Prediction with Expert Advice for the Brier Game

Vladimir Vovk VOVK@CS.RHUL.AC.UK
Fedor Zhdanov FEDOR@CS.RHUL.AC.UK
Computer Learning Research Centre, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

Abstract

We show that the Brier game of prediction is mixable and find the optimal learning rate and substitution function for it. The resulting prediction algorithm is applied to predict results of football and tennis matches. The theoretical performance guarantee turns out to be rather tight on these data sets, especially in the case of the more extensive tennis data.

1. Introduction

The paradigm of prediction with expert advice was introduced in the late 1980s (see, e.g., Littlestone & Warmuth, 1994, Cesa-Bianchi et al., 1997) and has been applied to various loss functions; see Cesa-Bianchi and Lugosi (2006) for a recent book-length review. An especially important class of loss functions is that of "mixable" ones, for which the learner's loss can be made as small as the best expert's loss plus a constant (depending on the number of experts). It is known (Haussler et al., 1998; Vovk, 1998) that the optimal additive constant is attained by the "strong aggregating algorithm" proposed in Vovk (1990) (we use the adjective "strong" to distinguish it from the "weak aggregating algorithm" of Kalnishkan & Vyugin, 2005).

There are several important loss functions that have been shown to be mixable and for which the optimal additive constant has been found. The prime examples in the case of binary observations are the log loss function and the square loss function. The log loss function, whose mixability is obvious, has been explored extensively, along with its important generalizations, the Kullback–Leibler divergence and Cover's loss function.

In this paper we concentrate on the square loss function. In the binary case, its mixability was demonstrated in Vovk (1990). There are two natural directions in which this result could be generalized:

Regression: observations are real numbers (square-loss regression is a standard problem in statistics).

Classification: observations take values in a finite set (this leads to the "Brier game", to be defined below, a standard way of measuring the quality of predictions in meteorology and other applied fields: see, e.g., Dawid, 1986).

The mixability of the square loss function in the case of observations belonging to a bounded interval of real numbers was demonstrated in Haussler et al. (1998); Haussler et al.'s algorithm was simplified in Vovk (2001). Surprisingly, the case of square-loss non-binary classification has never been analysed in the framework of prediction with expert advice. The purpose of this paper is to fill this gap. The full version (Vovk & Zhdanov, 2008) of this paper is available on arXiv.

2. Prediction Algorithm and Loss Bound

A game of prediction consists of three components: the observation space Ω, the decision space Γ, and the loss function $\lambda : \Omega \times \Gamma \to \mathbb{R}$. In this paper we are interested in the following *Brier game* (Brier, 1950): Ω is a finite and non-empty set, $\Gamma := \mathcal{P}(\Omega)$ is the set of all probability measures on Ω, and

$$\lambda(\omega, \gamma) = \sum_{o \in \Omega} \left(\gamma\{o\} - \delta_\omega\{o\}\right)^2 ,$$

where $\delta_\omega \in \mathcal{P}(\Omega)$ is the probability measure concentrated at ω: $\delta_\omega\{\omega\} = 1$ and $\delta_\omega\{o\} = 0$ for $o \neq \omega$. (For example, if $\Omega = \{1,2,3\}$, $\omega = 1$, $\gamma\{1\} = 1/2$, $\gamma\{2\} = 1/4$, and $\gamma\{3\} = 1/4$, $\lambda(\omega,\gamma) = (1/2-1)^2 + (1/4-0)^2 + (1/4-0)^2 = 3/8$.)

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

The game of prediction is being played repeatedly by a learner having access to decisions made by a pool of experts, which leads to the following prediction protocol:

Protocol 1 Prediction with expert advice

```
$L_0 := 0$.
$L_0^k := 0$, $k = 1, \dots, K$.
**for** $N = 1, 2, \dots$ **do**
    Expert $k$ announces $\gamma_N^k \in \Gamma$, $k = 1, \dots, K$.
    Learner announces $\gamma_N \in \Gamma$.
    Reality announces $\omega_N \in \Omega$.
    $L_N := L_{N-1} + \lambda(\omega_N, \gamma_N)$.
    $L_N^k := L_{N-1}^k + \lambda(\omega_N, \gamma_N^k)$, $k = 1, \dots, K$.
**end for**
```

At each step of Protocol 1 Learner is given K experts' advice and is required to come up with his own decision; L_N is his cumulative loss over the first N steps, and L_N^k is the kth expert's cumulative loss over the first N steps. In the case of the Brier game, the decisions are probability forecasts for the next observation.

An optimal (in the sense of Theorem 1 below) strategy for Learner in prediction with expert advice for the Brier game is given by the strong aggregating algorithm. For each expert k, the algorithm maintains its weight w^k, constantly slashing the weights of less successful experts.

