Calibrated Forecasts: The Minimax Proof*

Sergiu Hart†

May 30, 2023

Abstract

We provide 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 calibration error of at most $1/N$.

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 out of those days in which the forecast was p is equal to p in the long run.

The surprising result of Foster and Vohra (1998) is that calibration can be *guaranteed*, no matter what the weather will be. There are various proofs of this result, and there is a large literature on calibration and its uses; see the survey of Olszewski (2015) and the more recent paper of Foster and Hart (2021).

A simple proof of the existence of calibrated forecasts, based on the *minimax theorem*, was provided by the author in 1995.[1] The basic argument

*Dedicated with great admiration to David Gale, in commemoration of his 100th birthday. Previous versions: August 2018; October 2021. The author thanks Jérôme Renault for asking about the relation between the calibration error and the number of periods, and Benjy Weiss for providing inequality (2).

†Einstein Institute of Mathematics, Bogen 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] 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).

is as follows (see below for details). If the forecaster knew the strategy of the "rainmaker" (which could well be a mixed, i.e., probabilistic, strategy), then the forecaster could clearly get calibrated forecasts by announcing in every period the corresponding known probability of rain. Incorporating this into a finite game (by using a finite grid of forecasts and a finite horizon) yields, by von Neumann's (1928) minimax theorem for two-person zero-sum finite games, the existence of a strategy of the forecaster that guarantees calibration against *any* strategy of the rainmaker. This is a striking use of the minimax theorem, since the fact that there is a calibrated *reply* to any given strategy of the rainmaker is clear, whereas the consequence that there is a *single* strategy that is calibrated against *all* strategies of the rainmaker comes as a big surprise.[2]

More formally, consider a two-person finite game where player 1 has m strategies, player 2 has n strategies, and u_{ij} is the payoff when player 1 plays his i-th strategy and player 2 plays his j-th strategy.[3] A mixed strategy x of player 1 is a probability distribution over his pure strategies $\{1, ..., m\}$, i.e., $x = (x_1, ..., x_m)$, where $x_i \geq 0$ for every $i = 1, ..., m$ and $\sum_{i=1}^{m} x_i = 1$; similarly, a mixed strategy y of player 2 is a probability distributions over his pure strategies $\{1, ..., n\}$, i.e., $y = (y_1, ..., y_n)$, where $y_j \geq 0$ for every $j = 1, ..., n$ and $\sum_{j=1}^{n} y_j = 1$. When the two players play the mixed strategies x and y, respectively, the (expected) payoff is $U(x, y) := \sum_{i=1}^{m} \sum_{j=1}^{n} x_i y_j u_{ij}$.

We now state the minimax theorem, formulated in a useful but perhaps less standard way.

Theorem 1 (Minimax) *Assume that the real number v satisfies the following:*

(i) *for every mixed strategy y of player 2 there is a mixed strategy $x \equiv x(y)$ of player 1 such that the payoff is at least v (i.e., $U(x(y), y) \geq v$).*

Then

[2] Indeed, Foster and Vohra had a hard time getting their paper published: they got many desk rejections saying that the result "cannot be true" (the technical report came out in 1991, and the published paper only seven years later).

[3] It does not matter who gets this "payoff" (it could be, say, player 1's payoff); also, the game need not be a zero-sum game, as only one payoff function is considered.

(ii) *there is a mixed strategy x^* of player 1 that guarantees that the payoff is at least v (i.e., $U(x^*, y) \geq v$ for every mixed strategy y of player 2).*

Indeed, the premise (i) says that $\max_x U(x, y) \geq v$ for every y, i.e., $\min_y \max_x U(x, y) \geq v$; since $\max_x \min_y U(x, y) = \min_y \max_x U(x, y)$ by von Neumann's (1928) minimax theorem, we get $\max_x \min_y U(x, y) \geq v$, and so, taking x^* to be a maximizer there, $\min_y U(x^*, y) \geq v$, which is the conclusion (ii).[4]

Stated this way, the minimax theorem may look surprising, since from a premise of "for every y there is an x" it gets a conclusion of "there is an x such that for every y," a false logical argument in general (while "every child has a mother" is true, "there is a mother of all children" is not). Nevertheless, the result is correct (and far from trivial) under the assumptions that, first, there are finitely many pure strategies, and second, one uses mixed strategies (the result is easily seen to be false if either one of these assumptions fails[5]).

