Biometrika (1998), **85**, 2, *pp.* 379–390
Printed in Great Britain

Asymptotic calibration

BY DEAN P. FOSTER

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

foster@hellspark.wharton.upenn.edu

AND RAKESH V. VOHRA

Department of Management Science, Fisher College of Business, Ohio State University, Columbus, Ohio 43210, U.S.A.

vohra.1@osu.edu

SUMMARY

Can we forecast the probability of an arbitrary sequence of events happening so that the stated probability of an event happening is close to its empirical probability? We can view this prediction problem as a game played against Nature, where at the beginning of the game Nature picks a data sequence and the forecaster picks a forecasting algorithm. If the forecaster is not allowed to randomise, then Nature wins; there will always be data for which the forecaster does poorly. This paper shows that, if the forecaster can randomise, the forecaster wins in the sense that the forecasted probabilities and the empirical probabilities can be made arbitrarily close to each other.

Some key words: Brier score; Calibration; Competitive ratio; Regret; Universal prediction of sequences; Worst case.

1. INTRODUCTION

Probability forecasting is the act of assigning probabilities to an uncertain event. It is an activity widely practised in meteorological circles. For example, since 1965, the U.S. National Weather Service has been in the habit of making and announcing 'probability of precipitation' forecasts. Such a forecast is interpreted to be the probability that precipitation, defined to be at least 0·01 inches, will occur in a specified time period and area. These forecasts are now popularly accepted by the American public as meaningful and informative.

There are many criteria for judging the effectiveness of the probability forecast (Murphy & Epstein, 1967). In this paper we limit ourselves to the consideration of calibration, sometimes termed reliability. Dawid (1982) offers the following intuitive definition of calibration:

> 'Suppose that, in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of precipitation was, say, close to some given value ω and (assuming these form an infinite sequence) determine the long run proportion p of such days on which the forecast event (rain) in fact occurred. The plot of p against ω is termed the forecaster's empirical calibration curve. If the curve is the diagonal $p = \omega$, the forecaster may be termed (empirically) well calibrated'.

We give a rigorous definition later.

Calibration by itself is not a sufficient condition for a forecast to be deemed good. To see this, suppose there are two weather forecasters facing the following weather sequence: dry, wet, dry, wet, ... One always forecasts a probability of $\frac{1}{2}$ of rain each day and the other alternates 0, 1, 0, 1, ... Both forecasters are well calibrated, but the forecasts of the first are clearly less useful than those of the second. Now consider two uncalibrated forecasts, the first of which always forecasts a probability of $\frac{1}{3}$ and the second of which alternates 1, 0, 1, 0, ..., always generating an incorrect forecast. Which of these two is better is a matter of debate; the first has a lower quadratic error but the second gets the 'pattern' of rain correct. Both seem dominated by the two forecasts discussed previously. Thus, calibration does seem to be an appealing minimal property that any probability forecast should satisfy.

The notion of calibration only makes sense if one can construct forecasts that are calibrated. Regrettably, Oakes (1985) has proved that no deterministic forecasting sequence can be calibrated for all possible sequences; see Dawid (1985) for a different proof. Specifically, Oakes shows that it is impossible to construct a joint distribution for an infinite sequence of events whose posterior mean is guaranteed to be calibrated for every possible sequence of outcomes.

A way around this impossibility result is to relax the requirement that a forecast be calibrated against all possible sequences. Perhaps it is sufficient that the forecaster be calibrated for some restricted family of distributions. Dawid (1985) argues that this can result in forecasting schemes that are computationally burdensome and in some cases not computable at all. Alternatively, one can reject the notion that calibration is at all a desirable or useful notion. Schervish (1985), for example, offers two arguments for this view. The first is that calibration is a long run criterion: in the short run, when we are alive, a forecaster may do quite well. The second is that, while a malevolent Nature may be able to make one forecaster look bad according to the calibration criterion, it is harder for her to make many forecasters look bad at the same time.

Our goal in this paper is to rescue the notion of calibration. We get around the impossibility result of Oakes by broadening the definition of calibration to include randomised forecasts. By carefully choosing our definition of calibration for randomised forecasts, we show how to construct a forecast which is in fact approximately calibrated. Finally, we generalise our results to the case when what is being forecast is a distribution, not just a point.

2. NOTATION AND DEFINITIONS

For ease of exposition assume our forecasting method, F, is assigned the task of forecasting the probabilities of two states of nature, wet or dry. The main result holds for more than two states. The proof is the same, just involving more notation. Denote by X_t the outcome in period t: $X_t = 1$ if it is wet and $X_t = 0$ if it is dry. Denote by X^T the sequence of wet and dry days up to and including period T. Since we can interpret X^T to be the first T terms of an infinite sequence X^∞ that has been revealed to us, we will, when there is no ambiguity, write X for X^T.