Algorithm 1 Strong aggregating algorithm for the Brier game

```
$w_0^k := 1$, $k = 1, \dots, K$.
**for** $N = 1, 2, \dots$ **do**
    Read the Experts' predictions $\gamma_N^k$, $k = 1, \dots, K$.
    Set $G_N(\omega) := -\ln \sum_{k=1}^K w_{N-1}^k e^{-\lambda(\omega, \gamma_N^k)}$, $\omega \in \Omega$.
    Solve $\sum_{\omega \in \Omega} (s - G_N(\omega))^+ = 2$ in $s \in \mathbb{R}$.
    Set $\gamma_N\{\omega\} := (s - G_N(\omega))^+/2$, $\omega \in \Omega$.
    Output prediction $\gamma_N \in \mathcal{P}(\Omega)$.
    Read observation $\omega_N$.
    $w_N^k := w_{N-1}^k e^{-\lambda(\omega_N, \gamma_N^k)}$.
**end for**
```

The algorithm will be derived in Section 5. The following result (to be proved in Section 4) gives a performance guarantee for it that cannot be improved by any other prediction algorithm.

Theorem 1. *Using Algorithm 1 as Learner's strategy in Protocol 1 for the Brier game guarantees that*

$$L_N \le \min_{k=1,\dots,K} L_N^k + \ln K \tag{1}$$

for all $N = 1, 2, \dots$. If $A < \ln K$, Learner does not have a strategy guaranteeing

$$L_N \le \min_{k=1,\dots,K} L_N^k + A \tag{2}$$

for all $N = 1, 2, \dots$.

3. Experimental Results

In our first empirical study of Algorithm 1 we use historical data about 6416 matches in various English football league competitions, namely: the Premier League (the pinnacle of the English football system), the Football League Championship, Football League One, Football League Two, the Football Conference. Our data, provided by Football-Data, cover two full seasons, 2005/2006 and 2006/2007, and part of the 2007/2008 season (which ends in May shortly after the paper submission deadline). The matches are sorted first by date and then by league. In the terminology of our prediction protocol, the outcome of each match is the observation, taking one of three possible values, "home win", "draw", or "away win"; we will encode the possible values as 1, 2, and 3.

For each match we have forecasts made by a range of bookmakers. We chose eight bookmakers for which we have enough data over a long period of time, namely Bet365, Bet&Win, Gamebookers, Interwetten, Ladbrokes, Sportingbet, Stan James, and VC Bet. (And the seasons mentioned above were chosen because the forecasts of these bookmakers are available for them.)

A probability forecast for the next observation is essentially a vector (p_1, p_2, p_3) consisting of positive numbers summing to 1. The bookmakers do not announce these numbers directly; instead, they quote three betting odds, a_1, a_2, and a_3. Each number a_i is the amount which the bookmaker undertakes to pay out to a client betting on outcome i per unit stake in the event that i happens (the stake itself is never returned to the bettor, which makes all betting odds greater than 1; i.e., the odds are announced according to the "continental" rather than "traditional" system). The inverse value $1/a_i$, $i \in \{1, 2, 3\}$, can be interpreted as the bookmaker's quoted probability for the observation i. The bookmaker's quoted probabilities are usually slightly (because of the competition with other bookmakers) in his favour: the sum $1/a_1 + 1/a_2 + 1/a_3$ exceeds 1 by the amount called the *overround* (at most 0.15 in the vast majority of cases). We used

$$p_i := \frac{1/a_i}{1/a_1 + 1/a_2 + 1/a_3}, \quad i = 1, 2, 3, \tag{3}$$

as the bookmaker's forecasts; it is clear that $p_1 + p_2 + p_3 = 1$.

The results of applying Algorithm 1 to the football data, with 8 experts and 3 possible observations, are shown in Figure 1. Let L_N^k be the cumulative loss of Expert k, $k = 1, \dots, 8$, over the first N matches and L_N be the corresponding number for Algorithm 1 (i.e., we essentially continue to use the notation of Theorem 1). The dashed line corresponding to Expert k shows the excess loss $N \mapsto L_N^k - L_N$ of Expert k over Algorithm 1. The excess loss can be negative, but from Theorem 1 we know that it cannot be less than $-\ln 8$; this lower bound is also shown in Figure 1. Finally, the thick line (the positive part of the x axis) is drawn for comparison: this is the excess loss of Algorithm 1 over itself. We can see that at each moment in time the algorithm's cumulative loss is fairly close to the cumulative loss of the best expert (at that time; the best expert keeps changing over the time).

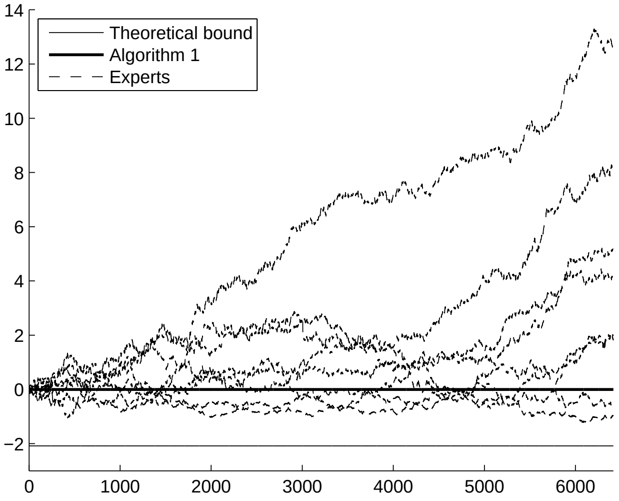

Figure 1. The difference between the cumulative loss of each of the 8 bookmakers (experts) and of Algorithm 1 on the football data. The theoretical lower bound $-\ln 8$ from Theorem 1 is also shown.

