Calibrated Forecasts: The Minimax Proof*

Sergiu Hart†

October 31, 2021

Abstract

A formal write-up of the simple proof (1995) of the existence of calibrated forecasts by the minimax theorem, which moreover shows that N^3 periods suffice to guarantee a $1/N$ calibration error.

Consider a weather forecaster who announces each day a probability p that there will be rain tomorrow. The forecaster is said to be *calibrated* if, for each forecast p that is used, the relative frequency of rainy days among the days where the forecast was p is, in the long run, (close to) p.

The surprising result of Foster and Vohra (1998) is that calibration can be *guaranteed*, no matter what the weather will be.[1] A simple proof of this result, based on the *minimax theorem*, was provided by the author in 1995.[2]

We provide here a formal write-up of this proof, which moreover shows that an expected calibration error of size ε is guaranteed after $1/\varepsilon^3$ periods.[3]

*First version: August 2018. Thanks to Jérôme Renault for asking about the relation between the calibration error and the number of periods, and to Benjy Weiss for providing inequality (2).

†Institute of Mathematics, Department of Economics, and 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]There is a large literature on calibration and its uses; see, e.g., the survey of Olszewski (2015) and the recent paper of Foster and Hart (2021).

[2]At a lecture given by Dean Foster at the Center for Rationality of the Hebrew University of Jerusalem; see Section 4, "An argument of Sergiu Hart," in Foster and Vohra (1998).

[3]Thus answering a question of Jérôme Renault.

For each period (day) $t = 1, 2, ...$, let $a_t \in \{0, 1\}$ be the *weather*, with 1 for rain and 0 for no rain, and let $c_t \in [0, 1]$ be the *forecast*. We will let our forecasts lie on the grid $D := \{1/(2N), 3/(2N), ..., (2N-1)/(2N)\}$ for some positive integer N (each point in $[0, 1]$ is thus within a distance of at most $1/(2N)$ from a point in D).

The *calibration score* K_T at time T is computed as follows. For each $d \in D$ let[4] $n(d) \equiv n_T(d) := \sum_{t=1}^{T} \mathbf{1}_{c_t=d}$ be the number of periods in which the forecast was d, and let $\overline{a}(d) \equiv \overline{a}_T(d) := (1/n(d)) \sum_{t=1}^{T} \mathbf{1}_{c_t=d}\, a_t$ be the (relative) frequency of rain in those $n(d)$ periods; the calibration score K_T is then the average distance between forecasts and rain frequencies, namely,[5]

$$K_T := \sum_{d \in D} \left(\frac{n(d)}{T} \right) |\overline{a}(d) - d| \,.$$

This setup can be viewed as a finite T-period game in which in every period $t = 1, ..., T$ the "rainmaker" chooses the weather $a_t \in \{0, 1\}$ and the "forecaster" chooses the forecast $c_t \in D$, and the payoff is the calibration score K_T. Both players are assumed to have perfect recall of past weather and forecasts (thus allowing for an "adversarial" rainmaker); since the number of periods T and the sets of choices of the players, $\{0, 1\}$ and D, are all finite, the game is a finite game (i.e., each player has finitely many pure strategies).

Theorem 1 *Let $T \geq N^3$. Then there exists a mixed strategy of the forecaster such that*[6] $\mathbb{E}[K_T] \leq 1/N$ *is guaranteed against any mixed strategy of the rainmaker.*

The result follows from Proposition 2 below by applying von Neumann's (1928) minimax theorem for two-person finite games. The minimax theorem says that if, for every mixed strategy of player 2, player 1 has a reply that

[4] We write $\mathbf{1}_X$ for the indicator of the event X; thus, $\mathbf{1}_{c_t=d}$ equals 1 if $c_t = d$ and 0 otherwise.

[5] An alternative score averages the squared errors: $\mathcal{K}_T := \sum_{d \in D} (n(d)/T)(\overline{a}(d) - d)^2$. The two scores are essentially equivalent, because $(K_T)^2 \leq \mathcal{K}_T \leq K_T$ (the first inequality is by Jensen's inequality, and the second is by $|\overline{a}(d) - d| \leq 1$, since $\overline{a}(d)$ and d are both in $[0, 1]$).

[6] The expectation is over the random choices of the two players.

yields an expected payoff of at least v, then player 1 has a mixed strategy that guarantees (no matter what player 2 does) an expected payoff of at least v.

Proposition 2 *Let $T \geq N^3$. Then for every mixed strategy of the rainmaker there is a strategy of the forecaster such that $\mathbb{E}[K_T] \leq 1/N$.*

Proof. Let τ be a mixed strategy of the rainmaker. For every $t \geq 1$ and history $h_{t-1} = (a_1, c_1, ..., a_{t-1}, c_{t-1}) \in (\{0,1\} \times D)^{t-1}$ of rain and forecasts before time t, let $p_t := \mathbb{P}[a_t = 1 | h_{t-1}] = \mathbb{E}[a_t | h_{t-1}]$ be the probability of rain induced by the rainmaker's strategy τ. We then let the forecast c_t after the history h_{t-1} be the rounding of p_t to the grid D, with a fixed tie-breaking rule when p_t is the midpoint of two consecutive points in D; this makes c_t a deterministic function of the history—i.e., c_t is h_{t-1}-measurable—and we always have $|c_t - p_t| \leq 1/(2N)$.