In our context a forecasting method is simply a function that associates with any binary sequence, from the space of all binary sequences, a unique number in the interval $[0, 1]$. A randomised forecasting method would associate with each binary sequence a probability distribution over $[0, 1]$ which governs the selection of a number in $[0, 1]$. The forecast

that F makes in period t will be denoted by $f_t = F(X^{t-1})$. Let $n_T(p; F, X)$ be the number of times F forecasts p up to time T. Let $\rho_T(p; X, F)$ be the fraction of those times that it actually rained. That is,

$$n_T(p; F, X) \equiv \sum_{t=1}^{T} I(f_t = p), \quad \rho_T(p; F, X) \equiv \sum_{t=1}^{T} \frac{I(f_t = p)}{n_t(p)},$$

where I is the indicator function. In the original definition of calibration it was assumed that F was restricted to selecting forecasts from a finite set, A, fixed a priori. One definition of calibration is the following: F is well calibrated with respect to X if and only if, for each $p \in A$,

$$\lim_{t \to \infty} \rho_t(p; X, F) = p.$$

Another definition is based on the calibration component of the Brier score (Brier, 1950; Murphy, 1972, 1973); see Blattenberger & Lad (1985) for an exposition. To introduce this definition, let the calibration score of F with respect to X after n periods be denoted by $C_t(F, X)$, where

$$C_t(F, X) = \sum_{p \in A} \{\rho_t(p; X, F) - p\}^2 \frac{n_t(p; F, X)}{t}.$$

Thus, F is well calibrated with respect to X if and only if $C_t(F, X)$ goes to zero as t goes to infinity.

The requirement that F select from a fixed set A is not a severe restriction for practical purposes. Many weather forecasters forecast probabilities to only one decimal place.

3. Rules of the game

So that the assumptions underlying our analysis are clear, we frame the analysis in terms of a repeated game between two players. One is the statistician (he) making the probability forecasts and the other is Nature (she) who chooses the outcomes. Nature picks the data X and the statistician picks the forecast function F. The payment from the statistician to Nature after t rounds of play is $C_t(F, X)$. The statistician would like to play so that $C_t(F, X)$ is vanishingly small in the limit. In the case when the statistician employs a randomised forecasting rule, the goal is to make $C_t(F, X)$ vanishingly small in some probabilistic sense, which we will specify later. For the moment, if the statistician succeeds in making $C_t(F, X)$ vanishingly small in an appropriate sense, we will say that he wins the game.

Whether or not the statistician can win the game depends on his forecasting scheme as well as the power of his adversary, Nature. We describe a number of different assumptions about the power of the adversary.

Prescient adversary. In this scenario Nature knows the forecast that the statistician will make on each round before she chooses the next element of the sequence, 'rain' or 'shine'. It is impossible for the statistician to win the game under these conditions (Oakes, 1985). To see why, suppose the statistician uses some deterministic forecasting scheme F. Consider the following procedure for generating a sequence X (Dawid, 1985):

$$X_t = \begin{cases} 1 & \text{if in period } i \text{ the forecaster predicts a probability } \leqslant 0{\cdot}5, \\ 0 & \text{otherwise.} \end{cases}$$

A straightforward calculation establishes that $C_t(F, X) \geqslant \frac{1}{4}$ for all deterministic forecasting methods F. The case of equality occurs when F is the forecast that generates $f_t = \frac{1}{2}$ for all i.

It makes no difference if the statistician employs a randomised forecasting scheme. Since Nature knows the forecast before she moves, this is essentially equivalent to the deterministic set-up once we have conditioned on the randomisation.

If the statistician's pay-off or loss function is changed, for example if the Brier score is used instead of the calibration score, then the statistician can win at this game against a prescient adversary (Foster, 1991; Littlestone & Warmuth, 1994; Vovk, 1990; Feder, Mehrav & Gutman, 1992; Cover, 1991).

Oblivious adversary. In this scenario Nature knows only the forecasting scheme that will be used by the statistician. Nature then picks the entire sequence at the start of the game. In the case when the statistician uses a deterministic forecasting scheme, there is no difference between the oblivious and prescient adversary. Nature need only run the statistician's forecasting algorithm to work out what he will predict on each round. When the statistician uses a randomised forecast, however, picking the sequence at the start of the game means that Nature will not know the results of the randomisation on each round, although she knows the distribution over the different forecasts that the statistician will use. It will follow as an immediate corollary of the main result that the statistician can win against an oblivious adversary. Results for other kinds of loss functions can be found in Hanan (1957) and Foster & Vohra (1993).

