# Scalable Utility-Aware Multiclass Calibration

Mahmoud Hegazy  Michael I. Jordan  Aymeric Dieuleveut

CMAP, Ecole polytechnique | Inria Paris | UC Berkeley

AISTATS 2026

Michael I. Jordan

Aymeric Dieuleveut

# Calibration

### Full multiclass

$$\mathbb{E}[Y \mid f(X)] = f(X)$$

audits the full probability vector.

Very challenging to achieve.

Vaicenavicius et al. (2019); Lee et al. (2023)

### Top-class reduction

$$p_* = \max_c f_c(X), \quad \mathbb{E}[Y_* \mid p_*] = p_*$$

reduces calibration to top-class confidence.

Assumes the user cares about the top class.

Guo et al. (2017); Gupta & Ramdas (2022)

### Weighted calibration

$$\sup_{w \in \mathcal{W}} \mathbb{E}[\langle w(f(X)), Y - f(X) \rangle]$$

chosen witnesses encode downstream errors.

General notion via chosen witness functions.

Jung et al. (2021); Hebert-Johnson et al. (2018)

### Distance to calibration

$$\mathrm{DC}(f) = \inf_{g:\,\mathrm{cal.}} \mathbb{E}\|f(X) - g(X)\|$$

distance to the closest calibrated predictor.

Challenging to achieve; useful as a benchmark.

Blasiok et al. (2023); Gopalan et al. (2024)

# Utility Calibration

**True utility:** $u(f(X), Y) \in [-1, 1]$

**Score:**

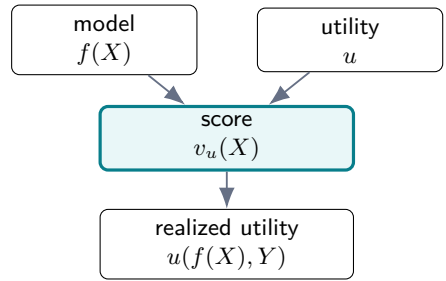
$$v_u(X) = \mathbb{E}_{\hat{Y} \sim f(X)} u(f(X), \hat{Y}) = \langle f(X), \vec{u}(X) \rangle .$$

- $Y$ is available, bootstrap it from $f$.
- Audit the bias of $v_u(X)$.

**Calibration error**

$$\mathrm{UC}(f, u) = \sup_{I \subset [-1,1]} \left| \mathbb{E}\left[ (u(f(X), Y) - v_u(X)) \, \mathbf{1}_{\{v_u(X) \in I\}} \right] \right|$$

**Audit worst-case interval of the scalar $v_u(X)$.**




**It recovers:**

top-class | class-wise | linear payoffs | rank / top-$K$

Jung et al. (2021); Rossellini et al. (2025).

# Audits

**Worst case over a class**

$$\mathrm{UC}(f,\mathcal{U}) = \sup_{u\in\mathcal{U}} \mathrm{UC}(f,u)$$

$$\mathrm{UC}(f,u) = \sup_{I\subset[-1,1]} \left|\mathbb{E}\left[(u(f(X),Y) - v_u(X))\mathbf{1}_{\{v_u(X)\in I\}}\right]\right|$$

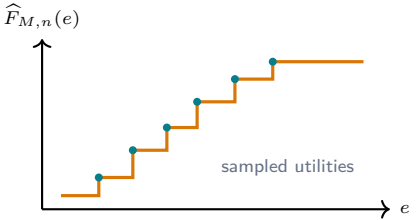
$$\sup_{u\in\mathcal{U}} \sup_{I} \left|\mathbb{E}\left[(u(f(X),Y) - v_u(X))\mathbf{1}_{\{v_u(X)\in I\}}\right]\right|$$

$\underbrace{\sup_{u\in\mathcal{U}}}$ class search can still be quite hard

**Distributional audit**

$$u_m \sim \Pi_{\mathcal{U}}, \qquad E_m = \widehat{\mathrm{UC}}(f,u_m)$$

$$\widehat{F}_{M,n}(e) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{E_m \le e\}$$