The calibration score K_T can be written as

$$K_T = \frac{1}{T} \sum_{d \in D} |G(d)| ,$$

where[7]

$$G(d) := n(d)(\overline{a}(d) - d) = \sum_{t=1}^{T} \mathbf{1}_{c_t = d}(a_t - d) = \sum_{t=1}^{T} \mathbf{1}_{c_t = d}(a_t - c_t)$$

for every $d \in D$. Replacing each c_t with p_t yields the scores

$$\begin{aligned} \widetilde{G}(d) &:= \sum_{t=1}^{T} \mathbf{1}_{c_t = d}(a_t - p_t) \quad \text{and} \\ \widetilde{K}_T &:= \frac{1}{T} \sum_{d \in D} \left| \widetilde{G}(d) \right| ; \end{aligned}$$

[7] $G(d)$ is the difference between the actual number of rainy days, $n(d)\overline{a}(d)$, and the predicted number of rainy days, $n(d)d$, in the $n(d)$ days in which the forecast was d; it is referred to as the (total) "gap" in Foster and Hart (2021).

since $|c_t - p_t| \le 1/(2N)$ it follows that $|G(d) - \widetilde{G}(d)| \le n(d)/(2N)$ and

$$\left|K_T - \widetilde{K}_T\right| \le \frac{1}{T}\sum_{d\in D}\frac{n(d)}{2N} = \frac{1}{2N} \tag{1}$$

(because $\sum_d n(d) = T$).

We claim that[8]

$$\mathbb{E}\left[\widetilde{G}(d)^2\right] \le \frac{1}{4}\mathbb{E}\left[n(d)\right] \tag{2}$$

for each $d \in D$. Indeed, $\widetilde{G}(d) = \sum_{t=1}^{T} \mathbf{1}_{c_t=d} Z_t$ where $Z_t := a_t - p_t$, for which we have $\mathbb{E}\left[Z_t|h_{t-1}\right] = 0$ (because $p_t = \mathbb{E}\left[a_t|h_{t-1}\right]$) and $\mathbb{E}\left[Z_t^2|h_{t-1}\right] \le 1/4$ (because this is the variance of a Bernoulli random variable, namely, $a_t|h_{t-1}$). Then, for $s < t$ we get

$$\begin{aligned}\mathbb{E}\left[(\mathbf{1}_{c_s=d}\, Z_s)\cdot(\mathbf{1}_{c_t=d}\, Z_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[(\mathbf{1}_{c_s=d}\, Z_s)\cdot(\mathbf{1}_{c_t=d}\, Z_t)|h_{t-1}\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}_{c_s=d}\, Z_s \mathbf{1}_{c_t=d}\, \mathbb{E}\left[Z_t|h_{t-1}\right]\right] = 0\end{aligned}$$

(because the random variables c_s, Z_s, and c_t are h_{t-1}-measurable), and for $s = t$ we get

$$\begin{aligned}\mathbb{E}\left[(\mathbf{1}_{c_t=d}\, Z_t)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(\mathbf{1}_{c_t=d}\, Z_t)^2\,|h_{t-1}\right]\right] \\ &= \mathbb{E}\left[\mathbf{1}_{c_t=d}\,\mathbb{E}\left[Z_t^2|h_{t-1}\right]\right] \le \frac{1}{4}\mathbb{E}\left[\mathbf{1}_{c_t=d}\right];\end{aligned}$$

summing all these terms yields $\mathbb{E}\left[\widetilde{G}(d)^2\right] \le (1/4)\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{1}_{c_t=d}\right] = (1/4)\mathbb{E}\left[n(d)\right]$, which is (2).

Therefore,

$$\begin{aligned}\mathbb{E}\left[\widetilde{K}_T\right] &= \frac{1}{T}\sum_{d\in D}\mathbb{E}\left[\left|\widetilde{G}(d)\right|\right] \le \frac{1}{T}\frac{1}{2}\sum_{d\in D}\left(\mathbb{E}\left[n(d)\right]\right)^{1/2} \\ &\le \frac{1}{T}\frac{1}{2}\left(N\sum_{d\in D}\mathbb{E}\left[n(d)\right]\right)^{1/2} = \frac{1}{2}\left(\frac{N}{T}\right)^{1/2},\end{aligned} \tag{3}$$

[8] If one does not care about the bound N^3 on T one may use at this point various simpler Chebyshev or law-of-large-numbers inequalities (see also Remarks (c) and (d) below).

where we have used $\mathbb{E}\left[\left|\widetilde{G}(d)\right|\right] \leq \left(\mathbb{E}\left[\widetilde{G}(d)^2\right]\right)^{1/2}$ and (2) for the first inequality, the Cauchy–Schwartz inequality for the second one, and finally $\sum_d \mathbb{E}\left[n(d)\right] = T$. When $T \geq N^3$ this gives $\mathbb{E}\left[\widetilde{K}_T\right] \leq 1/(2N)$, and hence $\mathbb{E}\left[K_T\right] \leq 1/(2N) + 1/(2N) = 1/N$ by (1). ■

Remarks. *(a)* Since the game between the rainmaker and the forecaster is a game of perfect recall, by Kuhn's (1953) theorem one can replace mixed strategies with their equivalent *behavior* strategies. A behavior strategy of the forecaster, which is referred to as a *forecasting procedure*, consists of a separate randomization after each history; i.e., it is a mapping from the set of histories to the set of probability distributions on D.