Figure 2 shows the results of another empirical study, involving data about a large number of tennis tournaments in 2004, 2005, 2006, and 2007, with the total number of matches 10,087. The tournaments include, e.g., Australian Open, French Open, Wimbledon, and US Open; the data is provided by Tennis-Data. The matches are sorted by date, then by tournament. The data contain information about the winner of each match and the betting odds of 4 bookmakers for his/her win and for the opponent's win. Therefore, now there are two possible observations (player 1's win and player 2's win). There are four bookmakers: Bet365, Centrebet, Expekt, and Pinnacle Sports. The results in Figure 2 are presented in the same way as in Figure 1. Typical values of the overround are below 0.1.

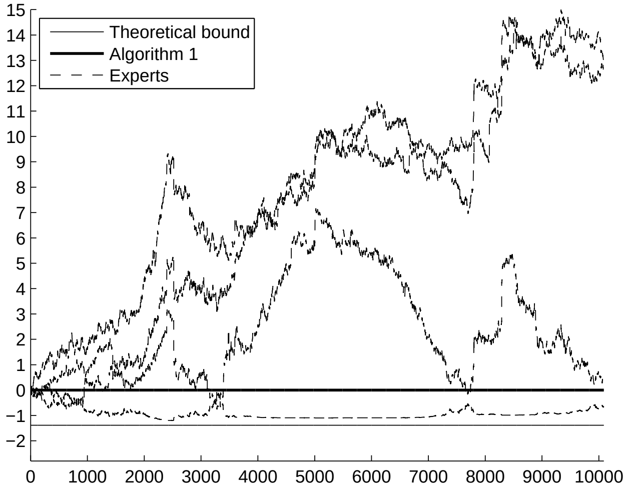

Figure 2. The difference between the cumulative loss of each of the 4 bookmakers and of Algorithm 1 on the tennis data. Now the theoretical bound is $-\ln 4$.

In both Figure 1 and Figure 2 the cumulative loss of Algorithm 1 is close to the cumulative loss of the best expert, despite the fact that some of the experts perform poorly. The theoretical bound is not hopelessly loose for the football data and is rather tight for the tennis data. The pictures look exactly the same when Algorithm 1 is applied in the more realistic manner where the weights w^k are not updated over the matches that are played simultaneously.

Our second empirical study (Figure 2) is about binary prediction, and so the algorithm of Vovk (1990) could have also been used (and would have given similar results). We included it since we are not aware of any empirical studies even for the binary case.

Other popular algorithms for prediction with expert advice that could be used instead of Algorithm 1 in our empirical studies are Kivinen and Warmuth's (1999) Weighted Average Algorithm (WAA) and Freund and Schapire's (1997) Hedge algorithm (HA). The performance guarantees for these two algorithms are much weaker than the optimal (1), especially in the case of the HA (even if the loss bound given in Freund & Schapire, 1997, is replaced by the stronger bound given in Vovk, 1998, Example 7). The weak performance guarantees show in the empirical performance of the algorithms. For the football data the maximal difference between the cumulative loss of both the WAA and the HA and the cumulative loss of the best expert is about twice as large as that for Algorithm 1 (and so is approximately equal to the optimal bound $\ln K$ given by (1)). For the tennis data the maximal difference for the WAA is about three times as large as for Algorithm 1, and for the HA it is about twice as large; therefore, both algorithms violate the optimal bound $\ln K$. For further details, see Vovk and Zhdanov (2008).

The data used for producing Figures 1 and 2 can be downloaded from `http://vovk.net/ICML2008`.

4. Proof of Theorem 1

This proof will use some basic notions of elementary differential geometry, especially those connected with the Gauss–Kronecker curvature of surfaces. (The use of curvature in this kind of results is standard: see, e.g., Vovk, 1990, and Haussler et al., 1998.) All definitions that we will need can be found in, e.g., Thorpe, 1979.

A vector $f \in \mathbb{R}^\Omega$ (understood to be a function $f : \Omega \to \mathbb{R}$) is a *superprediction* if there is $\gamma \in \Gamma$ such that, for all $\omega \in \Omega$, $\lambda(\omega, \gamma) \leq f(\omega)$; the set Σ of all superpredictions is the *superprediction set*. For each *learning rate* $\eta > 0$, let $\Phi_\eta : \mathbb{R}^\Omega \to (0, \infty)^\Omega$ be the homeomorphism defined by

$$\Phi_\eta(f) : \omega \in \Omega \mapsto e^{-\eta f(\omega)}, \quad f \in \mathbb{R}^\Omega.$$

The image $\Phi_\eta(\Sigma)$ of the superprediction set will be called the *η-exponential superprediction set*. It is known that

$$L_N \leq \min_{k=1,\dots,K} L_N^k + \frac{\ln K}{\eta}$$

can be guaranteed if and only if the η-exponential superprediction set is convex (part "if" for all K and part "only if" for $K \to \infty$ are proved in Vovk, 1998; part "only if" for all K is proved by Chris Watkins, and the details can be found in, e.g., Vovk, 2007, Appendix). Comparing this with (1) and (2) we can see that we are required to prove that

- $\Phi_\eta(\Sigma)$ is convex when $\eta \leq 1$;
- $\Phi_\eta(\Sigma)$ is not convex when $\eta > 1$.