Adaptive adversary. In this scenario Nature will know the forecasting rule used by the statistician. If the statistician uses a deterministic forecast, Nature will be able to work out the forecast that will be generated before she moves. If the forecast used is a randomised one, Nature will know the distribution over the possible forecasts before she moves but not the actual realisation. Unlike the oblivious adversary, Nature is not restricted to choosing the entire sequence at the start of the game. She can condition on the previous plays that the statistician has made. Thus, as time goes on, Nature can learn more about what the statistician's behaviour is, but the statistician can still randomise on the next move so that Nature does not know exactly what he will do.

There are two equivalent ways of viewing the strategies used by the statistician against an adaptive adversary. The first is to allow the statistician to randomise only before the first move. In other words, he picks a single forecasting scheme at random. Alternatively, he is allowed to randomise on each successive round. Assuming that Nature can only observe the actions taken by the statistician and not the actual randomisation, these two variations are equivalent. It is the second view we adopt in this paper.

The adaptive adversary appears to be less powerful than the prescient one, and this is a weakness on which the statistician can capitalise. Consider the randomised forecasting strategy defined as follows:

$$f_t = \begin{cases} \frac{2}{3} & \text{with probability } \frac{2}{3} \text{ if } X_{t-1} = 1, \\ \frac{1}{3} & \text{with probability } \frac{1}{3} \text{ if } X_{t-1} = 1, \\ \frac{2}{3} & \text{with probability } \frac{1}{3} \text{ if } X_{t-1} = 0, \\ \frac{1}{3} & \text{with probability } \frac{2}{3} \text{ if } X_{t-1} = 0. \end{cases}$$

For this particular strategy one can establish after tedious calculations that

$$\max_X C_t(F, X) = \tfrac{2}{3} + o_p(1),$$

where $o_p(1)$ tends to zero in probability as n tends to infinity. Hence, $\min_F \max_X C_t(F, X) < \frac{1}{4}$, with probability tending to 1. By randomising in this way we are implicitly operating two forecasting strategies instead of one. Thus, Nature finds it harder to miscalibrate F. The rest of this paper will show how to improve this $\frac{2}{3} + o_p(1)$ down to $o_p(1)$.

In this paper we assume that Nature is an adaptive adversary: Freund & Schapire (1995) discussed the relationship between the adaptive and oblivious adversary. Here is a precise statement of the rules we will follow.

Rule 1. The statistician begins by choosing a randomised forecasting method or function F and reveals only the distribution of this forecast to Nature.

Rule 2. In period $t \geqslant 1$, the statistician generates f_t $(= F(X^{t-1}))$, and, simultaneously, Nature selects the value of X_t. Nature knows the distribution of strategies that the statistician will use, but not the actual value of the randomisation.

Rule 3. The penalty the statistician incurs after t rounds is $C_t(F, X)$.

Alternatively, we can state our goal without recourse to using a game-theoretic model. The object is to find an F that is ε-calibrated, according to the following definition.

DEFINITION (ε-*calibration*). *A randomised forecast rule F is ε-calibrated if, for all X,*

$$\lim_{t \to \infty} \mathrm{pr}\{C_t(F, X) < \varepsilon\} > 1 - \varepsilon.$$

Note that X is allowed to be a stochastic process that may depend on the previous realisations of the F's and X's.

Within the context of a game, the existence of an F that is ε-calibrated is implied by showing that $\min_F \max_X E\{C_t(F, X)\}$ is less than ε^2 for all t sufficiently large, where the expectations are over the distributions chosen by Nature and by the statistician. The proof is via Jensen's inequality:

$$\min_F \max_X P\{C_t(F, X) > \varepsilon\} \leqslant \min_F \max_X E\{C_t(F, X)\}/\varepsilon = \varepsilon.$$

Likewise, if an F exists which is ε-calibrated then our security level is less than 2ε. Since both of these can be shown to go to zero, they are equivalent statements.

4. AN ARGUMENT OF SERGIU HART

Some readers of earlier versions of this paper have derived alternative proofs of the main result. Among these is the following elegant but non-constructive argument of Sergiu Hart for the existence of a winning strategy for the statistician.

The proof constructs a zero-sum game played between the statistician and Nature. Each will have a finite number of strategies and so the minimax theorem will hold. Fix at n the number of times to be forecast. In order that the statistician's strategy space is finite, we restrict him to picking one of the following as a forecast: $0, 1/k, 2/k, \ldots, 1$. Here k is some sufficiently large integer that will be chosen later. A pure strategy for the statistician consists of an n-vector of such forecasts. Thus his strategy space consists of $(k+1)^{2^n - 1}$ pure strategies. Nature's strategy space is then the set of all 2^n binary strings. In fact, technically, Nature has 2^n strategies only if we assume that she is an oblivious adversary. If Nature is an adaptive adversary, her strategy space is much larger since her strategy in each round can depend on what she saw in the previous rounds. In this case she has $2^{\{(k+1)^n - 1\}/k}$ strategies. Nevertheless, the argument is the same.

