Journal of Machine Learning Research 10 (2009) 2445-2471 Submitted 6/09; Published 11/09

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, England

Editor: Yoav Freund

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, with well-known bookmakers playing the role of experts. The theoretical performance guarantee is not excessively loose on the football data set and is rather tight on the tennis data set.

Keywords: Brier game, classification, on-line prediction, strong aggregating algorithm, weighted average algorithm

1. Introduction

The paradigm of prediction with expert advice was introduced in the late 1980s (see, e.g., DeSantis et al., 1988, Littlestone and 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 and Vyugin, 2008).

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 (see, e.g., the review by Vovk, 2001, Section 2.5).

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 shortly, 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. Its short conference version (Vovk and Zhdanov, 2008a) appeared in the ICML 2008 proceedings.

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

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

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 $\gamma\{3\} = 1/4$, $\lambda(\omega,\gamma) = (1/2-1)^2 + (1/4-0)^2 + (1/4-0)^2 = 3/8$.)

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 (see Algorithm 1). For each expert k, the algorithm maintains its weight w^k, constantly slashing the weights of less successful experts. Its description uses the notation $t^+ := \max(t,0)$.

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.

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
```

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

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

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

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

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

3. Experimental Results

In our first empirical study of Algorithm 1 we use historical data about 8999 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 four seasons, 2005/2006, 2006/2007, 2007/2008, and 2008/2009. The matches are sorted first by date, then by league, and then by the name of the home team. 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 $a_i > 1$ is the total amount which the bookmaker undertakes to pay out to a client betting on outcome i per unit stake in the event that i happens (if the bookmaker wishes to return the stake to the bettor, it should be included in a_i; 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 use Victor Khutsishvili's (2009) formula

$$p_i := a_i^{-\gamma}, \quad i = 1,2,3, \tag{3}$$