Define the *η-exponential superprediction surface* to be the part of the boundary of the η-exponential superprediction set $\Phi_\eta(\Sigma)$ lying inside $(0, \infty)^\Omega$. The idea of the proof is to check that, for all $\eta < 1$, the Gauss–Kronecker curvature of this surface is nowhere vanishing. Even when this is done, however, there is still uncertainty as to in which direction the surface is bulging (towards the origin or away from it). The standard argument (as in Thorpe, 1979, Chapter 12, Theorem 6) based on the continuity of the smallest principal curvature shows that the η-exponential superprediction set is bulging away from the origin for small enough η: indeed, since it is true at some point, it is true everywhere on the surface. By the continuity in η this is also true for all $\eta < 1$. Now, since the η-exponential superprediction set is convex for all $\eta < 1$, it is also convex for $\eta = 1$.

Let us now check that the Gauss–Kronecker curvature of the η-exponential superprediction surface is always positive when $\eta < 1$ and is sometimes negative when $\eta > 1$ (the rest of the proof, an elaboration of the above argument, will be easy). Set $n := |\Omega|$; without loss of generality we assume $\Omega = \{1, \dots, n\}$.

A convenient parametric representation of the η-exponential superprediction surface is

$$\begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{n-1} \\ x^n \end{pmatrix} = \begin{pmatrix} e^{-\eta((u^1-1)^2+(u^2)^2+\cdots+(u^n)^2)} \\ e^{-\eta((u^1)^2+(u^2-1)^2+\cdots+(u^n)^2)} \\ \vdots \\ e^{-\eta((u^1)^2+\cdots+(u^{n-1}-1)^2+(u^n)^2)} \\ e^{-\eta((u^1)^2+\cdots+(u^{n-1})^2+(u^n-1)^2)} \end{pmatrix}, \qquad (4)$$

where $u^1, \dots, u^{n-1}$ are the coordinates on the surface, $u^1, \dots, u^{n-1} \in (0, 1)$ subject to $u^1 + \cdots u^{n-1} < 1$, and u^n is a shorthand for $1 - u^1 - \cdots - u^{n-1}$. The derivative of (4) in u^1 is

$$\frac{\partial}{\partial u^1} \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{n-1} \\ x^n \end{pmatrix} = 2\eta \times$$

$$\begin{pmatrix} (u^n - u^1 + 1) e^{-\eta((u^1-1)^2+(u^2)^2+\cdots+(u^{n-1})^2+(u^n)^2)} \\ (u^n - u^1) e^{-\eta((u^1)^2+(u^2-1)^2+\cdots+(u^{n-1})^2+(u^n)^2)} \\ \vdots \\ (u^n - u^1) e^{-\eta((u^1)^2+(u^2)^2+\cdots+(u^{n-1}-1)^2+(u^n)^2)} \\ (u^n - u^1 - 1) e^{-\eta((u^1)^2+(u^2)^2+\cdots+(u^{n-1})^2+(u^n-1)^2)} \end{pmatrix}$$

$$\propto \begin{pmatrix} (u^n - u^1 + 1) e^{2\eta u^1} \\ (u^n - u^1) e^{2\eta u^2} \\ \vdots \\ (u^n - u^1) e^{2\eta u^{n-1}} \\ (u^n - u^1 - 1) e^{2\eta u^n} \end{pmatrix},$$

the derivative in u^2 is

$$\frac{\partial}{\partial u^2}\begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{n-1} \\ x^n \end{pmatrix} \propto \begin{pmatrix} (u^n - u^2)e^{2\eta u^1} \\ (u^n - u^2 + 1)e^{2\eta u^2} \\ \vdots \\ (u^n - u^2)e^{2\eta u^{n-1}} \\ (u^n - u^2 - 1)e^{2\eta u^n} \end{pmatrix},$$

and so on, up to

$$\frac{\partial}{\partial u^{n-1}}\begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{n-1} \\ x^n \end{pmatrix} \propto \begin{pmatrix} (u^n - u^{n-1})e^{2\eta u^1} \\ (u^n - u^{n-1})e^{2\eta u^2} \\ \vdots \\ (u^n - u^{n-1} + 1)e^{2\eta u^{n-1}} \\ (u^n - u^{n-1} - 1)e^{2\eta u^n} \end{pmatrix},$$

all coefficients of proportionality being equal and positive.

