The Annals of Statistics
1991, Vol. 19, No. 2, 1084–1090

PREDICTION IN THE WORST CASE[1]

BY DEAN P. FOSTER

University of Chicago

A predictor is a method of estimating the probability of future events over an infinite data sequence. One predictor is as strong as another if for all data sequences the former has at most the mean square error (MSE) of the latter. Given any countable set $\mathscr{D}$ of predictors, we explicitly construct a predictor S that is at least as strong as every element of $\mathscr{D}$. Finite sample bounds are also given which hold uniformly on the space of all possible data.

1. Introduction and summary. As in Rissanen (1986), the data will be an infinite sequence of binary variables R_i, in other words, $R = \{R_i, \text{ for } i = 1, 2, \dots\}$ is a $\{0, 1\}$-valued stochastic process. Ω is the set of infinite binary sequences. Let $\mathscr{F}_n$ be the natural σ-field induced by $R_1, R_2, \dots, R_n$. Call M a predictor if at time $n - 1$ it outputs a number $M(n)$ which estimates the probability $P(R_n = 1|\mathscr{F}_{n-1})$. Technically, M is a predictor if and only if it is an $\mathscr{F}_{n-1}$-adapted stochastic process taking values in the interval $[0, 1]$. This model of probability forecasting is discussed in Dawid (1984, 1986).

Consider two predictors: A and B. Which is better? Define

$$\overline{Q}_R(A, n) = \frac{1}{n}\sum_{i=1}^{n}(A(i) - R_i)^2,$$

the mean square error between prediction and observation—called the Brier score in meteorology [Brier (1950)]. We can compare predictors A and B by computing $d_R(A, B)$,

$$d_R(A, B) = \liminf_{n\to\infty}(\overline{Q}_R(A, n) - \overline{Q}_R(B, n)).$$

Note that $d_R(\cdot, \cdot)$ is not a metric, nor is it antisymmetric. Since $\overline{Q}_R$ is bounded between zero and one, $d_r(\cdot, \cdot)$ is finite. If $d_R(A, B) > 0$, then in the limit B forecasts better than A on the data sequence R.

DEFINITION. B is as strong as A if for all sequences R in Ω, $d_R(A, B) \geq 0$.

DEFINITION. B is stronger than A if it is as strong as A and there exists a sequence R in Ω such that $d_R(A, B) > 0$.

The results of this paper concern relationships among predictors, such as strong or stronger, which hold *for all* sample paths. Although not explicitly

Received October 1987; revised April 1990.
[1]Research supported by the Office of Naval Research under Contract FAS-01-5-28870 at the University of Maryland.
AMS 1980 *subject classifications*. Primary 62M20; secondary 62A99.
Key words and phrases. Comparing forecasts, worst-case behavior, mean square error.

mentioned, such properties will hold a fortiori almost surely or in expectation with respect to any measure $\mathscr{P}$ on Ω. For example, if B is as strong as A, then for all measures $\mathscr{P}$,

$$\liminf_{n\to\infty} E\big(\overline{Q}_R(A,n)\big) - E\big(\overline{Q}_R(B,n)\big) \geq 0.$$

The goal of this paper is to propose and justify a new method for combining predictors. More precisely, consider two predictors A and B, neither of which is as strong as the other. This new method will yield a predictor S which is as strong as *both* A and B. In other words, S combines the best properties of both A and B; if A predicts well on some sequence of R_i, then so does S.

Empirical properties of various other methods of combining forecasts have been discussed in Newbold and Granger (1974) and Gupta and Wilton (1987). A concept similar to strength which deals with portfolio selection in finance is due to Cover (1991).

2. Exploring strength.

EXAMPLE. Define two predictors: ZERO$(n) \equiv 0$ and ONE$(n) \equiv 1$. Then, ZERO is not as strong as ONE, nor is ONE as strong as ZERO. This is seen by considering the two data sequences $R_i = 1$ for all i and the sequence $R'_i = 0$ for all i. Then, $d_R(\text{ONE}, \text{ZERO}) = -1$, $d_{R'}(\text{ZERO}, \text{ONE}) = -1$.

This situation seems reasonable—each of these two predictors has some data that it does better on. Is stronger a vacuous definition? No; the following lemma gives a general construction.

