Statistical Calibration: a simplification of Foster's proof

Andrés Carvajal [*]

July 24, 2006

Consider the following problem: at each date in the future, a given event may or may not occur, and you will be asked to forecast, at each date, the probability that the event will occur in the next date. Unless you make degenerate forecasts (zero or one), the fact that the event does or does not occur does not prove your forecast wrong. But, in the long run, if your forecasts are accurate, the conditional relative frequencies of occurrence of the event should approach your forecast.

[4] has presented an algorithm that, whatever the sequence of realizations of the event, will meet the long-run accuracy criterion, even though it is completely ignorant about the real probabilities of occurrence of the event, or about the reasons why the event occurs or fails to occur. It is an adaptive algorithm, that reacts to the history of forecasts and occurrences, but does not learn from the history anything about the future: indeed, the past need not say anything about the future realizations of the event. The algorithm only looks at its own past inaccuracies and tries to make up for them in the future. The amazing result is that this (making up for past inaccuracies) can be done with arbitrarily high probability!

Alternative arguments for this result have been proposed in the literature, remarkably by [3], where a very simple algorithm has been proved to work, using a classical result in game theory: Blackwell's approachability result, [1]. Very recently, [2] has especialized Blackwell's theorem in a way that (under a minor modification of the algorithm) simplifies the argument of [3]. Here I present such modification and argument.

1 Preliminaries

At each future date $t \in \mathbb{N}$, an event may occur ($x_t = 1$) or not ($x_t = 0$). For each t, a forecast is a number $p_t \in [0,1]$ representing the probability that, one suggests, the event will occur at t. It is assumed that the forecast is made

[*] Department of Economics, Royal Holloway College, University of London; andres.carvajal@rhul.ac.uk. I thank Alvaro Riascos for comments that improved the presentation of the argument.

after observing $(x_q)_{q=1}^{t-1}$, and that only a subset of forecasts are acceptable.To formalize this, fix $M \in \mathbb{N}$, and denote $\mathcal{M} := \{1, ..., M\}$. For each $m \in \mathcal{M}$, define $I(m) := \left[\frac{m-1}{M}, \frac{m}{M}\right]$, and $p(m) := \frac{2m-1}{2M}$. Notice that $\bigcup_{m=1}^{M} I(M) = [0, 1]$ and that $p(m)$ is the middle point of $I(m)$. It is assumed that the forecast is restricted to be an element of the set $\{p(m)\}_{m\in\mathcal{M}}$. Since p defines a one-to-one correspondence, I will refer to m also as the forecast $p(m)$.

For each $T \in \mathbb{N}$, denote $H_T := (\mathcal{M} \times \{0, 1\})^T$, with generic element $h := (m_t, x_t)_{t=1}^{T}$, where m_t represents the forecast made for t. For simplicity, adopt the convention that $H_0 := \{(1, 0)\}$.

In the long-run, a good forecast should have the property that if $p(m)$ has been forecast infinitely many times, then the relative frequency of occurrence conditional on $p(m)$ having been forecast should approach $p(m)$, and, in particular, should lie in $I(m)$.

Define, for each $m \in \mathcal{M}$, $\rho_T^m : H_T \to [0, 1]$, $d_T^m : H_T \to \mathbb{R}$, and $e_T^m : H_T \to \mathbb{R}$ as

$$\rho_T^m(h) := \begin{cases} \frac{\sum_{t=1}^{T} x_t I(m_t=m)}{\sum_{t=1}^{T} I(m_t=m)}, & \text{if } \sum_{t=1}^{T} I(m_t = m) > 0; \\ \\ p(m), & \text{otherwise.} \end{cases}$$

$$d_T^m(h) := \left(\frac{m-1}{M} - \rho_T^m(h)\right) \sum_{t=1}^{T} \frac{I(m_t = m)}{T},$$

and

$$e_T^m(h) := \left(\rho_T^m(h) - \frac{m}{M}\right) \sum_{t=1}^{T} \frac{I(m_t = m)}{T}.$$

Define also $C_T : H_T \to \mathbb{R}_+$ as

$$C_T(h) := \sum_{m=1}^{M} \left(d_T^m(h)^+ + e_T^m(h)^+\right).$$

It is straightforward that $\rho_T^m(h) \in I(m)$ iff $d_T^m(h) \leq 0$ and $e_T^m(h) \leq 0$, so $(C_T)_{T=1}^{\infty}$ is a good measure of how inaccurate the performance of a sequence of forecasts is along different paths.