Let us set $v^{i,j} := (u^n - u^i)e^{2\eta u^j}$ and $w^i := (u^n - u^i)$, for purely typographical reasons. A normal vector to the surface can be found as

$$Z := \begin{vmatrix} e_1 & \cdots & e_{n-1} & e_n \\ v^{1,1} + e^{2\eta u^1} & \cdots & v^{1,n-1} & v^{1,n} - e^{2\eta u^n} \\ \vdots & \ddots & \vdots & \vdots \\ v^{n-1,1} & \cdots & \begin{matrix} v^{n-1,n-1} \\ +e^{2\eta u^{n-1}} \end{matrix} & v^{n-1,n} - e^{2\eta u^n} \end{vmatrix}.$$

The coefficient in front of e_1 is the $(n-1) \times (n-1)$ determinant

$$\begin{vmatrix} v^{1,2} & \cdots & v^{1,n-1} & v^{1,n} - e^{2\eta u^n} \\ v^{2,2} + e^{2\eta u^2} & \cdots & v^{2,n-1} & v^{2,n} - e^{2\eta u^n} \\ \vdots & \ddots & \vdots & \vdots \\ v^{n-1,2} & \cdots & \begin{matrix} v^{n-1,n-1} \\ +e^{2\eta u^{n-1}} \end{matrix} & v^{n-1,n} - e^{2\eta u^n} \end{vmatrix}$$

$$\propto e^{-2\eta u^1} \begin{vmatrix} w^1 & \cdots & w^1 & w^1 - 1 \\ w^2 + 1 & \cdots & w^2 & w^2 - 1 \\ \vdots & \ddots & \vdots & \vdots \\ w^{n-1} & \cdots & w^{n-1} + 1 & w^{n-1} - 1 \end{vmatrix}$$

$$= e^{-2\eta u^1} \begin{vmatrix} 1 & 1 & \cdots & 1 & w^1 - 1 \\ 2 & 1 & \cdots & 1 & w^2 - 1 \\ 1 & 2 & \cdots & 1 & w^3 - 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 2 & w^{n-1} - 1 \end{vmatrix}$$

$$= e^{-2\eta u^1} \begin{vmatrix} 1 & 1 & \cdots & 1 & u^n - u^1 - 1 \\ 1 & 0 & \cdots & 0 & u^1 - u^2 \\ 0 & 1 & \cdots & 0 & u^1 - u^3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & u^1 - u^{n-1} \end{vmatrix}$$

$$\begin{aligned} &= e^{-2\eta u^1} \big((-1)^n (u^n - u^1 - 1) + (-1)^{n+1}(u^1 - u^2) \\ &\quad + (-1)^{n+1}(u^1 - u^3) + \cdots + (-1)^{n+1}(u^1 - u^{n-1})\big) \\ &= e^{-2\eta u^1}(-1)^n \times \\ &\quad \big((u^2 + u^3 + \cdots + u^n) - (n-1)u^1 - 1\big) \\ &= -e^{-2\eta u^1}(-1)^n n u^1 \propto u^1 e^{-2\eta u^1} \end{aligned} \qquad (5)$$

(with a positive coefficient of proportionality, $e^{2\eta}$, in the first $\propto$; the third equality follows from the expansion of the determinant along the last column and then along the first row).

Similarly, the coefficient in front of e_i is proportional (with the same coefficient of proportionality) to $u^i e^{-2\eta u^i}$ for $i = 2, \dots, n-1$; indeed, the $(n-1) \times (n-1)$ determinant representing the coefficient in front of e_i can be reduced to the form analogous to (5) by moving the ith row to the top.

The coefficient in front of e_n is proportional to

$$e^{-2\eta u^n} \begin{vmatrix} w^1 + 1 & w^1 & \cdots & w^1 & w^1 \\ w^2 & w^2 + 1 & \cdots & w^2 & w^2 \\ \vdots & \vdots & \ddots & \vdots & \\ w^{n-2} & w^{n-2} & \cdots & w^{n-2} + 1 & w^{n-2} \\ w^{n-1} & w^{n-1} & \cdots & w^{n-1} & w^{n-1} + 1 \end{vmatrix}$$

$$= e^{-2\eta u^n} \begin{vmatrix} 1 & 0 & \cdots & 0 & w^1 \\ 0 & 1 & \cdots & 0 & w^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & w^{n-2} \\ -1 & -1 & \cdots & -1 & w^{n-1} + 1 \end{vmatrix}$$

$$= e^{-2\eta u^n} \begin{vmatrix} 1 & 0 & \cdots & 0 & u^n - u^1 \\ 0 & 1 & \cdots & 0 & u^n - u^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & u^n - u^{n-2} \\ 0 & 0 & \cdots & 0 & n u^n \end{vmatrix} = n u^n e^{-2\eta u^n}$$

(with the coefficient of proportionality $e^{2\eta}(-1)^{n-1}$).

The Gauss–Kronecker curvature at the point with coordinates $(u^1, \dots, u^{n-1})$ is proportional (with a positive coefficient of proportionality, possibly depending on the point) to

$$\begin{vmatrix} \frac{\partial Z^{\mathrm{T}}}{\partial u^1} \\ \vdots \\ \frac{\partial Z^{\mathrm{T}}}{\partial u^{n-1}} \\ Z^{\mathrm{T}} \end{vmatrix} \qquad (6)$$

(Thorpe, 1979, Chapter 12, Theorem 5, with $^{\mathrm{T}}$ standing for transposition).

