Statistical Calibration: A simplification of Foster's proof

Andrés Carvajal
Department of Economics, University of Warwick
a.m.carvajal@warwick.ac.uk

Forthcoming in *Mathematical Social Sciences*

Abstract: Foster (1999) has given a proof of the Calibration Theorem of Foster and Vohra (1998), using the Approachability Theorem proposed by Blackwell (1956). This note presents a simplified version of Foster's argument, invoking the specialization given by Greenwald et Al (2006) of Blackwell's Theorem.
Keywords: Calibration; Forecasting; Prediction of Sequences; Blackwell's Approachability Theorem.
JEL Classification Numbers: C5, C11, C73, D83.

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.

Foster and Vohra (1998) 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. A remarkable one is given by Foster (1999), where a very simple algorithm has been proved to work, using a classical result in game theory, the Approachability Theorem, proposed by Blackwell (1956). Blackwell's theorem gives sufficient conditions under which the average rewards of a player who learns from past plays can approach any closed a convex set, in a vector-valued repeated game. Recently, Greenwald et Al (2006) has specialized Blackwell's Theorem for the case where the set to be approached is the non-negative orthant of the payoff space, and has shown that a weaker condition than Blackwell's suffices in that case. It turns out that the case considered by Greenwald et Al is precisely the setting needed for Foster's argument, so,

in that sense, it provides a simplification of Foster's proof. In this note I present such an argument.

Incidentally, in this note I modify Foster's algorithm, to correct what seems to be a typographical error. Also, for the sake of clarity, I present the argument in all detail.

1. The Setting

At each future date t, an event may occur ($x_t = 1$) or not ($x_t = 0$). For each date t, a *forecast* is a number p_t representing the probability that, one suggests, the event will occur at t. The forecast for date t is made after observing the history of realizations of the event up to the previous period, $(x_s)_{s=1}^{t-1}$. Let us assume that only a subset of forecasts are acceptable: fix a positive integer M, and, for each $m \leq M$, define the interval $I(m) = [\frac{m-1}{M}, \frac{m}{M}]$ and the point $p(m) = \frac{2m-1}{2M}$;[1] it is assumed that the forecast that can be made is restricted to be an element of the set $\{p(1), \ldots, p(M)\}$.[2]

For simplicity, denote $X = (X_t)_{t=1}^{\infty}$, $\mathcal{M} = \{1, \ldots, M\}$, and, for each positive integer T, denote by H_T the set of all possible histories of forecasts and realizations up to date T, namely $H_t = (\mathcal{M} \times \{0, 1\})^T$; the generic history in that set is denoted by $h = (m_t, x_t)_{t=1}^T$, where m_t represents the forecast made for t. For definiteness, also adopt the convention that $H_0 = \{(1, 0)\}$.

2. Forecast Deficit and Excess

Given a history h of length T, define, for each possible forecast m in $\mathcal{M}$, the following numbers:

(i) The (observed) empirical frequency, conditional on m having been the forecast:

$$\rho_T^m(h) = \frac{\sum_{t=1}^T x_t I(m_t = m)}{\sum_{t=1}^T I(m_t = m)},$$

whenever the denominator of the expression is positive;[3] The denominator is zero if the forecast m has not been made along history h; in this case, simply let $\rho_T^m(h) = p(m)$.

(ii) The weighted deficit on the empirical conditional frequency relative to the lower bound of the forecast:

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

(iii) The weighted excess on the empirical conditional frequency relative to the upper bound of the forecast:

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

It is immediate that $\rho_T^m(h) \in I(m)$ if, and only if, $d_T^m(h) \leq 0$ and $e_T^m(h) \leq 0$. Also, notice that $d_T^m(h) \geq 0$ implies $e_T^m(h) < 0$, and $e_T^m(h) \geq 0$ implies $d_T^m(h) < 0$. Another useful property of these numbers is given by the following Lemma.

[1] Notice that $\cup_{m=1}^M I(M) = [0, 1]$ and that $p(m)$ is the middle point of $I(m)$.