for computing the bookmaker's probability forecasts, where $\gamma > 0$ is chosen such that $a_1^{-\gamma} + a_2^{-\gamma} + a_3^{-\gamma} = 1$. Such a value of γ exists and is unique since the function $a_1^{-\gamma} + a_2^{-\gamma} + a_3^{-\gamma}$ continuously and strictly decreases from 3 to 0 as γ changes from 0 to ∞. In practice, we usually have $\gamma > 1$ as $a_1^{-1} + a_2^{-1} + a_3^{-1} > 1$ (i.e., the overround is positive). The method of bisection was more than sufficient for us to solve $a_1^{-\gamma} + a_2^{-\gamma} + a_3^{-\gamma} = 1$ with satisfactory accuracy. Khutsishvili's argument for (3) is outlined in Appendix B.

Typical values of γ in (3) are close to 1, and the difference $\gamma - 1$ reflects the bookmaker's target profit margin. In this respect $\gamma - 1$ is similar to the overround; indeed, the approximate value of the overround is $(\gamma - 1)\sum_{i=1}^{3} a_i^{-1} \ln a_i$ assuming that the overround is small and none of a_i is too close to 0. The coefficient of proportionality $\sum_{i=1}^{3} a_i^{-1} \ln a_i$ can be interpreted as the entropy of the quoted betting odds.

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, $k = 1, \ldots, 8$, 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 the first part of Theorem 1 (Equation (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 time).

Figure 2 shows the distribution of the bookmakers' overrounds. We can see that in most cases overrounds are between 0.05 and 0.15, but there are also occasional extreme values, near zero or in excess of 0.3.

Figure 3 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, for example, Australian Open, French Open, US Open, and Wimbledon; the data is provided by Tennis-Data. The matches are sorted by date, then by tournament, and then by the winner's name. 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 3 are presented in the same way as in Figure 1.

Typical values of the overround are below 0.1, as shown in Figure 4 (analogous to Figure 2).

In both Figure 1 and Figure 3 the cumulative loss of Algorithm 1 is close to the cumulative loss of the best expert. The theoretical bound is not hopelessly loose for the football data and is rather tight for the tennis data. The pictures look almost the same when Algorithm 1 is applied in the more realistic manner where the experts' weights w^k are not updated over the matches that are played simultaneously.

Our second empirical study (Figure 3) 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.

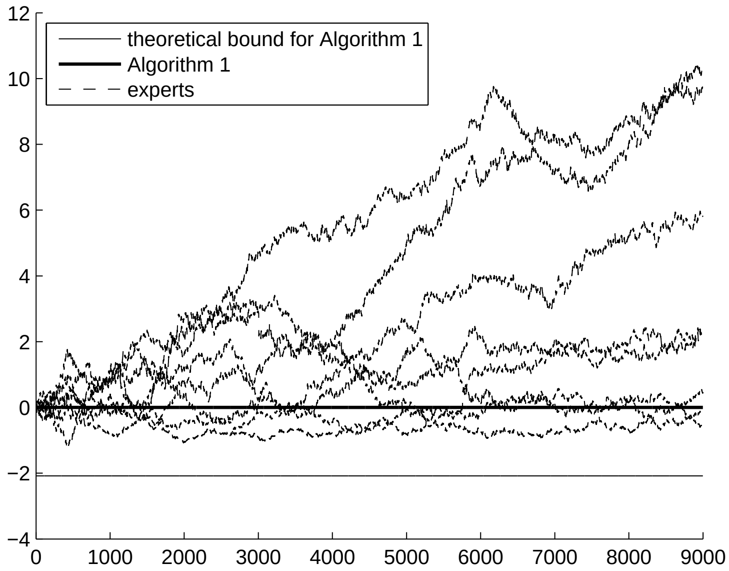

Figure 1: The difference between the cumulative loss of each of the 8 bookmakers (experts) and of Algorithm 1 on the football data. The theoretical lower bound $-\ln 8$ from Theorem 1 is also shown.

For comparison with several other popular prediction algorithms, see Appendix C. The data used for producing all the figures and tables in this section and in Appendix C 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, for example, 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

$$\Phi_\eta(f) : \omega \in \Omega \mapsto e^{-\eta f(\omega)}, \quad f \in \mathbb{R}^\Omega. \tag{4}$$

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

$$L_N \leq \min_{k=1,\dots,K} L_N^k + \frac{\ln K}{\eta}, \quad N = 1,2,\dots,$$