(b) N^3 is the right order of magnitude for the horizon T that guarantees a calibration error of $1/N$ when the forecaster rounds the rain probabilities p_t to the grid D, because if the rainmaker chooses p_t to be uniform on $[0,1]$ then each one of the N forecasts d in D is used about T/N times, and so in order to get an error of $1/N$ one needs T/N to be of the order of N^2.

(c) A tighter estimation in the proof of Proposition 2 uses $\mathbb{E}\left[Z_t^2|h_{t-1}\right] = p_t(1-p_t)$, which is close to $d(1-d)$, instead of $\mathbb{E}\left[Z_t^2|h_{t-1}\right] \leq 1/4$ (recall that $\mathbb{E}\left[Z_t^2|h_{t-1}\right]$ is the variance of a Bernoulli(p_t) random variable); this yields $\mathbb{E}\left[K_T\right] \leq 1/N$ for T starting approximately at $(2/3)N^3$. More precisely: let $f(d) := d'(1-d')$ where $d' = d + 1/(2N)$ for $d < 1/2$, $d' = d$ for $d = 1/2$, and $d' = d - 1/(2N)$ for $d > 1/2$; then $|p_t - d| \leq 1/(2N)$ implies $p_t(1-p_t) \leq f(d)$ (because $x(1-x)$ increases for $x < 1/2$ and decreases for $x > 1/2$), and then the coefficient $1/4$ in inequality (2) may be replaced with $f(d)$. This yields

$$\begin{aligned}\mathbb{E}\left[\widetilde{K}_T\right] &\leq \frac{1}{T}\sum_{d\in D}\left(f(d)\mathbb{E}\left[n(d)\right]\right)^{1/2} \leq \frac{1}{T}\left(\sum_{d\in D} f(d)\right)^{1/2}\left(\sum_{d\in D}\mathbb{E}\left[n(d)\right]\right)^{1/2} \\ &= \frac{1}{T^{1/2}}\left(\sum_{d\in D} f(d)\right)^{1/2}.\end{aligned}$$

Now it is a straightforward computation to see that $\sum_{d\in D} f(d) = N/6 + 1/4 - 1/(6N)$, and so for all $T \geq (2/3)N^3 + N^2 - (2/3)N$ we have $\mathbb{E}\left[\widetilde{K}_T\right] \leq 1/(2N)$, and thus $\mathbb{E}\left[K_T\right] \leq 1/N$.

(d) A looser but slightly simpler estimation in the proof of Proposition 2, using $n(d) \leq T$ for each d instead of $\sum_d n(d) = T$, gives $\mathbb{E}\left[\widetilde{K}_T\right] \leq (1/T)(1/2)NT^{1/2}$, and then $\mathbb{E}[K_T] \leq 1/N$ for $T \geq N^4$.

(e) If instead of D we would use the grid $D' = \{0, 1/N, 2/N, ..., 1\}$, we would need to replace N (the size of D) with $N+1$ (the size of D') in (3), which would yield $\mathbb{E}[K_T] \leq 1/N$ for $T \geq (N+1)N^2 = N^3 + N^2$.

(f) A *lower* bound on the guaranteed calibration error as a function of the number of periods T has recently been obtained by Qiao and Valiant (2021); it is of the order of $T^{-0.472}$ (improving on the trivial lower bound of the order of $T^{-1/2}$, which is obtained when the rain is an i.i.d. Bernoulli(1/2) process; note that what we showed here is an upper bound of $T^{-1/3}$).

References

Foster, D. P. and S. Hart (2021), "Forecast Hedging and Calibration," *Journal of Political Economy* 129 (forthcoming), doi.org/10.1086/716559. `http://www.ma.huji.ac.il/hart/publ.html#calib-int`

Foster, D. P. and R. V. Vohra (1998), "Asymptotic Calibration," *Biometrika* 85, 379–390.

Kuhn, H. W. (1953), "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games, Vol. II*, H. W. Kuhn and A. W. Tucker (editors), *Annals of Mathematics Studies* 28, Princeton University Press, 193–216.

Olszewski, W. (2015), "Calibration and Expert Testing," in *Handbook of Game Theory, Vol. 4*, H. P. Young and S. Zamir (editors), Springer, 949–984.

Qiao, M. and G. Valiant (2021), "Stronger Calibration Lower Bounds via Sidestepping," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21)*, doi.org/10.1145/3406325.3451050.

von Neumann, J. (1928), "Zur Theorie der Gesellschaftsspiele, " *Mathematische Annalen* 100, 295–320.