Set $v^i := (1 - 2\eta u^i)e^{-2\eta u^i}$ and $w^i = u^i e^{-2\eta u^i}$, again for typographical reasons. A straightforward calculation allows us to rewrite determinant (6) (ignoring the positive coefficient $((-1)^{n-1}ne^{2\eta})^n)$ as

$$\begin{vmatrix} v^1 & 0 & \cdots & 0 & -v^n \\ 0 & v^2 & \cdots & 0 & -v^n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & v^{n-1} & -v^n \\ w^1 & w^2 & \cdots & w^{n-1} & w^n \end{vmatrix} \propto$$

$$\begin{vmatrix} 1-2\eta u^1 & 0 & \cdots & 0 & -1+2\eta u^n \\ 0 & 1-2\eta u^2 & \cdots & 0 & -1+2\eta u^n \\ \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & \cdots & 1-2\eta u^{n-1} & -1+2\eta u^n \\ u^1 & u^2 & \cdots & u^{n-1} & u^n \end{vmatrix}$$
$$= u^1(1-2\eta u^2)(1-2\eta u^3)\cdots(1-2\eta u^n) + u^2(1-2\eta u^1)(1-2\eta u^3)\cdots(1-2\eta u^n) + \cdots + u^n(1-2\eta u^1)(1-2\eta u^2)\cdots(1-2\eta u^{n-1}) \qquad (7)$$

(with a positive coefficient of proportionality; to avoid calculation of the parities of various permutations, the reader might prefer to prove the last equality by induction in n, expanding the last determinant along the first column). Our goal is to show that the last expression in (7) is positive when $\eta < 1$ but can be negative when $\eta > 1$.

If $\eta > 1$, set $u^1 = u^2 := 1/2$ and $u^3 = \cdots = u^n := 0$. The last expression in (7) becomes negative. Therefore, the η-exponential superprediction set is not convex (Thorpe, 1979, Chapter 13, Theorem 1).

It remains to consider the case $\eta < 1$. Set $t_i := 1 - 2\eta u^i$, $i = 1, \ldots, n$; the constraints on the t_i are

$$-1 < 1 - 2\eta < t_i < 1, \quad i = 1, \ldots, n,$$
$$t_1 + \cdots + t_n = n - 2\eta > n - 2. \qquad (8)$$

Our goal is to prove

$$(1-t_1)t_2t_3\cdots t_n + \cdots + (1-t_n)t_1t_2\cdots t_{n-1} > 0,$$

i.e.,

$$t_2t_3\cdots t_n + \cdots + t_1t_2\cdots t_{n-1} > nt_1\cdots t_n. \qquad (9)$$

This reduces to

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} > n \qquad (10)$$

if $t_1 \cdots t_n > 0$, and to

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} < n \qquad (11)$$

if $t_1 \cdots t_n < 0$. The remaining case is where some of the t_i are zero; for concreteness, let $t_n = 0$. By (8) we have $t_1 + \cdots + t_{n-1} > n - 2$, and so all of $t_1, \ldots, t_{n-1}$ are positive; this shows that (9) is indeed true.

Let us prove (10). Since $t_1 \cdots t_n > 0$, all of $t_1, \ldots, t_n$ are positive (if two of them were negative, the sum $t_1 + \cdots + t_n$ would be less than $n-2$; cf. (8)). Therefore,

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} > \underbrace{1 + \cdots + 1}_{n \text{ times}} = n.$$

To establish (9) it remains to prove (11). Suppose, without loss of generality, that $t_1 > 0$, $t_2 > 0$,…, $t_{n-1} > 0$, and $t_n < 0$. Since the function $t \in (0, 1] \mapsto 1/t$ is convex, we can also assume, without loss of generality, $t_1 = \cdots = t_{n-2} = 1$. Then $t_{n-1} + t_n > 0$, and so

$$\frac{1}{t_{n-1}} + \frac{1}{t_n} < 0;$$

therefore,

$$\frac{1}{t_1} + \cdots + \frac{1}{t_{n-2}} + \frac{1}{t_{n-1}} + \frac{1}{t_n} < n - 2 < n.$$

Finally, let us check that the positivity of the Gauss–Kronecker curvature implies the convexity of the η-exponential superprediction set, for $\eta \le 1$. Because of the continuity of the η-exponential superprediction surface in η we can and will assume, without loss of generality, that $\eta < 1$. The η-exponential superprediction surface will be oriented by choosing the normal vector field directed towards the origin; this can be done since

$$\begin{pmatrix} x^1 \\ \vdots \\ x^n \end{pmatrix} \propto \begin{pmatrix} e^{2\eta u^1} \\ \vdots \\ e^{2\eta u^n} \end{pmatrix}, \quad Z \propto \begin{pmatrix} -u^1 e^{-2\eta u^1} \\ \vdots \\ -u^n e^{-2\eta u^n} \end{pmatrix}, \qquad (12)$$

with the first coefficient of proportionality positive (cf. (4) and the bottom row of the first determinant in (7)), and the scalar product of the two vectors in (12) is always negative.