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 Appendix A). Comparing this with (1) and (2) we can see that we are required to prove that

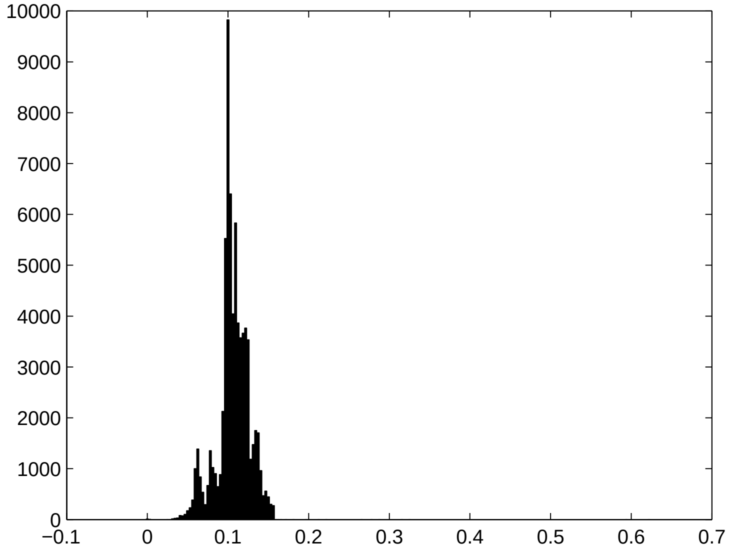

Figure 2: The overround distribution histogram for the football data, with 200 bins of equal size between the minimum and maximum values of the overround.

- $\Phi_\eta(\Sigma)$ is convex when $\eta \le 1$;
- $\Phi_\eta(\Sigma)$ 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

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} e^{-\eta((u_1-1)^2+u_2^2+\cdots+u_n^2)} \\ e^{-\eta(u_1^2+(u_2-1)^2+\cdots+u_n^2)} \\ \vdots \\ e^{-\eta(u_1^2+\cdots+(u_{n-1}-1)^2+u_n^2)} \\ e^{-\eta(u_1^2+\cdots+u_{n-1}^2+(u_n-1)^2)} \end{pmatrix}, \tag{5}$$

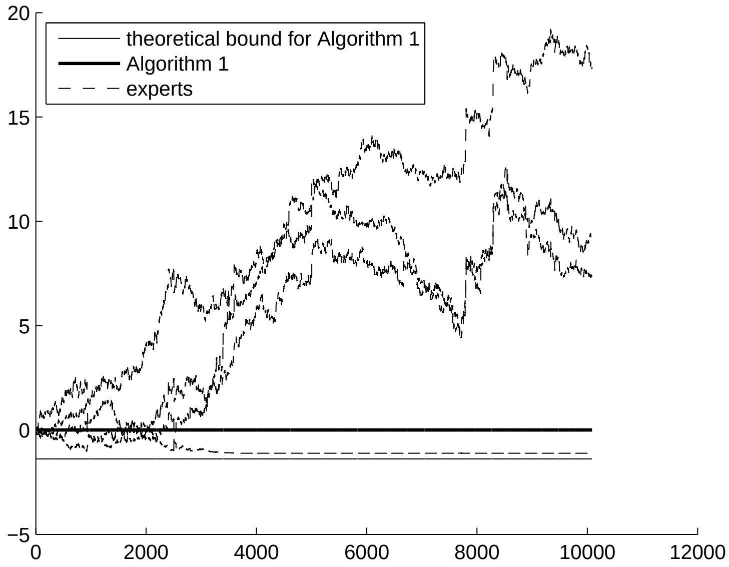

Figure 3: The difference between the cumulative loss of each of the 4 bookmakers and of Algorithm 1 on the tennis data. Now the theoretical bound is $-\ln 4$.

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 (5) in u_1 is

$$\frac{\partial}{\partial u_1}\begin{pmatrix}x_1\\x_2\\\vdots\\x_{n-1}\\x_n\end{pmatrix}=2\eta\begin{pmatrix}(u_n-u_1+1)e^{-\eta((u_1-1)^2+u_2^2+\cdots+u_{n-1}^2+u_n^2)}\\(u_n-u_1)e^{-\eta(u_1^2+(u_2-1)^2+\cdots+u_{n-1}^2+u_n^2)}\\\vdots\\(u_n-u_1)e^{-\eta(u_1^2+u_2^2+\cdots+(u_{n-1}-1)^2+u_n^2)}\\(u_n-u_1-1)e^{-\eta(u_1^2+u_2^2+\cdots+u_{n-1}^2+(u_n-1)^2)}\end{pmatrix}\propto\begin{pmatrix}(u_n-u_1+1)e^{2\eta u_1}\\(u_n-u_1)e^{2\eta u_2}\\\vdots\\(u_n-u_1)e^{2\eta u_{n-1}}\\(u_n-u_1-1)e^{2\eta u_n}\end{pmatrix},$$

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.

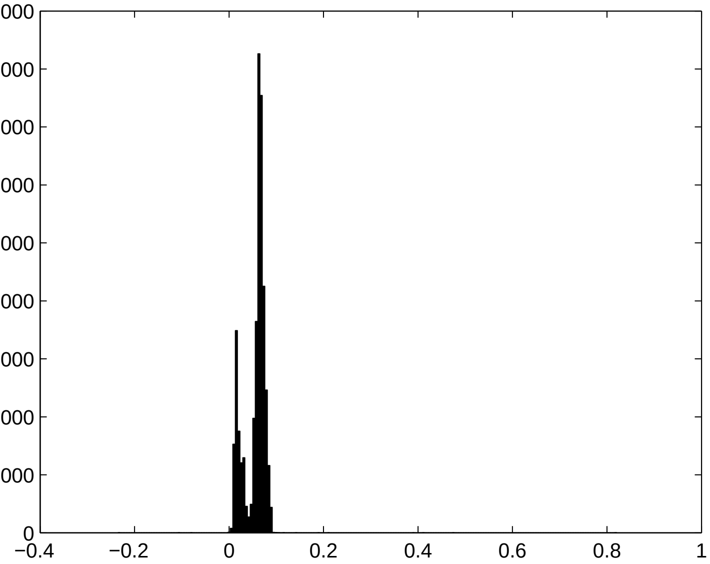

Figure 4: The overround distribution histogram for the tennis data.

A normal vector to the surface can be found as

$$Z := \begin{vmatrix} e_1 & \cdots & e_{n-1} & e_n \\ (u_n-u_1+1)e^{2\eta u_1} & \cdots & (u_n-u_1)e^{2\eta u_{n-1}} & (u_n-u_1-1)e^{2\eta u_n} \\ \vdots & \ddots & \vdots & \vdots \\ (u_n-u_{n-1})e^{2\eta u_1} & \cdots & (u_n-u_{n-1}+1)e^{2\eta u_{n-1}} & (u_n-u_{n-1}-1)e^{2\eta u_n} \end{vmatrix},$$

where e_i is the ith vector in the standard basis of $\mathbb{R}^n$ and $|\cdot|$ stands for the determinant (the matrix contains both scalars and vectors, but its determinant can still be computed using the standard rules). The coefficient in front of e_1 is the $(n-1)\times(n-1)$ determinant

$$\begin{vmatrix} (u_n-u_1)e^{2\eta u_2} & \cdots & (u_n-u_1)e^{2\eta u_{n-1}} & (u_n-u_1-1)e^{2\eta u_n} \\ (u_n-u_2+1)e^{2\eta u_2} & \cdots & (u_n-u_2)e^{2\eta u_{n-1}} & (u_n-u_2-1)e^{2\eta u_n} \\ \vdots & \ddots & \vdots & \vdots \\ (u_n-u_{n-1})e^{2\eta u_2} & \cdots & (u_n-u_{n-1}+1)e^{2\eta u_{n-1}} & (u_n-u_{n-1}-1)e^{2\eta u_n} \end{vmatrix}$$

$$\propto e^{-2\eta u_1} \begin{vmatrix} u_n-u_1 & \cdots & u_n-u_1 & u_n-u_1-1 \\ u_n-u_2+1 & \cdots & u_n-u_2 & u_n-u_2-1 \\ \vdots & \ddots & \vdots & \vdots \\ u_n-u_{n-1} & \cdots & u_n-u_{n-1}+1 & u_n-u_{n-1}-1 \end{vmatrix}$$