Now suppose that Nature has to pick her strategy first. To achieve her minimax value she will randomise among her choice of pure strategies. We can now assume that the statistician knows the randomisation policy that Nature will follow. To use the minimax theorem we now need to specify a strategy for the statistician which will keep his loss below ε. If we can do this for all possible strategies of Nature then there must exist a strategy for the statistician which will guarantee him a loss below ε.

How should the statistician behave if he knows what random policy Nature will follow? At each point in time he can compute the conditional probability of the next item in the sequence. He can then round this probability to the nearest i/k value, which he then forecasts. If we assume that k is much less than $n^{1/3}$, his calibration score will be less than $1/k$ with high probability. Here is an outline of why this must be so. The forecaster's calibration score is

$$C_t(F, X) = \sum_{j=0}^{k} \{\rho_t(j/k;\, X, F) - j/k\}^2 \frac{n_t(j/k;\, F, X)}{t}.$$

Now consider all the times on which the statistician forecasts j/k. He did so because the probability that Nature would pick a 1 on that round was some number q such that $|q - j/k| \leqslant 1/(2k)$. By a law of large numbers type of argument we would expect then that $|\rho_t(j/k;\, X, F) - j/k| \leqslant 1/(2k)$. Hence

$$C_t(F, X) \leqslant \sum_{j=0}^{k} \left(\frac{1}{k}\right)^2 \frac{n_t(j/k;\, F, X)}{t} = O\left(\frac{1}{k}\right).$$

Thus there exists a strategy which will guarantee him a win. The drawback is that determination of the strategy is impractical in that it requires the solution of an enormous linear programme. In the next section we describe a different and more efficient way of constructing an ε-calibrated forecast.

5. CONSTRUCTING AN ε-CALIBRATED FORECAST

In round t the algorithm randomly selects one element from the set

$$A = \{0, 1/k, 2/k, \ldots, (k-1)/k, 1\}$$

according to the distribution μ_t. We shall find a distribution μ_t over the set A such that the random forecast F which forecasts $f_t = i/k$ with probability μ_t^i will be ε-calibrated. These μ_t^i may be history dependent.

First define the expected Brier score for a randomised forecast to be

$$\text{EBS}_t(F, X) \equiv \sum_{s=0}^{t} \sum_{i=0}^{k} \frac{\mu_s^i (X_s - i/k)^2}{t}.$$

This score is averaged over the randomisation of the forecast, that is μ_t, but not over the data X, making it a 'prequential expectation' (Dawid, 1984, 1993). Define a new random forecast $F^{i\to j}$ to be exactly the same as F except that, whenever F makes a forecast of i/k, $F^{i\to j}$ makes a forecast of j/k. It might happen that $F^{i\to j}$ has a lower expected Brier score than F and hence it is a better forecast than F. When this happens, the difference in their Brier scores is called the regret of changing i/k to j/k.

DEFINITION (*Regret*). *Define the regret of changing i/k to j/k to be*

$$R_t^{i\to j} \equiv t\{\text{EBS}_t(F, X) - \text{EBS}_t(F^{i\to j}, X)\}^+,$$

where $(x)^+$ *is the positive part of* x. *In other words, if we define* S_t^{ij} *to be the signed difference in the Brier score, namely,*

$$S_t^{ij} = \sum_{s=1}^{t} \mu_s^i \left(X_s - \frac{i}{k} \right)^2 - \sum_{s=1}^{t} \mu_s^i \left(X_s - \frac{j}{k} \right)^2, \tag{1}$$

then the regret from changing i/k *to* j/k *is*

$$R_t^{i \to j} = \begin{cases} S_t^{ij} & \text{if } S_t^{ij} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

In Theorem 3 we show that the calibration score of F is closely related to the maximal regret: in particular,

$$C_t(F, X) = \sum_i \max_j \frac{R_t^{i \to j}}{t} + O(1/k^2).$$

We pick our distribution μ_t so that it satisfies the following conservation condition for all i:

$$\sum_{j \neq i} \mu_t^j R_{t-1}^{j \to i} = \mu_t^i \sum_{j \neq i} R_{t-1}^{i \to j}. \tag{2}$$

THEOREM 1. *A randomised forecast using the* μ_t *defined by equation* (2) *with* $k \doteq 1/\varepsilon^2$ *is an* ε*-calibrated forecast.*

We prove this theorem in the next section. The remainder of this section is devoted to showing that the algorithm is well defined. For convenience we suppress the dependence on t. Let A be a matrix with elements

$$a_{ij} = R^{j \to i}, \tag{3}$$

for all $i \neq j$, and

$$a_{ii} = -\sum_{j \neq i} R^{i \to j}. \tag{4}$$

Note that the row sums of A are all zero. Equation (2) is equivalent to $Ax = 0$. We need to show that system $Ax = 0$ admits a nontrivial and nonnegative solution, which can be normalised to turn it into a probability vector.