LEMMA 1. *For any arbitrary predictor M, there exists a predictor S which is stronger than M.*

PROOF. We need to find two things to prove this lemma: a predictor S which is as strong as M and a point ω in Ω for which this predictor does better than M. Find a point ω in Ω with the following property:

$$\omega_i = \begin{cases} 1, & \text{if } M(i)(\omega) \leq 0.5, \\ 0, & \text{if } M(i)(\omega) > 0.5. \end{cases}$$

This point can be recursively constructed because $M(i)$ is $\mathscr{F}_{i-1}$-measurable. Note that $\overline{Q}_\omega(M,n) \geq 0.25$ for all n. Now let

$$S(i)(R) = \begin{cases} \omega_i, & \text{if } R_j = \omega_j \text{ for all } j < i, \\ M(i), & \text{otherwise.} \end{cases}$$

Thus, for all points $R \neq \omega$ in Ω, $d_R(M,S) = 0$. Also, $d_\omega(M,S) \geq 0.25$. Thus, S is stronger than M. □

One would expect from the name that strength is transitive. This is so. It follows from superadditivity of limit inferior: $d_R(C,A) \geq d_R(C,B) + d_R(B,A) \geq 0$.

3. Results. Suppose we have a finite set $\mathscr{D}$ of predictors. If a strongest element of $\mathscr{D}$ does not exist, then we cannot decide which predictor from $\mathscr{D}$ to use. Thus, the search is on for finding a predictor $S_{\mathscr{D}}$ which is as strong as every element of $\mathscr{D}$. The subscript $\mathscr{D}$ in $S_{\mathscr{D}}$ stresses that S depends on which set is being dominated.

Call the elements of $\mathscr{D}$: $f_1, f_2, f_3, \ldots, f_d$. For instance, the prediction made by the second predictor in $\mathscr{D}$ at time n is $f_2(n)$. Define the convex hull of $\mathscr{D}$ to be the set of predictors created by convex weightings of predictors in $\mathscr{D}$.

DEFINITION. Convex hull of $\mathscr{D}$ is $\{M | \dot{M}(i) = w_1 f_1(i) + w_2 f_2(i) + \cdots + w_d f_d(i)$ for $w_i \geq 0$ and $w_1 + w_2 + \cdots + w_d = 1\}$.

In vector notation, $M = w^T f$, where $w = [w_1, w_2, \ldots, w_d]^T$ and $f = [f_1, f_2, \ldots, f_d]^T$. At time n, we write the prediction made by M as $M(n) = w^T f(n)$.

DEFINITION. Let B represent the simplex

$$B = \left\{ w \in \mathbb{R}^d \middle| \sum_{i=1}^{d} w_i = 1, w_i \geq 0 \right\}.$$

The convex hull of $\mathscr{D}$ can be written as $\{M | M = w^T f, w \in B\}$.

We now state the main theorem.

THEOREM 1. *For $\mathscr{D}$ a collection of d predictors, let*

$$S_{\mathscr{D}}(n) = \hat{w}(n-1)^T f(n),$$

where $\hat{w}(n-1)$ minimizes $(n-1)\overline{Q}(w^T f, n-1) + w^T w$ over w in B. Then $S_{\mathscr{D}}$ is as strong as every element of the convex hull of $\mathscr{D}$. In particular: $\overline{Q}_R(S_{\mathscr{D}}, n) - \overline{Q}_R(M, n) \leq (2 + d \log d(n+1))/n$ for every R in Ω and for all M in the convex hull of $\mathscr{D}$.

We will prove theorem later in the section.

Because $(n-1)\overline{Q} + w^T w$ is a quadratic over a compact set, the minimum exists but it need not be unique. If it is not unique, just pick one arbitrarily. No matter how this is done, we have

$$(n-1)\overline{Q}\big((\hat{w}(n-1) + x)^T f, n-1\big) + (\hat{w}(n-1) + x)^T(\hat{w}(n-1) + x)$$
$$-(n-1)\overline{Q}\big(\hat{w}(n-1)^T f, n-1\big) - \hat{w}(n-1)^T \hat{w}(n-1) \geq 0$$

for all $x \in \mathbb{R}^d$ with $\hat{w}(n-1) + x \in B$. Let $L_{n-1}(x)$ denote the linear part in x of this expression. $L_{n-1}(x)$ can be thought of as the directional derivative of

$(n-1)\overline{Q}+w^Tw$ at its minimum point $(\hat{w}(n-1))$. If $\hat{w}(n-1)$ happens to be in the interior B, then $L_{n-1}(x)=0$. More generally, it follows that $L_{n-1}(x)\geq 0$ if $\hat{w}(n-1)+x\in B$.