Let us first check that the smallest principal curvature

$$k_1 = k_1(u^1, \ldots, u^{n-1}, \eta)$$

of the η-exponential superprediction surface is always positive (among the arguments of k_1 we list not only the coordinates $u^1, \ldots, u^{n-1}$ of a point on the surface (4) but also the learning rate $\eta \in (0, 1)$). At least at some $(u^1, \ldots, u^{n-1}, \eta)$ the value of $k_1(u^1, \ldots, u^{n-1}, \eta)$ is positive: take a sufficiently small η and the point on the surface (4) at which the maximum of $x^1 + \cdots + x^n$ is attained (the point of the η-exponential superprediction set at which the maximum is attained will lie on the surface since the maximum is attained at $(x^1, \dots, x^n) = (1, \dots, 1)$ when $\eta = 0$). Therefore, for all $(u^1, \dots, u^{n-1}, \eta)$ the value of $k_1(u^1, \dots, u^{n-1}, \eta)$ is positive: if k_1 had different signs at two points in the set

$$\bigl\{(u^1, \dots, u^{n-1}, \eta) \,|\, u^1 \in (0,1), \dots, u^{n-1} \in (0,1), \\ u^1 + \dots + u^{n-1} < 1, \eta \in (0,1)\bigr\}, \quad (13)$$

we could connect these points by a continuous curve lying completely inside (13); at some point on the curve, k_1 would be zero, in contradiction to the positivity of the Gauss–Kronecker curvature $k_1 \cdots k_{n-1}$.

Now it is easy to show that the η-exponential superprediction set is convex. Suppose there are two points A and B on the η-exponential superprediction surface such that the interval $[A, B]$ contains points outside the η-exponential superprediction set. The intersection of the plane OAB, where O is the origin, with the η-exponential superprediction surface is a planar curve; the curvature of this curve at the point between A and B closest to the origin will be negative (with the curve oriented by directing the normal vector field towards the origin), contradicting the positivity of k_1 at that point and Meusnier's theorem (cf. (12)).

5. Derivation of the Prediction Algorithm

To achieve the loss bound (1) in Theorem 1 Learner can use, as discussed earlier, the strong aggregating algorithm (see, e.g., Vovk, 2001, Section 2.1, (15)) with $\eta = 1$. In this section we will find a substitution function for the strong aggregating algorithm for the Brier game with $\eta \le 1$, which is the only component of the algorithm not described explicitly in Vovk (2001). Our substitution function will not require that its input, the generalized prediction, should be computed from the normalized distribution $(w^k)_{k=1}^K$ on the experts; this is a valuable feature for generalizations to an infinite number of experts (as demonstrated in, e.g., Vovk, 2001, Appendix A.1).

Suppose that we are given a generalized prediction $(l_1, \dots, l_n)^{\mathrm{T}}$ computed by the aggregating pseudo-algorithm from a normalized distribution on the experts. Since $(l_1, \dots, l_n)^{\mathrm{T}}$ is a superprediction (remember that we are assuming $\eta \le 1$), we are only required to find a permitted prediction

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} (u^1-1)^2 + (u^2)^2 + \dots + (u^n)^2 \\ (u^1)^2 + (u^2-1)^2 + \dots + (u^n)^2 \\ \vdots \\ (u^1)^2 + (u^2)^2 + \dots + (u^n-1)^2 \end{pmatrix} \quad (14)$$

(cf. (4)) satisfying

$$\lambda_1 \le l_1, \dots, \lambda_n \le l_n. \quad (15)$$

Now suppose we are given a generalized prediction $(L_1, \dots, L_n)^{\mathrm{T}}$ computed by the APA from an unnormalized distribution on the experts; in other words, we are given

$$\begin{pmatrix} L_1 \\ \vdots \\ L_n \end{pmatrix} = \begin{pmatrix} l_1 + c \\ \vdots \\ l_n + c \end{pmatrix}$$

for some $c \in \mathbb{R}$. To find (14) satisfying (15) we can first find the largest $t \in \mathbb{R}$ such that $(L_1 - t, \dots, L_n - t)^{\mathrm{T}}$ is still a superprediction and then find (14) satisfying

$$\lambda_1 \le L_1 - t, \dots, \lambda_n \le L_n - t. \quad (16)$$

Since $t \ge c$, it is clear that $(\lambda_1, \dots, \lambda_n)^{\mathrm{T}}$ will also satisfy the required (15).

Proposition 1. *Define $s \in \mathbb{R}$ by the requirement*

$$\sum_{i=1}^{n} (s - L_i)^+ = 2. \quad (17)$$

The unique solution to the optimization problem $t \to \max$ under the constraints (16) with $\lambda_1, \dots, \lambda_n$ as in (14) will be

$$u^i = \frac{(s - L_i)^+}{2}, \ \ i = 1, \dots, n, \quad (18)$$

$$t = s - 1 - (u^1)^2 - \dots - (u^n)^2. \quad (19)$$

There exists a unique s satisfying (17) since the left-hand side of (17) is a continuous, increasing (strictly increasing when positive) and unbounded above function of s. The substitution function is given by (18).