In the above calibration setup, the premise (i) is that for every strategy of the rainmaker there is a strategy of the forecaster that yields a small calibration score,[6] and the conclusion (ii) is that there is a strategy of the forecaster that yields a small calibration score for every strategy of the rainmaker (apply the minimax theorem, taking as payoff the negative of the calibration score). Let us show how to get a calibration score of, say, 10%. To see that the premise (i) holds, assume that the strategy of the rainmaker is given. We will round each forecast to a multiple of 10% (the finite grid of forecasts is thus 0%, 10%, ..., 100%); therefore, the forecast of, say, 70%, is announced when the probability of rain is between 65% and 75%. Assume that this has occurred on a large number of days so that by the law of large numbers (i.e., Chebyshev's inequality) the expected error between expectation and realization of at most 5%; the relative frequency of rain out of these days will then

[4] Since (ii) trivially implies (i), the two conditions (i) and (ii) are in fact equivalent. Also, the premise (i) is easily seen to be equivalent to "for every mixed strategy y of player 2 there is a *pure* strategy $i \equiv i(y)$ of player 1 such that $U(i(y), y) \geq v$."

[5] Consider the "choose the higher integer" infinite game, and the "matching pennies" game with pure strategies only.

[6] The "calibration score" will be formally defined below, as the average distance between forecasts and relative frequencies (and so being calibrated means that the calibration score is equal to zero).

be between $65\% - 5\% = 60\%$ and $75\% + 5\% = 80\%$—i.e., with a calibration error of at most 10%. Since the same holds for every forecast (that is used nonnegligibly often), taking an appropriately large horizon proves the premise (i)—and thus the conclusion (ii).

We now provide 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.

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*. For convenience, we will let our forecasts lie in the grid $D := \{1/(2N), 3/(2N), ..., (2N-1)/(2N)\}$ for some positive integer N; thus, each point in $[0, 1]$ is within a distance of at most $1/(2N)$ from a point in D (see Remark (e) below for the standard $1/N$-grid).

The *calibration score* K_T at time T is computed as follows. For each $d \in D$ let[7]

$$n(d) \equiv n_T(d) \; := \; \sum_{t=1}^{T} \mathbf{1}_{c_t=d}$$

be the number of periods in which the forecast was d, and let

$$\overline{a}(d) \equiv \overline{a}_T(d) \; := \; \frac{1}{n(d)} \sum_{t=1}^{T} \mathbf{1}_{c_t=d}\, a_t$$

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,[8]

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

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

[7] 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.

[8] 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 $[0, 1]$).

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 2 (Calibration) *Let $T \geq N^3$. Then there exists a mixed strategy of the forecaster that guarantees that*[9] $\mathbb{E}\left[K_T\right] \leq 1/N$ *against any mixed strategy of the rainmaker.*

This follows from the proposition below by applying the minimax theorem to the payoff function $-K_T$.

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

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 $p_t := \mathbb{P}\left[a_t = 1 | h_{t-1}\right] = \mathbb{E}\left[a_t | h_{t-1}\right]$ 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 expressed as

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

where[10]

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

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

[10] $G(d)$ 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).

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

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

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[11]

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

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).

[11] 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).

Therefore,

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

where we have used $\mathbb{E}\left[\left|\widetilde{G}(d)\right|\right] \le \left(\mathbb{E}\left[\widetilde{G}(d)^2\right]\right)^{1/2}$ 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 $T \ge N^3$ this gives $\mathbb{E}\left[\widetilde{K}_T\right] \le 1/(2N)$, and hence $\mathbb{E}\left[K_T\right] \le 1/(2N) + 1/(2N) = 1/N$ 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 3 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 $\mathbb{E}\left[Z_t^2 | h_{t-1}\right] \le 1/4$ (recall that $\mathbb{E}\left[Z_t^2 | h_{t-1}\right]$ is the variance of a Bernoulli(p_t) random variable); this yields $\mathbb{E}\left[K_T\right] \le 1/N$ 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 $|p_t - d| \le 1/(2N)$ implies $p_t(1-p_t) \le f(d)$ (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 3 that uses $n(d) \leq T$ for each d instead of $\sum_d n(d) = T$ yields $\mathbb{E}\left[\widetilde{K}_T\right] \leq (1/T)(1/2)NT^{1/2}$, and so $\mathbb{E}\left[K_T\right] \leq 1/N$ for $T \geq N^4$.

(e) If instead of D we were to use the standard $1/N$-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 $\mathbb{E}\left[K_T\right] \leq 1/N$ 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 have shown here is an upper bound of $T^{-1/3}$).

(g) The minimax approach can be further used to obtain calibrated forecasts that are "calibeating," a concept introduced by Foster and Hart (2022): they are guaranteed to beat the Brier score of any other forecast by that forecast's calibration score. See Appendix A.2 of the `arxiv` version of Foster and Hart (2022).

(h) The minimax proof does not construct a calibrated procedure; it only shows its existence. There are various specific such constructions in the literature, the simplest being the one in Section V of Foster and Hart (2021).

References

Foster, D. P. and S. Hart (2021), "Forecast Hedging and Calibration," *Journal of Political Economy* 129, 3447–3490.

Foster, D. P. and S. Hart (2022), " 'Calibeating': Beating Forecasters at Their Own Game," `http://arxiv.org/abs/2209.04892v2`; *Theoretical Economics* (forthcoming).

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.