Also, notice that $d_T^m(h) \geq 0$ implies $e_T^m(h) < 0$, and $e_T^m(h) \geq 0$ implies $d_T^m(h) < 0$.

Lemma 1 (Foster)**.** *Fix $T \in \mathbb{N}$ and $h \in H_T$ such that for all $m \in \mathcal{M}$, $\rho_T^m(h) \notin I(m)$. Then, there exists $m \in \mathcal{M}$ such that $d_T^m(h) > 0$ and $e_T^{m-1}(h) > 0$*

Proof. By assumption, $\forall m \in \mathcal{M}$, either $d_T^m(h) > 0$ or $e_T^m(h) > 0$. By construction, $d_T^1(h) \leq 0$ and $e_T^M(h) \leq 0$, so $e_T^1(h) > 0$ and $d_T^M(h) > 0$. If $d_T^2(h) > 0$, we are done. Otherwise, $d_T^2(h) \leq 0$ and, hence $e_T^2(h) > 0$, and we can follow the search. The result follows since $M \in \mathbb{N}$: at the latest, $d_T^{M-1}(h) \leq 0$, so $e_T^{M-1}(h) > 0$, which suffices since $d_T^M(h) > 0$. □

2 Randomized forecasts and calibration

One does not need to impose structure on how the sequence $\mathtt{x} := (x_t)_{t=1}^{\infty}$ is determined, with the only exception that it is assumed that x_t cannot be determined as a function of m_t. This is so, because the choice of the forecast is allowed to be made randomly.

Let Δ denote the $(M-1)$-dimensional unit simplex.

A (randomized) **forecast** (or learning algorithm) is a sequence

$$\mathcal{L} := \left(L_t : H^{T-1} \to \Delta\right)_{t=1}^{\infty}$$

That is, given a history $h \in H_{t-1}$, $\mathcal{L}$ induces a random variable on $\mathcal{M}$, with distribution $L_t(h)$.

Given a forecast $\mathcal{L}$ and a sequence $\mathtt{x} := (x_t)_{t=1}^{\infty} \in \{0,1\}^{\infty}$, let $\mathrm{P}_{\mathcal{L},\mathtt{x}}$ denote the probability measure induced on $\mathcal{M}^{\infty}$ (see the Appendix).

A forecast $\mathcal{L}$ is (asymptotically) **calibrated** if for every $\epsilon > 0$, there exists $T_\epsilon \in \mathbb{N}$ such that, for any $\mathtt{x} \in \{0,1\}^{\infty}$,

$$\mathrm{P}_{\mathcal{L},\mathtt{x}}\left(\{h \in \mathcal{M}^{\infty} : \exists T \geq T_\epsilon : C_T\left((m_t, x_t)_{t=1}^{T}\right) \geq \epsilon\}\right) < \epsilon.$$

3 A calibrated forecast

The following forecast is a very minor modification of the one presented by [3]: define $\mathcal{L}$ as follows: for $T \in \mathbb{N}$, given $h \in H_{T-1}$,

1. If there exists $\bar{m} \in \mathcal{M}$ such that $\rho_{T-1}^{\bar{m}}(h) \in I(\bar{m})$, then

$$L_T(h)(m) := \begin{cases} 1, \text{ if } m = \bar{m}; \\ \\ 0, \text{ otherwise.} \end{cases}$$

2. Otherwise, find $\bar{m} \in \mathcal{M}$ such that $d_{T-1}^{\bar{m}}(h) > 0$ and $e_{T-1}^{\bar{m}-1}(h) > 0$, and let

$$L_T(h)(m) := \begin{cases} \frac{e_{T-1}^{m-1}(h)}{d_{T-1}^{m}(h)+e_{T-1}^{m-1}(h)}, \text{ if } m = \bar{m}; \\ \\ \frac{d_{T-1}^{m+1}(h)}{d_{T-1}^{m+1}(h)+e_{T-1}^{m}(h)}, \text{ if } m = \bar{m}-1; \\ \\ 0, \text{ otherwise.} \end{cases}$$

It follows from the lemma that $\mathcal{L}$ is well defined.