[2] Since p defines a one-to-one correspondence, I will refer to m also as the forecast $p(m)$.

[3] Here, I denotes the standard indicator function.

LEMMA **1** (Foster)**.** *If $\rho_T^m(h) \notin I(m)$ for every forecast $m \in \mathcal{M}$, then there exists some forecast $m \in \mathcal{M}$ such that $d_T^m(h) > 0$ and $e_T^{m-1}(h) > 0$.*

Proof: By assumption, for all $m \in \mathcal{M}$, 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, it must be that $d_T^2(h) \leq 0$ and, hence $e_T^2(h) > 0$, and we can follow the search. The result follows since M is finite: at the latest, $d_T^{M-1}(h) \leq 0$, so $e_T^{M-1}(h) > 0$, which suffices since $d_T^M(h) > 0$. *Q.E.D.*

3. Randomized Forecasts and Calibration

Let Δ denote the set of probability distributions over the set of forecasts $\mathcal{M}$.[4] A *forecasting rule* is a sequence $\mathcal{L} = (L_t : H_{T-1} \to \Delta)_{t=1}^{\infty}$. That is, for a date t and given a history $h \in H_{t-1}$, the forecasting rule $\mathcal{L}$ gives a probability distribution $L_t(h)$ over $\mathcal{M}$; the interpretation is that, when forecasting for that date and after that history, forecast m is going to be chosen with probability $L_t(h)(m)$.

Given a forecast $\mathcal{L}$ and a sequence $X = (x_t)_{t=1}^{\infty} \in \{0,1\}^{\infty}$, let $\mathrm{P}_{\mathcal{L},X}$ denote the probability measure induced on $\mathcal{M}^{\infty}$.[5] In the long-run, a sequence of good forecasts 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 $I(m)$. We capture this property as follows. First, for a history h of length T, define the aggregate mistake made by the forecasts by aggregating the deficits and surpluses whenever they are positive, by letting

$$C_T(h) = \sum_{m=1}^{M} (d_T^m(h)^+ + e_T^m(h)^+);$$

a forecasting rule $\mathcal{L}$ is *calibrated* if for every $\epsilon > 0$, there exists a date T_ϵ such that, for any sequence of events X,

$$\mathrm{P}_{\mathcal{L},X}(\{h \in \mathcal{M}^{\infty} : C_T((m_t, x_t)_{t=1}^{T}) \geq \epsilon \text{ for some } T \geq T_\epsilon\}) < \epsilon.$$

Note that no structure is imposed on how the sequence $X = (x_t)_{t=1}^{\infty}$ is determined, with the only exception that it is assumed that x_t cannot be determined as a function of m_t, because the choice of the forecast is allowed to be made randomly.

[4] Namely, Δ is the unit simplex in $\mathbb{R}^M$.

[5] Let $\mathcal{S}$ be the algebra of finite collections of finite histories, and define the outer measure $\mathrm{P}^* : \mathcal{S} \to [0,1]$ by

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

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

$$\mathrm{P}^*(S) = \inf\{\sum_{n=1}^{\infty} \mathrm{P}^*(S_n) : \{S_n\}_{n=1}^{\infty} \subseteq \mathcal{S} \text{ and } S \subseteq \cup_{n=1}^{\infty} S_n\}.$$

4. A Calibrated Forecasting Rule

The following forecasting rule is a very minor modification of the one presented by Foster (1999): define $\bar{\mathcal{L}}$ as follows: for a date T, given a previous history $h \in H_{T-1}$,
(1) if there exists $\bar{m} \in \mathcal{M}$ such that $\rho_{T-1}^{\bar{m}}(h) \in I(\bar{m})$, then let $L_T(h)(\bar{m}) = 1$ and $L_T(h)(m) = 0$ for every other m;
(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[6]

$$L_T(h)(\bar{m}) = \frac{e_{T-1}^{\bar{m}-1}(h)}{d_{T-1}^{\bar{m}}(h) + e_{T-1}^{\bar{m}-1}(h)} \text{ and } L_T(h)(\bar{m}-1) = \frac{d_{T-1}^{\bar{m}}(h)}{d_{T-1}^{\bar{m}}(h) + e_{T-1}^{\bar{m}-1}(h)},$$

while $L_T(h)(m) = 0$ for every other m.