Let A' be the matrix with elements $a'_{ij} = a_{ij}/B$, where $B = \max_{i,j} |a_{ij}|$. Note that $|a'_{ij}| \leqslant 1$ and $\sum_i a'_{ij} = 0$. Let $P = A' + I$. Then P will be a nonnegative row-stochastic matrix. Hence there is a nonnegative probability vector x such that $Px = x$: since we do not require that x be unique, we do not need any restrictions on the matrix P. Since $P = A' + I$ we deduce that

$$A'x + Ix = x \Rightarrow A'x = 0 \Rightarrow Ax = 0.$$

The vector x gives the required distribution, and it can easily be found by Gaussian elimination.

6. PROOF THAT THE ALGORITHM WORKS

As F is essentially fixed, we can for convenience suppress the dependence on F in our notation. We write $n_t(p)$, $\rho_t(p)$ and C_t for $n_t(p; F, X)$, $\rho_t(p; F, X)$ and $C_t(F, X)$ respectively.

Table 1. *Modified versions of n, ρ and C*

Base definitions	Modified definitions
$n_t(i/k) \equiv \sum_{s=1}^{t} I(f_t = i/k)$	$\tilde{n}_t(i/k) \equiv \sum_{s=1}^{t} \mu_s^i$
$\rho_t(i/k) \equiv \sum_{s=1}^{t} \frac{I(\hat{f}_t = i/k)X_s}{n_t(i/k)}$	$\tilde{\rho}_t(i/k) \equiv \sum_{s=1}^{t} \frac{\mu_s^i X_s}{\tilde{n}_t(i/k)}$
$C_t \equiv \sum_{j=0}^{k} \frac{n_t(j/k)}{t}\{\rho_t(i/k) - j/k\}^2$	$\tilde{C}_t \equiv \sum_{j=0}^{k} \frac{\tilde{n}_t(j/k)}{t}\{\tilde{\rho}_t(j/k) - j/k\}^2$

The proof divides into two steps. In the first step we show that C_t can be closely approximated by something akin to its average value. To this end define modified versions of n, ρ and C as given in Table 1. Note that $n_t(i/k) - \tilde{n}_t(i/k)$ and $\rho_t(i/k)n_t(i/k) - \tilde{\rho}_t(i/k)\tilde{n}_t(i/k)$ are both martingales. This allows us to approximate C_t by $\tilde{C}_t$, as follows.

THEOREM 2. *In probability, $C_t - \tilde{C}_t \to 0$ as $t \to \infty$.*

Proof. The function $\mathscr{C}(.)$, defined by

$$\mathscr{C}(a_0, a_1, \ldots, a_k, b_0, b_1, \ldots, b_k) = \sum_{j=0}^{k} a_j \left(\frac{b_j}{a_j} - \frac{j}{k}\right)^2,$$

is a continuous function over the compact set $0 \leqslant b_i \leqslant a_i \leqslant 1$; it is uniformly continuous. We can rewrite C_t and $\tilde{C}_t$ as

$$C_t = \mathscr{C}\left(\frac{n_t(0)}{t}, \frac{n_t(1/k)}{t}, \ldots, \frac{n_t(1)}{t}, \frac{\rho_t(0)n_t(0)}{t}, \frac{\rho_t(1/k)n_t(1/k)}{t}, \ldots, \frac{\rho_t(1)n_t(1)}{t}\right),$$

$$\tilde{C}_t = \mathscr{C}\left(\frac{\tilde{n}_t(0)}{t}, \frac{\tilde{n}_t(1/k)}{t}, \ldots, \frac{\tilde{n}_t(1)}{t}, \frac{\tilde{\rho}_t(0)\tilde{n}_t(0)}{t}, \frac{\tilde{\rho}_t(1/k)\tilde{n}_t(1/k)}{t}, \ldots, \frac{\tilde{\rho}_t(1)\tilde{n}_t(1)}{t}\right).$$

It is sufficient to show that the differences in the arguments converge to zero in order to establish that $C_t - \tilde{C}_t$ converges to zero.