The forecast is different from the one presented by [3] in that, in case 2, for the same $\bar{m}$, it randomizes between $\bar{m}$ and $\bar{m}-1$, with probabilities proportional to $e_{T-1}^{\bar{m}-1}(h)$ and $d_{T-1}^{\bar{m}}(h)$, respectively, while the forecast of [3] randomizes between $\bar{m}$ and $\bar{m}+1$ with probabilities proportional to $d_{T-1}^{\bar{m}}(h)$ and $e_{T-1}^{\bar{m}-1}(h)$, respectively. While this difference is subtle, it is not, I think, trivial.

Following [3], but using [2] instead of [1], I now show that $\mathcal{L}$ is calibrated:

Theorem 1 (Foster and Vohra)**.** *$\mathcal{L}$ is calibrated*

Proof. As in [1] and [2], define a vector-valued game $\Gamma = (A, A', V, \gamma)$ as follows: the set of actions of a player, A, is finite; A', which is the set of actions of the opponent(s), is arbitrary; set V is a vector space over $\mathbb{R}$, endowed with an inner product and the outcome function is $\gamma : A \times A' \to V$.

Now, consider the infinite, sequential repetition of the game defined by $A := \mathcal{M}$, $A' := \{0, 1\}$, $V := \mathbb{R}^{2M}$, and, as in [3], γ defined as: for each $l \in \{1, ..., 2M\}$,

$$\gamma_l(m, x) := \begin{cases} \frac{m-1}{M} - x, & \text{if } m = l; \\ x - \frac{m}{M}, & \text{if } m = l - M; \\ 0, & \text{otherwise.} \end{cases}$$

It is obvious that γ is bounded, and it follows, by construction, that for every $l \in \{1, ..., M\}$,

$$\begin{aligned} \sum_{t=1}^{T} \gamma_l(m_t, x_t) &= \sum_{t=1}^{T} I(x_t = 0) I(m_t = l) \left(\frac{l-1}{M} \right) \\ &\quad + \sum_{t=1}^{T} I(x_t = 1) I(m_t = l) \left(\frac{l-1}{M} - 1 \right) \\ &= \sum_{t=1}^{T} I(m_t = l) \left(\frac{l-1}{M} \right) - \sum_{t=1}^{T} x_t I(m_t = l) \end{aligned}$$

so, if $\sum_{t=1}^{T} I(m_t = l) = 0$ then $\sum_{t=1}^{T} \gamma_l(m_t, x_t) = 0$, while if $\sum_{t=1}^{T} I(m_t = l) \neq 0$ then

$$\begin{aligned} \sum_{t=1}^{T} \gamma_l(m_t, x_t) &= \sum_{t=1}^{T} I(m_t = l) \left(\frac{l-1}{M} - \rho_T^l(h) \right) \\ &= d_T^l(h) T. \end{aligned}$$

Similarly, for every $l \in \{M + 1, ..., 2M\}$, if $\sum_{t=1}^{T} I(m_t = l - M) = 0$, then $\sum_{t=1}^{T} \gamma_l(m_t, x_t) = 0$, while if $\sum_{t=1}^{T} I(m_t = l - M) \neq 0$ then

$$\sum_{t=1}^{T} \gamma_l(m_t, x_t) = e_T^l(h) T$$

Now, I want to show that for every $T \in \mathbb{N}$, $h \in H_{T-1}$ and $x \in \{0, 1\}$,

$$\left(\sum_{t=1}^{T-1} \gamma(m_t, x_t) \right)^+ \cdot \sum_{m=1}^{M} L_T(h)(m) \gamma(m, x) \leq 0$$

For this, we consider two cases:

1. There exists $m \in \mathcal{M}$ such that $\rho^m_{T-1}(h) \in I(m)$;
2. For all $m \in \mathcal{M}$, $\rho^m_{T-1}(h) \notin I(m)$.

In the first case, there exists $\bar{m} \in \mathcal{M}$ such that $\rho^{\bar{m}}_{T-1}(h) \in I(\bar{m})$, and

$$L_T(h)(m) := \begin{cases} 1 \text{ if } m = \bar{m} \\ \\ 0 \text{ otherwise} \end{cases}$$

Then,

$$\begin{aligned} & \left(\sum_{t=1}^{T-1} \gamma(m_t, x_t)\right)^+ \cdot \sum_{m=1}^{M} L_T(h)(m)\gamma(m,x) \\ = \quad & \left(\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t)\right)^+ L_T(h)(\bar{m})\gamma_{\bar{m}}(\bar{m},x) \\ & + \left(\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t)\right)^+ L_T(h)(\bar{m})\gamma_{\bar{m}+M}(\bar{m},x) \end{aligned}$$