It follows from Lemma 1 that forecasting rule $\bar{\mathcal{L}}$ is well defined. It is different from the one presented by Foster (1999) 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 forecasting rule of Foster (1999) 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.

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

The proof of this result given by Foster (1999) represents the problem as an infinitely-repeated game played between the forecaster and Nature (who chooses the realizations of the event). In this representation, Blackwell's Approachability Theorem is used to show that the forecaster can force his average payoff to approach any closed and convex set, and in particular the non-negative orthant, which is equivalent to calibration, as we will see below. The result in Greenwald et Al (2006) specializes Blackwell's Theorem for the specific case needed here: it gives a sufficient condition under which the forecaster's average payoff will approximate the non-negative orthant.

Before giving the argument, recall the definition of a *a vector-valued game* given by Blackwell (1956): it is a 4-tuple $\Gamma = (A, A', V, \gamma)$ consisting of: (i) the set of actions of a player, A, which is assumed to be finite; (ii) the set of actions of the opponent(s), A'; (iii) a vector space over $\mathbb{R}$, set V, endowed with an inner product; and (iv) an outcome function $\gamma : A \times A' \to V$.

Now, consider the infinite, sequential repetition of the vector-valued game defined by $A = \mathcal{M}$, $A' = \{0, 1\}$, $V = \mathbb{R}^{2M}$, and, as in Foster (1999), γ defined as follows: for each component $l \in \{1, \ldots, 2M\}$,

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

It is obvious that γ is bounded.

[6] The following expression corrects a typo that appeared in the published version of this paper.

For a horizon T and a history h of length T, it is immediate that if $\sum_{t=1}^{T} I(m_t = l) = 0$, namely if forecast $l \in \{1, \ldots, M\}$ has not been made along history h, then $\sum_{t=1}^{T} \gamma_l(m_t, x_t) = 0$. Importantly, when forecast l has been made, so that $\sum_{t=1}^{T} I(m_t = l) \neq 0$, by construction we have that

$$\sum_{t=1}^{T} \gamma_l(m_t, x_t) = \sum_{t=1}^{T} I(m_t = l)(\frac{l-1}{M}) - \sum_{t=1}^{T} x_t I(m_t = l) = d_T^l(h)T.$$

We can apply a similar argument for the excess functions, to get that for every $l \in \{M + 1, \ldots, 2M\}$, 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 $\sum_{t=1}^{T} \gamma_l(m_t, x_t) = e_T^l(h)T$. With these results, the following Lemma will yield the Theorem.

LEMMA **2.** *For every date T, every history $h \in H_{T-1}$ and every realization $x \in \{0, 1\}$,*

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

Proof: Consider two cases:
Case 1: there exists some forecast $m \in \mathcal{M}$ such that $\rho_{T-1}^m(h) \in I(m)$. By construction, for some forecast $\bar{m} \in \mathcal{M}$ such that $\rho_{T-1}^{\bar{m}}(h) \in I(\bar{m})$, we have that $L_T(h)(\bar{m}) = 1$ and $L_T(h)(m) = 0$ for every other m. Immediately, the left-hand side of expression (1) becomes simply

$$(\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t))^+ L_T(h)(\bar{m})\gamma_{\bar{m}}(\bar{m}, x) + (\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t))^+ L_T(h)(\bar{m})\gamma_{\bar{m}+M}(\bar{m}, x).$$

