U-Calibration: Forecasting for an Unknown Agent

Robert Kleinberg RDK@CS.CORNELL.EDU
Cornell University

Renato Paes Leme RENATOPPL@GOOGLE.COM
Google Research

Jon Schneider JSCHNEI@GOOGLE.COM
Google Research

Yifeng Teng YIFENGT@GOOGLE.COM
Google Research

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the problem of evaluating forecasts of binary events whose predictions are consumed by rational agents who take an action in response to a prediction, but whose utility is unknown to the forecaster. We show that optimizing forecasts for a single scoring rule (e.g., the Brier score) cannot guarantee low regret for all possible agents. In contrast, forecasts that are well-calibrated guarantee that all agents incur sublinear regret. However, calibration is not a necessary criterion here (it is possible for miscalibrated forecasts to provide good regret guarantees for all possible agents), and calibrated forecasting procedures have provably worse convergence rates than forecasting procedures targeting a single scoring rule.

Motivated by this, we present a new metric for evaluating forecasts that we call *U-calibration*, equal to the maximal regret of the sequence of forecasts when evaluated under any bounded scoring rule. We show that sublinear U-calibration error is a necessary and sufficient condition for all agents to achieve sublinear regret guarantees. We additionally demonstrate how to compute the U-calibration error efficiently and provide an online algorithm that achieves $O(\sqrt{T})$ U-calibration error (on par with optimal rates for optimizing for a single scoring rule, and bypassing lower bounds for the traditionally calibrated learning procedures). Finally, we discuss generalizations to the multiclass prediction setting.[1]

Keywords: Calibration, forecasting, scoring rules, online learning.

References

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

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*, pages 1727–1740, 2023.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

[1] Extended abstract. Full version appears as arXiv:2307.00168v1.

Avrim Blum, MohammadTaghi Hajiaghayi, Katrina Ligett, and Aaron Roth. Regret minimization and the price of total anarchy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 373–382, 2008.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982.

Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. *Advances in neural information processing systems*, 32, 2019.

Dean P Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games and Economic Behavior*, 109:271–293, 2018.

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

Dean P Foster and Sergiu Hart. " calibeating": beating forecasters at their own game. *arXiv preprint arXiv:2209.04892*, 2022.

Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40, 1997.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Parikshit Gopalan, Lunjia Hu, Michael P Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. *arXiv preprint arXiv:2210.08649*, 2022a.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022b.

Parikshit Gopalan, Michael P Kim, and Omer Reingold. Characterizing notions of omniprediction via multicalibration. *arXiv preprint arXiv:2302.06726*, 2023.

Sergiu Hart. Calibrated forecasts: The minimax proof. *arXiv preprint arXiv:2209.05863*, 2022.

Jason D Hartline, Liren Shan, Yingkai Li, and Yifan Wu. Optimal scoring rules for multidimensional effort. *arXiv preprint arXiv:2211.03302*, 2022.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Ulf Johansson, Tuwe Löfström, and Henrik Boström. Calibrating multi-class models. In *Conformal and Probabilistic Prediction and Applications*, pages 111–130. PMLR, 2021.

Sham M Kakade and Dean P Foster. Deterministic calibration and nash equilibrium. In *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings 17*, pages 33–48. Springer, 2004.

Yingkai Li, Jason D Hartline, Liren Shan, and Yifan Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 2022.

Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in bayesian games. In *Conference on Learning Theory*, pages 5221–5252. PMLR, 2022.

Eric Neyman, Georgy Noarov, and S Matthew Weinberg. Binary scoring rules that incentivize precision. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 718–733, 2021.

David Oakes. Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339, 1985.

Mingda Qiao and Gregory Valiant. Stronger calibration lower bounds via sidestepping. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 456–466, 2021.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Mark J Schervish. A general method for comparing probability assessors. *The annals of statistics*, 17(4):1856–1879, 1989.

Teddy Seidenfeld. Calibration, coherence, and scoring rules. *Philosophy of Science*, 52(2):274–294, 1985.