If $\sum_{t=1}^{T-1} I(m_t = \bar{m}) = 0$, then $\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t) = 0$ and $\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t) = 0$, so the result is obvious. Else, $\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t) = d^{\bar{m}}_{T-1}(h)(T-1)$ and $\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t) = e^{\bar{m}}_{T-1}(h)(T-1)$, which implies that $\left(\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t)\right)^+ = 0$ and $\left(\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t)\right)^+ = 0$, since $\rho^{\bar{m}}_{T-1}(h) \in I(\bar{m})$ implies that $d^{\bar{m}}_{T-1}(h) \leq 0$ and $e^{\bar{m}}_{T-1}(h) \leq 0$.

In the second case, there exists some $\bar{m} \in \mathcal{M}$ such that $d^{\bar{m}}_{T-1}(h) > 0$, $e^{\bar{m}-1}_{T-1}(h) > 0$, and

$$L_T(h)(m) := \begin{cases} \frac{e^{\bar{m}-1}_{T-1}(h)}{d^{\bar{m}}_{T-1}(h) + e^{\bar{m}-1}_{T-1}(h)}, \text{ if } m = \bar{m}; \\ \\ \frac{d^{\bar{m}}_{T-1}(h)}{d^{\bar{m}}_{T-1}(h) + e^{\bar{m}-1}_{T-1}(h)}, \text{ if } m = \bar{m} - 1; \\ \\ 0, \text{ otherwise.} \end{cases}$$

Then,

$$\begin{aligned}
& \left(\sum_{t=1}^{T-1}\gamma(m_t,x_t)\right)^+\cdot\sum_{m=1}^{M}L_T(h)(m)\gamma(m,x) \\
= & \left(\sum_{t=1}^{T-1}\gamma_{\bar{m}}(m_t,x_t)\right)^+ L_T(h)(\bar{m})\gamma_{\bar{m}}(\bar{m},x) \\
& +\left(\sum_{t=1}^{T-1}\gamma_{\bar{m}-1}(m_t,x_t)\right)^+ L_T(h)(\bar{m}-1)\gamma_{\bar{m}-1}(\bar{m}-1,x) \\
& +\left(\sum_{t=1}^{T-1}\gamma_{\bar{m}+M}(m_t,x_t)\right)^+ L_T(h)(\bar{m})\gamma_{\bar{m}+M}(\bar{m},x) \\
& +\left(\sum_{t=1}^{T-1}\gamma_{\bar{m}+M-1}(m_t,x_t)\right)^+ L_T(h)(\bar{m}-1)\gamma_{\bar{m}+M-1}(\bar{m}-1,x)
\end{aligned}$$

Since , $d_{T-1}^{\bar{m}}(h)>0$ and $e_{T-1}^{\bar{m}-1}(h)>0$, it follows that $\sum_{t=1}^{T-1}I(m_t=\bar{m}-1)\neq 0$, $\sum_{t=1}^{T-1}I(m_t=\bar{m})(h)\neq 0$, $e_{T-1}^{\bar{m}}(h)<0$, and $d_{T-1}^{\bar{m}-1}(h)<0$.This implies that

$$\begin{aligned}
\sum_{t=1}^{T-1}\gamma_{\bar{m}}(m_t,x_t) &= d_{T-1}^{\bar{m}}(h)(T-1)\geq 0 \\
\sum_{t=1}^{T-1}\gamma_{\bar{m}-1}(m_t,x_t) &= d_{T-1}^{\bar{m}-1}(h)(T-1)\leq 0 \\
\sum_{t=1}^{T-1}\gamma_{\bar{m}+M}(m_t,x_t) &= e_{T-1}^{\bar{m}}(h)(T-1)\leq 0 \\
\sum_{t=1}^{T-1}\gamma_{\bar{m}+M-1}(m_t,x_t) &= e_{T-1}^{\bar{m}-1}(h)(T-1)\geq 0
\end{aligned}$$

and, hence, that

$$\begin{aligned}
& \left(\sum_{t=1}^{T-1}\gamma(m_t,x_t)\right)^+\cdot\sum_{m=1}^{M}L_T(h)(m)\gamma(m,x) \\
= & d_{T-1}^{\bar{m}}(h)\frac{e_{T-1}^{\bar{m}-1}(h)}{d_{T-1}^{\bar{m}}(h)+e_{T-1}^{\bar{m}-1}(h)}\left(\frac{\bar{m}-1}{M}-x\right)(T-1) \\
& +e_{T-1}^{\bar{m}-1}(h)\frac{d_{T-1}^{\bar{m}}(h)}{d_{T-1}^{\bar{m}}(h)+e_{T-1}^{\bar{m}-1}(h)}\left(x-\frac{\bar{m}-1}{M}\right)(T-1) \\
= & 0
\end{aligned}$$