Since $n_t(i/k) - \tilde{n}_t(i/k)$ and $\rho_t(i/k)n_t(i/k) - \tilde{\rho}_t(i/k)\tilde{n}_t(i/k)$ are both counting processes, their jumps are bounded by 1, and hence the variances of the jumps are trivially bounded by 1. In other words,

$$\operatorname{var}[\{n_t(i/k) - \tilde{n}_t(i/k)\} - \{n_{t-1}(i/k) - \tilde{n}_{t-1}(i/k)\}] \leqslant 1,$$

$$\operatorname{var}[\{\rho_t(i/k)n_t(i/k) - \tilde{\rho}_t(i/k)\tilde{n}_t(i/k)\} - \{\rho_{t-1}(i/k)n_{t-1}(i/k) - \tilde{\rho}_{t-1}(i/k)\tilde{n}_{t-1}(i/k)\}] \leqslant 1,$$

which leads to

$$\operatorname{var}\left\{\frac{n_t(i/k) - \tilde{n}_t(i/k)}{t}\right\} \leqslant 1/t, \quad \operatorname{var}\left\{\frac{\rho_t(i/k)n_t(i/k) - \tilde{\rho}_t(i/k)\tilde{n}_t(i/k)}{t}\right\} \leqslant 1/t.$$

Since L_2 convergence implies convergence in probability, we see that $C_t - \tilde{C}_t \to 0$ in probability. If almost sure convergence is desired, it will follow from a similar argument using 4th moments. □

In the second step we show that this 'average' calibration score goes to zero with t. This is done by using the regret to bound the average calibration score and then proving

that the regret is asymptotically small. There is a technical difficulty to be overcome. Regret as we have defined it is a function of the form $\max\{x, 0\}$, giving it a 'kink' at zero. To smooth away this kink we introduce a function g_δ that is differentiable at 0 and approximates $\max\{x, 0\}$.

Define

$$\tilde{R}_t^\delta \equiv \sum_{i,j} g_\delta(R_t^{i\to j}),$$

where

$$g_\delta(x) \equiv \begin{cases} \frac{1}{2}\delta x^2 & (x \geqslant 0), \\ 0 & (x \leqslant 0). \end{cases} \tag{5}$$

Note that $g_\delta(R_t^{i\to j}) = g_\delta(S_t^{ij})$, where S_t^{ij} is defined by equation (1).

THEOREM 3. *The calibration score is related to the regret by*

$$\sum_i \max_j R^{i\to j} \leqslant t\tilde{C}_t \leqslant \sum_i \max_j R^{i\to j} + \frac{t}{4k^2} \leqslant \tilde{R}_t^\delta + \frac{k}{2\delta} + \frac{t}{4k^2}.$$

Proof. First note that

$$\begin{aligned} S_t^{ij} &= \frac{2(j-i)}{k} \sum_{s=1}^{t} \mu_s^i \left\{ X_s - \frac{1}{2}\left(\frac{i+j}{k}\right)\right\} \\ &= \frac{2(j-k)\tilde{n}_t(i/k)}{k} \left\{ \tilde{\rho}_t\left(\frac{i}{k}\right) - \frac{1}{2}\left(\frac{i+j}{k}\right)\right\} \\ &= \tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - i/k\}^2 - \tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - j/k\}^2. \end{aligned}$$

Thus,

$$\tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - i/k\}^2 = S_t^{ij} + \tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - j/k\}^2 \geqslant S_t^{ij} \geqslant \max_j S_t^{ij} = \max_j R_t^{i\to j}.$$

Now summing over both sides provides the first inequality.

For the second inequality observe that the maximum regret occurs at the point where $\tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - j/k\}^2$ is smallest. Thus

$$\tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - i/k\}^2 = \max_j(S_t^{ij}) + \min_j \tilde{n}_t(i/k)\{\tilde{\rho}_t(i/k) - j/k\}^2 \leqslant \max_j(S_t^{ij}) + \frac{\tilde{n}_t(i/k)}{4k^2}.$$

Summing over i provides the second inequality.

Since $1/(2\delta) + g_\delta(x) \geqslant x$ we see that

$$\max_j R^{i\to j} \leqslant \frac{1}{2\delta} + \max_j g_\delta(R^{i\to j}) \leqslant \frac{1}{2\delta} + \sum_j g_\delta(R^{i\to j}),$$

and summing over i leads to the last inequality. □

THEOREM 4. *For the μ_t defined in equation* (2), $\tilde{R}_t^\delta \leqslant tk\delta$.