$$= e^{-2\eta u_1} \begin{vmatrix} 1 & 1 & \cdots & 1 & u_n-u_1-1 \\ 2 & 1 & \cdots & 1 & u_n-u_2-1 \\ 1 & 2 & \cdots & 1 & u_n-u_3-1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 2 & u_n-u_{n-1}-1 \end{vmatrix} = e^{-2\eta u_1} \begin{vmatrix} 1 & 1 & \cdots & 1 & u_n-u_1-1 \\ 1 & 0 & \cdots & 0 & u_1-u_2 \\ 0 & 1 & \cdots & 0 & u_1-u_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & u_1-u_{n-1} \end{vmatrix}$$

$$\begin{aligned}
&= e^{-2\eta u_1}\bigl((-1)^n(u_n-u_1-1)+(-1)^{n+1}(u_1-u_2)\\
&\quad+(-1)^{n+1}(u_1-u_3)+\cdots+(-1)^{n+1}(u_1-u_{n-1})\bigr)\\
&= e^{-2\eta u_1}(-1)^n\left((u_2+u_3+\cdots+u_n)-(n-1)u_1-1\right)\\
&\qquad\qquad = -e^{-2\eta u_1}(-1)^n n u_1 \propto u_1 e^{-2\eta u_1}
\end{aligned}\tag{6}$$

(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 $i = 2,\ldots,n-1$; indeed, the $(n-1)\times(n-1)$ determinant representing the coefficient in front of e_i can be reduced to the form analogous to (6) by moving the ith row to the top.

The coefficient in front of e_n is proportional to

$$e^{-2\eta u_n}\begin{vmatrix}
u_n-u_1+1 & u_n-u_1 & \cdots & u_n-u_1 & u_n-u_1\\
u_n-u_2 & u_n-u_2+1 & \cdots & u_n-u_2 & u_n-u_2\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
u_n-u_{n-2} & u_n-u_{n-2} & \cdots & u_n-u_{n-2}+1 & u_n-u_{n-2}\\
u_n-u_{n-1} & u_n-u_{n-1} & \cdots & u_n-u_{n-1} & u_n-u_{n-1}+1
\end{vmatrix}$$

$$= e^{-2\eta u_n}\begin{vmatrix}
1 & 0 & \cdots & 0 & u_n-u_1\\
0 & 1 & \cdots & 0 & u_n-u_2\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & \cdots & 1 & u_n-u_{n-2}\\
-1 & -1 & \cdots & -1 & u_n-u_{n-1}+1
\end{vmatrix}
= e^{-2\eta u_n}\begin{vmatrix}
1 & 0 & \cdots & 0 & u_n-u_1\\
0 & 1 & \cdots & 0 & u_n-u_2\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & \cdots & 1 & u_n-u_{n-2}\\
0 & 0 & \cdots & 0 & nu_n
\end{vmatrix}$$

$$= nu_n e^{-2\eta u_n}$$

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

The Gauss-Kronecker curvature at the point with coordinates $(u_1,\ldots,u_{n-1})$ is proportional (with a positive coefficient of proportionality, possibly depending on the point) to

$$\begin{vmatrix}
\frac{\partial Z^{\mathrm{T}}}{\partial u_1}\\
\vdots\\
\frac{\partial Z^{\mathrm{T}}}{\partial u_{n-1}}\\
Z^{\mathrm{T}}
\end{vmatrix}\tag{7}$$

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

A straightforward calculation allows us to rewrite determinant (7) (ignoring the positive coefficient $((-1)^{n-1}ne^{2\eta})^n$) as

$$\begin{vmatrix}
(1-2\eta u_1)e^{-2\eta u_1} & 0 & \cdots & 0 & (2\eta u_n-1)e^{-2\eta u_n}\\
0 & (1-2\eta u_2)e^{-2\eta u_2} & \cdots & 0 & (2\eta u_n-1)e^{-2\eta u_n}\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & \cdots & (1-2\eta u_{n-1})e^{-2\eta u_{n-1}} & (2\eta u_n-1)e^{-2\eta u_n}\\
u_1e^{-2\eta u_1} & u_2e^{-2\eta u_2} & \cdots & u_{n-1}e^{-2\eta u_{n-1}} & u_ne^{-2\eta u_n}
\end{vmatrix}$$

$$\propto \begin{vmatrix} 1-2\eta u_1 & 0 & \cdots & 0 & 2\eta u_n - 1 \\ 0 & 1-2\eta u_2 & \cdots & 0 & 2\eta u_n - 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-2\eta u_{n-1} & 2\eta u_n - 1 \\ u_1 & u_2 & \cdots & u_{n-1} & u_n \end{vmatrix}$$