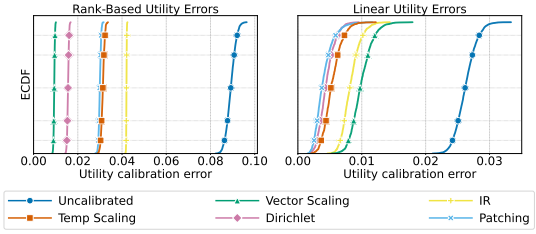



**sample $u$** **audit each** **plot eCDF**

# ImageNet

| Method | Brier $\times 10^2$ | $\text{CWE}_{\text{bin}} \times 10^4$ | $\text{TCE}_{\text{bin}} \times 10^3$ | $\mathcal{U}_{\text{comb}} \times 10^3$ |
|---|---|---|---|---|
| Uncal. | 22.6 | 2.46 | 94.2 | 124.0 |
| Dirichlet | **21.3** | 1.34 | 13.7 | 26.1 |
| Isotonic | 22.9 | **1.10** | 33.1 | 54.1 |
| Temp. | 21.8 | 1.26 | 30.0 | 45.2 |
| Vector | 22.8 | 1.54 | 35.1 | 37.4 |
| Patching | 21.6 | 1.56 | **10.3** | **19.4** |

$\arg\min \mathcal{U}_{\text{comb}} = \text{Patching}, \qquad \arg\min \text{Brier} = \text{Dirichlet}$

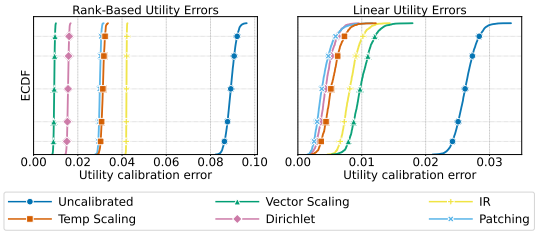



**Takeaways: utility choice changes the best predictors; distributional audits identify hidden trends.**

# ImageNet

| Method | Brier $\times 10^2$ | $\text{CWE}_{\text{bin}} \times 10^4$ | $\text{TCE}_{\text{bin}} \times 10^3$ | $\mathcal{U}_{\text{comb}} \times 10^3$ |
|---|---|---|---|---|
| Uncal. | 22.6 | 2.46 | 94.2 | 124.0 |
| Dirichlet | **21.3** | 1.34 | 13.7 | 26.1 |
| Isotonic | 22.9 | **1.10** | 33.1 | 54.1 |
| Temp. | 21.8 | 1.26 | 30.0 | 45.2 |
| Vector | 22.8 | 1.54 | 35.1 | 37.4 |
| Patching | 21.6 | 1.56 | **10.3** | **19.4** |

$\arg\min \mathcal{U}_{\text{comb}} = \text{Patching}, \qquad \arg\min \text{Brier} = \text{Dirichlet}$




**Takeaways: utility choice changes the best predictors; distributional audits identify hidden trends.**

Hot-take: we should move to a Bayesian way of reporting calibration.

# References


**Vaicenavicius et al. (2019)** Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. *Evaluating Model Calibration in Classification*. AISTATS 2019.

**Lee et al. (2023)** Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. *T-Cal: An Optimal Test for the Calibration of Predictive Models*. JMLR 2023.

**Guo et al. (2017)** Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. *On Calibration of Modern Neural Networks*. ICML 2017.

**Gupta & Ramdas (2022)** Chirag Gupta and Aaditya Ramdas. *Top-label Calibration and Multiclass-to-Binary Reductions*. ICLR 2022.

**Jung et al. (2021)** Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. *Moment Multicalibration for Uncertainty Estimation*. COLT 2021.

**Rossellini et al. (2025)** Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. *Can a Calibration Metric Be Both Testable and Actionable?*. arXiv 2025.

**Hebert-Johnson et al. (2018)** Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. *Multicalibration: Calibration for the (Computationally-Identifiable) Masses*. ICML 2018.

**Blasiok et al. (2023)** Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. *A Unifying Theory of Distance from Calibration*. STOC 2023.

**Gopalan et al. (2024)** Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. *On Computationally Efficient Multi-Class Calibration*. arXiv 2024.