Proof. Note that $g_\delta(x+\alpha) - g_\delta(x) \leqslant \delta\alpha(x)^+ + \delta\alpha^2$. Define $L_t^j \equiv (X_t - j/k)^2$, so that $S_t^{ij} = S_{t-1}^{ij} + \mu_t^i(L_t^j - L_t^i)$. Since

$$g_\delta(S_t^{ij}) - g_\delta(S_{t-1}^{ij}) \leqslant \delta\mu_t^i(L_t^j - L_t^i)R_{t-1}^{i\to j} + \delta(\mu_t^i)^2(L_t^j - L_t^i)^2$$

we obtain

$$\tilde{R}_t^\delta - \tilde{R}_{t-1}^\delta = \sum_{i,j} \{g_\delta(S_t^{ij}) - g_\delta(S_{t-1}^{ij})\} \leqslant \delta \sum_{i,j} [\{\mu_t^i L_t^i - \mu_t^i L_t^j\} R_{t-1}^{i \to j} + \{\mu_t^i L_t^i - \mu_t^i L_t^j\}^2].$$

From equation (2) we see that

$$\sum_{i,j:i \neq j} \{\mu_t^i L_t^i - \mu_t^i L_t^j\} R_{t-1}^{i \to j} = \sum_i L_t^i \left\{ \sum_j \mu_t^i R_{t-1}^{i \to j} - \mu_t^j R_{t-1}^{j \to i} \right\} = 0,$$

where the equality follows by interchanging the dummy arguments i and j.

Thus,

$$\tilde{R}_t^\delta - \tilde{R}_{t-1}^\delta \leqslant \delta \sum_{i,j} (\mu_t^i L_t^i - \mu_t^i L_t^j)^2 \leqslant \delta \sum_{i,j} (\mu_t^i)^2 = \delta k \sum_i (\mu_t^i)^2 \leqslant \delta k,$$

which gives $\tilde{R}_t^\delta - \tilde{R}_0^\delta \leqslant tk\delta$. However, we know that $\tilde{R}_0^\delta = 0$, so that $\tilde{R}_t^\delta \leqslant tk\delta$. □

Combining Theorems 1–4 yields the following obvious but technical corollary.

COROLLARY. *For all $\varepsilon > 0$, if $k > \varepsilon^{-\frac{1}{4}}$ and $t_0 > 8k^2/\varepsilon$, then, for all $t \geqslant t_0$, we have that $\tilde{C}_t \leqslant \varepsilon$. Further, there exists a $t_1 > t_0$ such that, for all $t \geqslant t_1$, $\mathrm{pr}(C_t < \varepsilon) \geqslant 1 - \varepsilon$.*

With care, ε can be chosen to be $O(t_0^{-1/3})$. Theorem 1 now follows directly from this corollary.

Theorem 1 can be strengthened to almost sure convergence. We sketch the argument here. First run a 2^{-i}-calibrated algorithm for a 'long time' and then switch to a $2^{-(i+1)}$-calibrated algorithm. Repeat indefinitely. The hard part is defining what a 'long time' means. It must be sufficiently long such that each stage has a probability of at most 2^{-i} of ever being above $2^{-(i-1)}$. It also must be sufficiently long that we can amortise the 'phase in' period of the $2^{-(i+1)}$-calibrated algorithm. Combining this with the almost sure version of Theorem 2 yields a non-constructive proof of the existence of an algorithm such that $C_t \to 0$ almost surely.

7. FORECASTING WITH DISTRIBUTIONS

Suppose that, instead of making a point forecast f_t of the probability that $X_t = 1$, we forecast a distribution $\mu_t(.)$. For example, suppose some sort of hierarchical model is considered such that there is a parameter p_t for each time t. Then we could think of a posterior for p_t, that is a distributional forecast of a binary event. Clearly, the definition of calibration must be generalised if it is to be applied to a distributional forecast. A reasonable definition of $\rho_t(.)$ is

$$\rho_t(p; X) = \frac{\sum_{s \leqslant t} d\mu_s(p) X_s}{\sum_{s \leqslant t} d\mu_s(p)}.$$

Hence, if μ is a distributional forecast, its calibration with respect to X after n periods is

$$C_T(\mu, X) = \frac{1}{T} \sum_{t=1}^{T} \int_0^1 \{\rho_t(p) - p\}^2 \, d\mu_t(p).$$

Note that, if the distributional forecast is a degenerate one, i.e. a point forecast, the definition of calibration reduces to the one given earlier in the paper. Given what we know about deterministic point forecasts, we can assert that some distributional forecasts are not calibrated.

Any randomised point forecast can be viewed as a distributional forecast. If this is done, the calibration score is exactly $\tilde{C}_t$, as defined in § 6. Simply treat the randomisation at each period as the distribution being forecast. In this case the calibration of the distributional forecast is a number, in contrast to the calibration of the associated randomised point forecast, which is a random variable. This observation yields the following corollary to Theorems 1–3.