It follows, then, from [2, §5], that for every $\epsilon>0$, there exists $T_\epsilon\in\mathbb{N}$ such that for every $x \in \{0,1\}^\infty$,

$$\mathrm{P}_{\mathcal{L},\mathbf{x}}\left(\{h \in \mathcal{M}^\infty : \exists T \geq T_\epsilon : \sum_{m=1}^{2M}\left(\sum_{t=1}^{T}\frac{\gamma_m(m_t,x_t)}{T}\right)^+ \geq \epsilon\right) < \epsilon.$$

This suffices, since, again

$$\sum_{m=1}^{2M}\left(\sum_{t=1}^{T}\frac{\gamma_m(m_t,x_t)}{T}\right)^+ = \sum_{m=1}^{M}\left(d_T^m\left((m_t,x_t)_{t=1}^T\right)\right)^+ + \sum_{m=1}^{M}\left(e_T^m\left((m_t,x_t)_{t=1}^T\right)\right)^+.$$

□

Appendix

A forecast $\mathcal{L}$ and a sequence $x := (x_t)_{t=1}^\infty \in \{0,1\}^\infty$, define a probability distribution on $\{1\} \times \mathcal{M}^\infty$ as follows. Let $\mathcal{S}$ be the algebra of finite collections of finite histories:

$$\left\{ S \subseteq \{1\} \times \mathcal{M}^\infty : \begin{array}{c} |S| \in (\mathbb{N} \cup \{0\}) \\ \wedge \\ \forall C \in S, \exists T \in (\mathbb{N} \cup \{0\}) : \exists m \in \mathcal{M}^T : C = \{1\} \times \{m\} \times \mathcal{M}^\infty \end{array} \right\}$$

and define the outer measure $\mathrm{P}^* : \mathcal{S} \to [0,1]$ by

$$\mathrm{P}^*(\{\{1\}\times\{(m_t^s)_{t=1}^{T_s}\}\times\mathcal{M}^\infty\}_{s=1}^{S}) := \sum_{s=1}^{S}\left(L_1((1))(m_1^s)\prod_{t=2}^{T_s}L_t((m_q^s,x_q)_{q=1}^{t-1})(m_t^s)\right)$$

Then, construct the probability space $(\{1\}\times\mathcal{M}^\infty, \Sigma, \mathrm{P}_{\mathcal{L},\mathbf{x}})$, using Carathéodory's extension procedure: Σ is the set of P^*-measurable subsets of $\{1\} \times \mathcal{M}^\infty$ and $\mathrm{P}_{\mathcal{L},\mathbf{x}}$ is the restriction to Σ of the extension of P^* as

$$\mathrm{P}^*(S) := \inf\left\{\sum_{n=1}^{\infty}\mathrm{P}^*(S_n) : \{S_n\}_{n=1}^\infty \subseteq \mathcal{S} \wedge S \subseteq \bigcup_{n=1}^{\infty} S_n\right\}$$

Obviously, we can drop $\{1\}$ from the notation. Informally, we can simply consider the probability induced, recursively, as:

$$\mathrm{P}_{\mathcal{L},\mathbf{x}}\left(m_T = m|\,(m_t)_{t=1}^{T-1}, x\right) := L_T\left((m_t,x_t)_{t=1}^{T-1}\right)(m).$$

References

[1] Blackwell, D., An analog of the Minimax Theorem for vector payoffs, *Pacific Journal of Mathematics* 6, 1-8, 1956

[2] Greenwald, A., A. Jafari and C. Marks, Blackwell's Approachability Theorem: a generalization in a special case, Dept. of Computer Science, Brown University, CS-06-01, 2006.

[3] Foster, D., A prrof of calibration via Blackwell's approachability theorem, *Games and Economic Behavior* 29, 73-78, 1999.

[4] Foster, D. and R. Vohra, Asymptotic calibration, *Biometrika* 85, 379-390, 1998.