Define Δ, a function of time i and an $\mathbb{R}^d$-vector x, as

$$\Delta(i,x)=i\Big(\overline{Q}_R\big(\hat{w}(i-1)^Tf,i\big)-\overline{Q}_R\big((\hat{w}(i-1)+x)^Tf,i\big)\Big)$$
$$+\hat{w}(i-1)^T\hat{w}(i-1)-(\hat{w}(i-1)+x)^T(\hat{w}(i-1)+x).$$

Using the fact that $\hat{w}(i-1)$ minimizes $\sum_{j=1}^{i}(R_i-w^Tf)^2+w^Tw$ over B, we have for $\hat{w}(i-1)+x\in B$,

$$\Delta(i,x)=-x^TG_ix+a_ix^Tf(i)-L_{i-1}(x)\leq -x^TG_ix+a_ix^Tf(i),$$

where $f(i)$ is the vector of predictions at time i, $G_i=I+\sum_{j=1}^{i}f(j)f(j)^T$, $a_i=2(R_i-f(i)^T\hat{w}(i-1))$ is a real number with $|a_i|\leq 2$ and I is the identity matrix.

Our first claim relates Δ to our total error.

CLAIM A.

$$\overline{Q}_R(S_{\mathscr{D}},n)-\overline{Q}_R\big(\hat{w}(n)^Tf,n\big)$$
$$=\frac{1}{n}\left(\hat{w}(n)^T\hat{w}(n)-\hat{w}(0)^T\hat{w}(0)+\sum_{i=1}^{n}\max_{\hat{w}(i-1)+x\in B}\Delta(i,x)\right).$$

PROOF.

$$\overline{Q}_R(S_{\mathscr{D}},n)=\frac{1}{n}\sum_{i=1}^{n}\Big(i\overline{Q}_R\big(\hat{w}(i-1)^Tf,i\big)$$
$$-(i-1)\overline{Q}_R\big(\hat{w}(i-1)^Tf,i-1\big)\Big)$$
$$=\overline{Q}_R\big(\hat{w}(n)^Tf,n\big)$$
$$+\frac{1}{n}\sum_{i=1}^{n}i\Big(\overline{Q}_R\big(\hat{w}(i-1)^Tf,i\big)-\overline{Q}_R\big(\hat{w}(i)^Tf,i\big)\Big)$$
$$+\frac{1}{n}\left(\hat{w}(n)^T\hat{w}(n)-\hat{w}(0)^T\hat{w}(0)\right.$$
$$\left.-\sum_{i=1}^{n}\Big[\hat{w}(i-1)^T\hat{w}(i-1)-\hat{w}(i)^T\hat{w}(i)\Big]\right)$$
$$=\overline{Q}_R\big(\hat{w}(n)^Tf,n\big)$$
$$+\frac{1}{n}\left(\hat{w}(n)^T\hat{w}(n)-\hat{w}(0)^T\hat{w}(0)\right.$$
$$\left.+\sum_{i=1}^{n}\Delta(i,\hat{w}(i)-\hat{w}(i-1))\right).$$

Notice that $\Delta(i,\cdot)$ attains the maximum at $x=\hat{w}(i)-\hat{w}(i-1)$. □

In Claim B, we bound the size of Δ.

CLAIM B.

$$\max_{\hat{w}(i-1)+x\in B} \Delta \le \max_{x\in\mathbb{R}^d} \left(-x^T G_i x + a_i x^T f(i)\right) \le \frac{|G_i| - |G_{i-1}|}{|G_i|}.$$

PROOF. Note $G_i = G_{i-1} + f(i) f(i)^T$. Because G_i is the identity plus a nonnegative definite matrix, we know that G_i is invertible. Let $|G_i|$ be the determinant of G_i.

$$|G_{i-1}| = \left|G_i - f(i) f(i)^T\right|,$$

$$|G_{i-1}| = |G_i|\left(1 - f(i)^T G_i^{-1} f(i)\right),$$

$$f(i)^T G_i^{-1} f(i) = \frac{|G_i| - |G_{i-1}|}{|G_i|}.$$

But the maximum value of Δ is less than or equal to $a_i f(i)^T G_i^{-1} f(i) a_i/4$. (Maximizing the quadratic $-x^T A x + b^T x$, where A is a positive definite matrix, yields $b^T A^{-1} b/4$.) Recalling that $|a_i| \le 2$, we get the desired result. □

CLAIM C.

$$\sum_{i=1}^{n} \max_{\hat{w}(i-1)+x\in B} \Delta(i, x) \le d \log d(n+1).$$

PROOF. Using Claim B and the equation $1 - x \le -\log(x)$, we get

$$\sum_{i=1}^{n} \max \Delta(i, x) \le \log(|G_n|).$$

$|G_n| \le L^d$, where L is the maximum eigenvalue of G_n. The elements of G_n are bounded by $n+1$. Thus, L is bounded by $d(n+1)$. □

