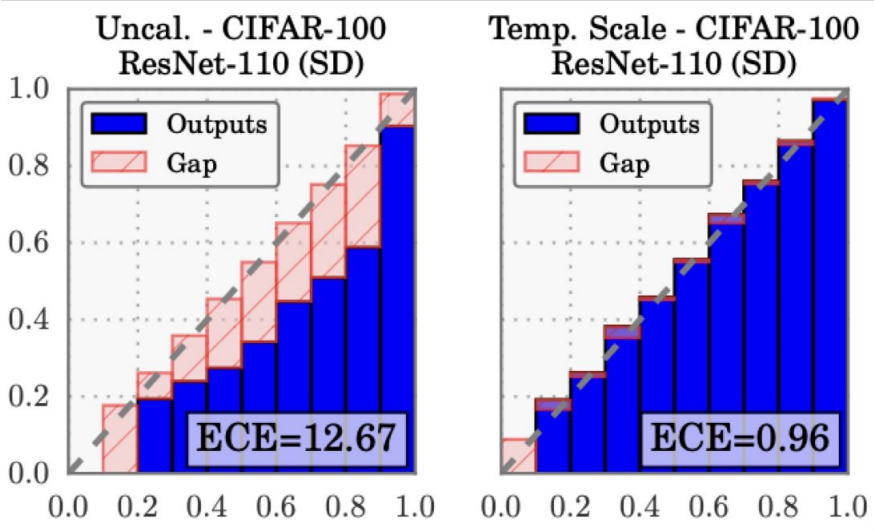

Figure 1: Reliability diagrams for ResNet-110 on CIFAR-100 before and after temperature scaling. Source: [GPSW17].

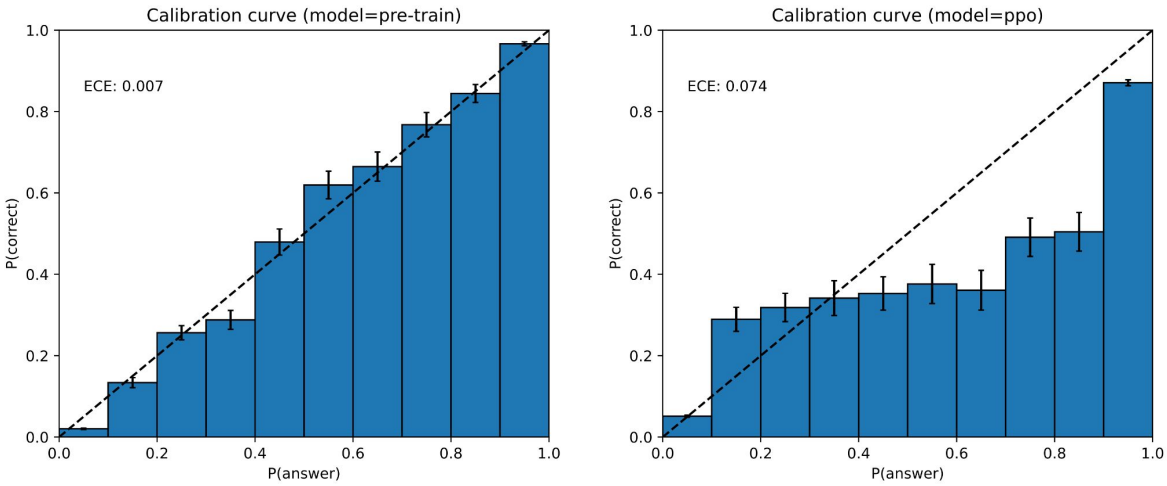

Figure 2: Calibration curves for the pre-trained and PPO models. Source: [Ope23].

Binned ECE is not robust to the choice of the number of bins k. Figure 3 illustrates this issue for models after temperature scaling. With 15 or 100 bins, models with higher classification error tend to have lower binned ECE. With 5000 bins, the trend reverses: models with higher classification error tend to have higher binned ECE. Thus, changing only the number of bins can reverse which predictor appears better calibrated, even though the predictions and outcomes are unchanged.

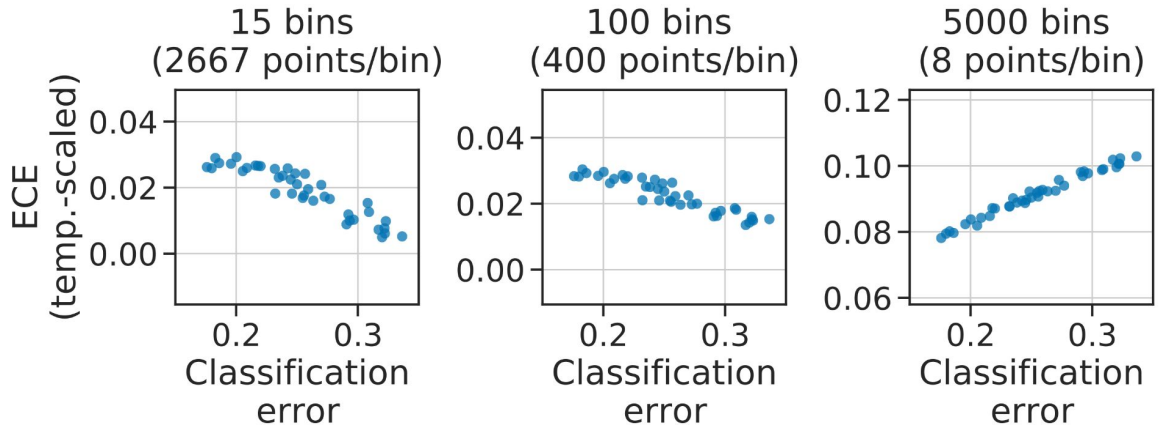

Figure 3: Binned ECE versus classification error after temperature scaling, using 15, 100, and 5000 bins. The direction of the relationship depends on the bin count. Source: [MDR+21].