COROLLARY. *There is a distributional forecast* $\mu(.)$ *such that, for all* X *and* $\varepsilon > 0$, $C_t(\mu, X) \leqslant \varepsilon$ *for all* t *sufficiently large.*

If we think of $\mu_t(.)$ as a posterior distribution for $\text{pr}(X_t = 1)$, then we can combine Oakes' (1985) result with the corollary to conclude that a posterior mean might not be calibrated for all X but that there are some posterior distributions that are always calibrated. Thus, in terms of calibration, the posterior distribution is a better statistic than the posterior mean.

8. DISCUSSION

Our goal in this paper has been to rescue the notion of calibration. We have done this by generalising the original definition of calibration offered by Dawid to allow for randomised forecasts. Further, we have shown that this weakened definition is not vacuous, by exhibiting a forecasting scheme that satisfies it.

The scheme we propose, while ε-calibrated, achieves this at the cost of a higher Brier score. To see why, consider the forecast that at each round forecasts $\sum_{j=1}^{k} \mu_t^j$. By the convexity of the Brier score, this forecast will have a Brier score that is smaller than the expected Brier score of the randomised forecast that uses the distribution μ_t.

While this paper was under review a number of other researchers have been moved to find alternative proofs of our main result. One of these approaches, due to Sergiu Hart, we have already described. A constructive version of Hart's proof has been derived independently by Drew Fudenberg and David Levine. They view each step to be forecast as a game and then use the minimax theorem to compute the value of that step. Sergiu Hart and Andreu Mas-Colell have recently shown that our Theorem 4 is a corollary of David Blackwell's (1956) approachability theorem and so provide an alternative proof that our algorithm will in fact be ε-calibrated.

ACKNOWLEDGEMENT

We thank A. P. Dawid, Arnold Zellner, Daniel Nelson and the referees for some useful comments, a number of which have been incorporated into this paper.

REFERENCES

BLACKWELL, D. (1956). An analog of the minimax theorem for vector payoffs. *Pac. J. Math.* **6**, 1–8.
BLATTENBERGER, G. & LAD, F. (1985). Separating the Brier score into calibration and refinement components: A graphical exposition. *Am. Statistician* **39**, 26–32.
BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **75**, 1–3.
COVER, T. (1991). Universal portfolios. *Math. Finan.* **16**, 876–98.
DAWID, A. P. (1982). The well calibrated Bayesian. *J. Am. Statist. Assoc.* **77**, 605–13.
DAWID, A. P. (1984). Statistical theory, the prequential approach. *J. R. Statist. Soc.* A **147**, 278–92.
DAWID, A. P. (1985). The impossibility of inductive inference. *J. Am. Statist. Assoc.* **80**, 340–1.

DAWID, A. P. (1993). Prequential data analysis. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, Ed. M. Ghosh and P. Pathak, IMS Lecture Notes Monograph Series, **17**, pp. 113–26. Hayward, CA: Inst. Math. Statist.

FEDER, M., MEHRAV, N. & GUTMAN, M. (1992). Universal prediction of individual sequences. *IEEE Trans. Info. Theory* **38**, 1258–70.

FOSTER, D. P. (1991). A worst case analysis of prediction. *Ann. Statist.* **21**, 625–44.

FOSTER, D. P. & VOHRA, R. (1993). A randomization rule for selecting forecasts. *Oper. Res.* **41**, 704–9.

FREUND, Y. & SCHAPIRE, R. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pp. 23–37. Springer-Verlag.

HANAN, J. (1957). Approximation to Bayes risk in repeated plays. In *Contributions to the Theory of Games of Games*, Ed. M. Dresher, A. W. Tucker and P. Wolfe, **3**, pp. 97–139. Princeton University Press.

LITTLESTONE, N. & WARMUTH, M. (1994). The weighted majority algorithm. *Info. Comp.* **108**, 212–61.

MURPHY, A. H. (1972). Scalar and vector partitions of the probability score. Part I: Two-state situation. *J. Appl. Meteor.* **11**, 273–8.

MURPHY, A. H. (1973). A new vector partition of the probability score. *J. Appl. Meteor.* **12**, 595–600.

MURPHY, A. H. & EPSTEIN, E. (1967). Verification of probabilistic predictions: A brief review. *J. Appl. Meteor.* **6**, 748–55.

OAKES, D. (1985). Self-calibrating priors do not exist. *J. Am. Statist. Assoc.* **80**, 339.

SCHERVISH, M. (1985). Comment on paper by Oakes. *J. Am. Statist. Assoc.* **80**, 341–2.

VOVK, V. (1990). Aggregating strategies. In *Proceedings of the 3rd Annual Conference on Computational Learning Theory*, pp. 71–83. San Francisco, CA: Morgan Kaufmann.

[*Received August* 1994. *Revised February* 1997]