PROOF OF THEOREM 1. For all M in the convex hull of $\mathscr{D}$,

$$\begin{aligned}
&\overline{Q}_R(M, n) - \overline{Q}_R(S_{\mathscr{D}}, n) \\
&\quad \ge \overline{Q}_R\left(\hat{w}(n)^T f, n\right) - \overline{Q}_R(S_{\mathscr{D}}, n) - \frac{1}{n} \\
&\quad \ge \frac{1}{n}\left(-\sum_{i=1}^{n} \max \Delta(i, x) + \hat{w}(0)^T \hat{w}(0) - \hat{w}(n)^T \hat{w}(n) - 1\right) \\
&\quad \ge \frac{1}{n}\left(-d \log d(n+1) - 2\right),
\end{aligned}$$

where the first inequality follows from $\overline{Q}_R(\hat{w}(n)^T f, n) \leq \overline{Q}_R(M, n) + 1/n$ for all M in $\mathscr{D}$. The second line is by Claim A. For $w \in B$, $w^T w \leq 1$. This plus Claim C yields the last inequality. Thus

$$d_R(M, S_{\mathscr{D}}) \geq \liminf_{n\to\infty} \frac{-d \log d(n+1) - 2}{n} = 0. \qquad \square$$

Since f_i is in the convex hull of $\mathscr{D}$, we see that $S_{\mathscr{D}}$ is as strong as f_i. So, of the predictors, $\{f_1, f_2, \ldots, f_d, S_{\mathscr{D}}\}$, the strongest is $S_{\mathscr{D}}$.

The bound in Theorem 1 holds for all R, so in particular it holds for the worst possible R—called the worst case. Thus it deals with prediction in the worst case.

THEOREM 2. *For any countable set $\mathscr{D}$ of predictors, there exists a predictor $T_{\mathscr{D}}$ which is as strong as every member of $\mathscr{D}$.*

PROOF. Define T_i for $i = 1, 2, 3, \ldots$ as

$$T_i(n) = \begin{cases} f_i(n), & \text{if } i \geq n, \\ S_{\mathscr{D}_i}(n), & \text{if } i < n, \end{cases}$$

where $\mathscr{D}_i = \{T_{i+1}, f_i\}$. Notice that for all times n, all T_i are well-defined. By Theorem 1, T_i is as strong as f_i and as strong as T_{i+1}. By transitivity, T_1 is as strong as f_i. In other words, T_1 is as strong as T_2, which is as strong as $T_3, \ldots$, which is as strong as T_i, which is as strong as f_i, so T_1 is as strong as f_i. Letting $T_{\mathscr{D}} = T_1$ proves our result. $\square$

EXAMPLE. Theorem 2 is only an asymptotic result. In other words, a set $\mathscr{D}$ can be constructed such that at all finite times n, there exist a perfect $f_i \in \mathscr{D}$ with $\overline{Q}(f_i, n) = 0$, but $\overline{Q}(T_{\mathscr{D}}, n) \cong \frac{1}{2}$. Theorem 2 forces $\liminf_{n\to\infty} \overline{Q}(f_i, n) \geq \frac{1}{2}$ for all f_i in $\mathscr{D}$. So, in this case,

$$0 = \lim_{n\to\infty} \inf_i \overline{Q}(f_i, n) < \liminf_{n\to\infty} \overline{Q}(T_{\mathscr{D}}, n) \leq \inf_i \liminf_{n\to\infty} \overline{Q}(f_i, n).$$

Acknowledgments. I am grateful for the help of Grace Yang, Michael Cohen, Rick Vohra and the referees. But the most thanks go to Peyton Young who provided encouragement, support and a general guiding hand.

REFERENCES

BRIER, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78** 1–3.

COVER, T. M. (1991). Universal portfolios. *Math. Finance* **1** 1–12.

DAWID, A. P. (1984). Statistical theory: The prequential approach (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

DAWID, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences* **7** 210–218. Wiley, New York.

GUPTA, S. and WILTON, P. C. (1987). Combination of forecasts: An extension. *Management Sci.* **3** 356–372.

NEWBOLD, P. and GRANGER, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *J. Roy. Statist. Soc. Ser. A* **137** 131–146.

RISSANEN, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080–1100.

UNIVERSITY OF CHICAGO
GRADUATE SCHOOL OF BUSINESS
CHICAGO, ILLINOIS 60637