Proof of Proposition 1. Let us denote the u^i and t defined by (18) and (19) as $\overline{u}^i$ and $\overline{t}$, respectively. To see that they satisfy the constraints (16), notice that the ith constraint can be spelt out as

$$(\overline{u}^1)^2 + \dots + (\overline{u}^n)^2 - 2\overline{u}^i + 1 \le L_i - \overline{t},$$

which immediately follows from (18) and (19). As a by-product, we can see that the inequality becomes an equality, i.e.,

$$\overline{t} = L_i - 1 + 2\overline{u}^i - (\overline{u}^1)^2 - \dots - (\overline{u}^n)^2, \quad (20)$$

for all i with $\overline{u}^i > 0$.

We can rewrite (16) as

$$\begin{cases} t \le L_1 - 1 + 2u^1 - (u^1)^2 - \dots - (u^n)^2, \\ \qquad\vdots \\ t \le L_n - 1 + 2u^n - (u^1)^2 - \dots - (u^n)^2, \end{cases} \quad (21)$$

and our goal is to prove that these inequalities imply $t < \bar{t}$ (unless $u^1 = \overline{u}^1, \ldots, u^n = \overline{u}^n$). Choose $\overline{u}^i$ (necessarily $\overline{u}^i > 0$ unless $u^1 = \overline{u}^1, \ldots, u^n = \overline{u}^n$; in the latter case, however, we can, and will, also choose $\overline{u}^i > 0$) for which $\epsilon_i := \overline{u}^i - u^i$ is maximal. Then every value of t satisfying (21) will also satisfy

$$\begin{aligned} t &\le L_i - 1 + 2u^i - \sum_{j=1}^n (u^j)^2 \\ &= L_i - 1 + 2\overline{u}^i - 2\epsilon_i - \sum_{j=1}^n (\overline{u}^j)^2 + 2\sum_{j=1}^n \epsilon_j \overline{u}^j - \sum_{j=1}^n \epsilon_j^2 \\ &\le L_i - 1 + 2\overline{u}^i - \sum_{j=1}^n (\overline{u}^j)^2 - \sum_{j=1}^n \epsilon_j^2 \le \bar{t}, \end{aligned}$$

with the last $\le$ following from (20) and becoming $<$ when not all u^j coincide with $\overline{u}^j$. □

The detailed description of the resulting prediction algorithm was given as Algorithm 1 in Section 2. As discussed, that algorithm uses the generalized prediction $G_N(\omega)$ computed from unnormalized weights.

6. Conclusion

In this paper we only considered the simplest prediction problem for the Brier game: competing with a finite pool of experts. In the case of square-loss regression, it is possible to find efficient closed-form prediction algorithms competitive with linear functions (see, e.g., Cesa-Bianchi & Lugosi, 2006, Chapter 11). Such algorithms can often be "kernelized" to obtain prediction algorithms competitive with reproducing kernel Hilbert spaces of prediction rules. This would be an appealing research programme in the case of the Brier game as well.

Acknowledgments

We are grateful to Football-Data and Tennis-Data for providing access to the data used in this paper. This work was partly supported by EPSRC (grant EP/F002998/1). Comments by Alexey Chernov, Alex Gammerman, Yuri Kalnishkan, and anonymous referees have helped us improve the presentation.

References

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

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., & Warmuth, M. K. (1997). How to use expert advice. *Journal of the Association for Computing Machinery*, *44*, 427–485.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge, England: Cambridge University Press.

Dawid, A. P. (1986). Probability forecasting. In S. Kotz, N. L. Johnson and C. B. Read (Eds.), *Encyclopedia of statistical sciences*, vol. 7, 210–218. New York: Wiley.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Haussler, D., Kivinen, J., & Warmuth, M. K. (1998). Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, *44*, 1906–1925.

Kalnishkan, Y., & Vyugin, M. V. (2005). The Weak Aggregating Algorithm and weak mixability. *Proceedings of the Eighteenth Annual Conference on Learning Theory* (pp. 188–203). Berlin: Springer.

Kivinen, J., & Warmuth, M. K. (1999). Averaging expert predictions. *Proceedings of the Fourth European Conference on Computational Learning Theory* (pp. 153–167). Berlin: Springer.

Littlestone, N., & Warmuth, M. K. (1994). The Weighted Majority Algorithm. *Information and Computation*, *108*, 212–261.

Thorpe, J. A. (1979). *Elementary topics in differential geometry.* New York: Springer.

Vovk, V. (1990). Aggregating strategies. *Proceedings of the Third Annual Workshop on Computational Learning Theory* (pp. 371–383). San Mateo, CA: Morgan Kaufmann.

Vovk, V. (1998). A game of prediction with expert advice. *Journal of Computer and System Sciences*, *56*, 153–173.

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review*, *69*, 213–248.

Vovk, V. (2007). *Defensive forecasting for optimal prediction with expert advice* (Technical Report `arXiv:0708.1503` [cs.LG]). `arXiv.org` e-Print archive.

Vovk, V., & Zhdanov, F. (2008). *Prediction with expert advice for the Brier game* (Technical Report `arXiv:0708.2502v2` [cs.LG]). `arXiv.org` e-Print archive.