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, $\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t) = d_{T-1}^{\bar{m}}(h)(T-1)$ and $\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t) = e_{T-1}^{\bar{m}}(h)(T-1)$, which implies that $(\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t))^+ = 0$ and $(\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t))^+ = 0$, since $\rho_{T-1}^{\bar{m}}(h) \in I(\bar{m})$ implies that $d_{T-1}^{\bar{m}}(h) \leq 0$ and $e_{T-1}^{\bar{m}}(h) \leq 0$.
Case 2: for every forecast $m \in \mathcal{M}$, $\rho_{T-1}^m(h) \notin I(m)$. Again by construction, there exists some $\bar{m} \in \mathcal{M}$ such that $d_{T-1}^{\bar{m}}(h) > 0$, $e_{T-1}^{\bar{m}-1}(h) > 0$,

$$L_T(h)(\bar{m}) = \frac{e_{T-1}^{\bar{m}-1}(h)}{d_{T-1}^{\bar{m}}(h) + e_{T-1}^{\bar{m}-1}(h)} \text{ and } L_T(h)(\bar{m}-1) = \frac{d_{T-1}^{\bar{m}}(h)}{d_{T-1}^{\bar{m}}(h) + e_{T-1}^{\bar{m}-1}(h)},$$

while $L_T(h)(m) = 0$ for every other m. The left-hand side of (1) now equals the sum of the four following terms:

$$(\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t))^+ L_T(h)(\bar{m})\gamma_{\bar{m}}(\bar{m}, x),$$

$$(\sum_{t=1}^{T-1} \gamma_{\bar{m}-1}(m_t, x_t))^+ L_T(h)(\bar{m}-1)\gamma_{\bar{m}-1}(\bar{m}-1, x),$$

$$(\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t))^+ L_T(h)(\bar{m})\gamma_{\bar{m}+M}(\bar{m}, x),$$

and

$$(\sum_{t=1}^{T-1} \gamma_{\bar{m}+M-1}(m_t, x_t))^+ L_T(h)(\bar{m}-1)\gamma_{\bar{m}+M-1}(\bar{m}-1, x).$$

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

$$\sum_{t=1}^{T-1} \gamma_{\bar{m}}(m_t, x_t) = d^{\bar{m}}_{T-1}(h)(T-1) \geq 0 \text{ and } \sum_{t=1}^{T-1} \gamma_{\bar{m}-1}(m_t, x_t) = d^{\bar{m}-1}_{T-1}(h)(T-1) \leq 0,$$

while

$$\sum_{t=1}^{T-1} \gamma_{\bar{m}+M}(m_t, x_t) = e^{\bar{m}}_{T-1}(h)(T-1) \leq 0 \text{ and } \sum_{t=1}^{T-1} \gamma_{\bar{m}+M-1}(m_t, x_t) = e^{\bar{m}-1}_{T-1}(h)(T-1) \geq 0.$$

It follows that the left-hand side of equation (1) is simply

$$d^{\bar{m}}_{T-1}(h)\frac{e^{\bar{m}-1}_{T-1}(h)}{d^{\bar{m}}_{T-1}(h) + e^{\bar{m}-1}_{T-1}(h)}(\frac{\bar{m}-1}{M} - x)(T-1) + e^{\bar{m}-1}_{T-1}(h)\frac{d^{\bar{m}}_{T-1}(h)}{d^{\bar{m}}_{T-1}(h) + e^{\bar{m}-1}_{T-1}(h)}(x - \frac{\bar{m}-1}{M})(T-1),$$

which is 0 by direct computation. *Q.E.D.*

Now, to prove the Theorem, it suffices to observe that, from Lemma 2 and Theorem 5 in Greenwald et Al (2006), we have that for every $\epsilon > 0$, there exists $T_\epsilon \in \mathbb{N}$ such that, for every $X \in \{0,1\}^\infty$,

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

This suffices, since, once again,

$$\sum_{m=1}^{2M}(\sum_{t=1}^{T} \frac{\gamma_m(m_t, x_t)}{T})^+ = \sum_{m=1}^{M}(d^m_T((m_t, x_t)^T_{t=1}))^+ + \sum_{m=1}^{M}(e^m_T((m_t, x_t)^T_{t=1}))^+.$$

References

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

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.

Foster, D., "A proof of calibration via Blackwell's approachability theorem," *Games and Economic Behavior* 29, 73-78, 1999.

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