4 Distance to Calibration

Calibration measures we've seen are discontinuous. We now introduce a continuous calibration measure: the distance to calibration (DTC).

Definition 4.1 (Distance to calibration (DTC) [BGHN23])**.** *Let f be a predictor and $\mathcal{D}$ be a distribution. We define the DTC of f on $\mathcal{D}$ as*

$$\inf_{f'} \mathbb{E}_{\mathcal{D}}[|f(x) - f'(x)|],$$

where the infimum is over all calibrated predictors f' on $\mathcal{D}$.

Example 4.1. *Let $\mathcal{D}$ be the uniform distribution on $\{(x_0, 0), (x_1, 1)\}$. Let f_α be the predictor such that $f(x_0) = \alpha, f(x_1) = 1 - \alpha$. Then*

$$\mathsf{DTC}(f, \mathcal{D}) = \begin{cases} \alpha, & \text{if } \alpha \in [0, 1/4]; \\ |\alpha - 1/2|, & \text{if } \alpha \in [1/4, 1]. \end{cases}$$

DTC is continuous, but it is challenging to compute due to the infimum in the definition.

5 Smooth Calibration Error

We now introduce a calibration measure that is continuous *and* easy to compute:

Definition 5.1 ([KF08, FH18])**.** *Let f be a predictor and $\mathcal{D}$ be a distribution. Let $\mathcal{W}$ be the family of 1-Lipschitz functions $w : [0, 1] \to [-1, 1]$ (these functions are called* weight functions*). We define the smooth calibration error of f on $\mathcal{D}$ as*

$$\mathsf{smCE}(f, \mathcal{D}) := \sup_{w \in \mathcal{W}} \mathbb{E}[(f(x) - y)w(f(x))].$$

Theorem 5.1 ([BGHN23])**.** *For every $(f, \mathcal{D})$ pair,*

$$\frac{1}{32}\,\mathsf{DTC}(f, \mathcal{D})^2 \le \mathsf{smCE}(f, \mathcal{D}) \le 2\,\mathsf{DTC}(f, \mathcal{D}).$$

Fast algorithms exist for computing the smooth calibration error (see e.g. [HJTY24]).

6 Desired Properties of Calibration Measures

The following is a collection of desired properties that we would like a calibration measure CAL to satisfy. It is an active research question to satisfy as many of these properties as possible. Many of my recent papers aim to address this question.

Completeness. If f is calibrated on $\mathcal{D}$, then $\mathsf{CAL}(f, \mathcal{D}) = 0$.

Soundness. If f is mis-calibrated on $\mathcal{D}$, then $\mathsf{CAL}(f, \mathcal{D}) > 0$.

Consistency. There exist constants $c_1, c_2, \alpha_1, \alpha_2 > 0$ such that for every pair of $(f, \mathcal{D})$,

$$c_1\,\mathsf{DTC}(f, \mathcal{D})^{\alpha_1} \le \mathsf{CAL}(f, \mathcal{D}) \le c_2\,\mathsf{DTC}(f, \mathcal{D})^{\alpha_2}.$$

Consistency implies completeness and soundness for CAL, because DTC is complete and sound.

Continuity. $\mathsf{CAL}(f, \mathcal{D})$ is a continuous function of the predictor f. In other words, small changes in the predictions only cause a small change in the calibration error.

Testability. $\mathsf{CAL}(f, \mathcal{D})$ can be estimated on a small set of data points drawn from $\mathcal{D}$.

Efficiency. $\mathsf{CAL}(f, \mathcal{D})$ can be estimated using an efficient algorithm.

Actionability. $\mathsf{CAL}(f, \mathcal{D})$ is a valid upper bound on the *swap regret* of downstream decision makers. We will define this property more formally in the next lecture.

Truthfulness. The expected value of $\mathsf{CAL}(f, \mathcal{D})$ *on a finite data set* is minimized when $f(x) = \mathbb{E}_{\mathcal{D}}[y|x]$.

References

[BGHN23] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1727–1740, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3564246.3585182.

[FH18] Dean P. Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0899825618300113, doi:10.1016/j.geb.2017.12.022.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL: https://proceedings.mlr.press/v70/guo17a.html.

[HJTY24] Lunjia Hu, Arun Jambulapati, Kevin Tian, and Chutong Yang. Testing calibration in nearly-linear time. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 116022–116059. Curran Associates, Inc., 2024. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/d20e3c35bedb17a9f6f01fc434a30fa3-Paper-Conference.pdf, doi:10.52202/079017-3683.

[KF08] Sham M. Kakade and Dean P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008. Learning Theory 2004. URL: https://www.sciencedirect.com/science/article/pii/S0022000007000633, doi:10.1016/j.jcss.2007.04.017.

[MDR+21] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural

networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html.

[Ope23] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL: https://arxiv.org/abs/2303.08774, doi:10.48550/arXiv.2303.08774.