“Calibeating”: Beating Forecasters at Their Own Game*

Dean P. Foster† Sergiu Hart‡

November 1, 2024

Addendum:
Multi-Calibeating Beats the Stronger Expert

We show here that calibeating is a stronger notion than the so-called “stronger expert.”

As shown in the paper (see (2) and the last paragraph of Section 2), the refinement score is the minimal Brier score over all relabelings of the bins; i.e.,

$$\mathcal{R}_t = \min_{\phi} \mathcal{B}_t^{\phi(\mathbf{c})}, \tag{2}$$

where the minimum is taken over all functions $\phi : C \to C$ (from current labels to new labels), and we write $\mathcal{B}_t^{\phi(\mathbf{c})}$ for the Brier score where the sequence $\mathbf{c}$ is replaced by $\phi(\mathbf{c}) = (\phi(c_s))_{s=1,2,...}$.

Therefore, taking $C = \Delta(A)$, we have:

- $\mathbf{c}$ is *multi-calibeating* $\mathbf{b}_1, ..., \mathbf{b}_N$ if

  $$\mathcal{B}_t^{\mathbf{c}} \leq \min_{\phi} \mathcal{B}_t^{\phi(\mathbf{b}_1,...,\mathbf{b}_N)} + o(1)$$

  (as $t \to \infty$; for simplicity we ignore here the uniformity on $\mathbf{a}$), where the minimum is taken over all functions $\phi : \Pi_{n=1}^N B^n \to \Delta(A)$.

By comparison, when all forecasts are probability distributions on A, i.e., all B^n are subsets of $\Delta(A)$, Foster (1991) defines:[1]

---

*Published in: *Theoretical Economics* 18 (2003), 4, 1441-1474. doi.org/10.3982/TE5330

†Department of Statistics, Wharton, University of Pennsylvania, Philadelphia, and Amazon, New York. *e-mail*: `dean@foster.net` *web page*: `http://deanfoster.net`

‡Institute of Mathematics, Department of Economics, and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem. *e-mail*: `hart@huji.ac.il` *web page*: `http://www.ma.huji.ac.il/hart`

[1]In the expansive literature on experts, these notions are referred to as “prediction with no regret”; see Appendix A.9 of the full paper for the parallel results with the logarithmic scoring rule.

- $\mathbf{c}$ is *as strong as* $\mathbf{b}_1, ..., \mathbf{b}_N$ if

$$\mathcal{B}_t^{\mathbf{c}} \leq \min_{1 \leq n \leq N} \mathcal{B}_t^{\mathbf{b}_n} + o(1),$$

  and

- $\mathbf{c}$ is *as strong as the convex hull* of $\mathbf{b}_1, ..., \mathbf{b}_N$ if

$$\mathcal{B}_t^{\mathbf{c}} \leq \min_{w} \mathcal{B}_t^{w_1 \mathbf{b}_1 + ... w_N \mathbf{b}_N} + o(1),$$

  where the minimum is taken over all $w = (w_1, ..., w_N)$ with $w_n \geq 0$ and $\sum_{n=1}^{N} w_n = 1$.

Whereas multi-calibeating takes into account *all* bin relabelings, stronger-experts notions do not go beyond linear-combination relabelings. Therefore, as claimed,

Calibeating is stronger than being the "stronger expert."