$$= u_1(1-2\eta u_2)(1-2\eta u_3)\cdots(1-2\eta u_n) + u_2(1-2\eta u_1)(1-2\eta u_3)\cdots(1-2\eta u_n) + \cdots \\ + u_n(1-2\eta u_1)(1-2\eta u_2)\cdots(1-2\eta u_{n-1}) \quad (8)$$

(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 next goal is to show that the last expression in (8) 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 (8) becomes negative. It will remain negative if u_1 and u_2 are sufficiently close to $1/2$ and $u_3, \ldots, u_n$ are sufficiently close to 0.

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

$$\begin{gathered} -1 < 1 - 2\eta < t_i < 1, \quad i = 1, \ldots, n, \\ t_1 + \cdots + t_n = n - 2\eta > n - 2. \end{gathered} \quad (9)$$

Our goal is to prove

$$(1-t_1)t_2 t_3 \cdots t_n + \cdots + (1-t_n)t_1 t_2 \cdots t_{n-1} > 0,$$

that is,

$$t_2 t_3 \cdots t_n + \cdots + t_1 t_2 \cdots t_{n-1} > n t_1 \cdots t_n. \quad (10)$$

This reduces to

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} > n \quad (11)$$

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

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} < n \quad (12)$$

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 (9) 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 (10) is indeed true.

Let us prove (11). 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. (9)). Therefore,

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

To establish (10) it remains to prove (12). Suppose, without loss of generality, that $t_1 > 0$, $t_2 > 0, \ldots, t_{n-1} > 0$, and $t_n < 0$. We will prove a slightly stronger statement allowing $t_1, \ldots, t_{n-2}$ to take value 1 and removing the lower bound on t_n. 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 in the case $\eta \le 1$, and the lack of positivity of the Gauss-Kronecker curvature implies the lack of convexity of the η-exponential superprediction set in the case $\eta > 1$. The η-exponential superprediction surface will be oriented by choosing the normal vector field directed towards the origin. This can be done since

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \propto \begin{pmatrix} e^{2\eta u_1} \\ \vdots \\ e^{2\eta u_n} \end{pmatrix}, \quad Z \propto (-1)^{n-1} \begin{pmatrix} u_1 e^{-2\eta u_1} \\ \vdots \\ u_n e^{-2\eta u_n} \end{pmatrix}, \tag{13}$$

with both coefficients of proportionality positive (cf. (5) and the bottom row of the first determinant in (8)), and the sign of the scalar product of the two vectors on the right-hand sides in (13) does not depend on the point $(u_1,\dots,u_{n-1})$. Namely, we take $(-1)^n Z$ as the normal vector field directed towards the origin. The Gauss-Kronecker curvature will not change sign after the re-orientation: if n is even, the new orientation coincides with the old, and for odd n the Gauss-Kronecker curvature does not depend on the orientation.

In the case $\eta > 1$, the Gauss-Kronecker curvature is negative at some point, and so the η-exponential superprediction set is not convex (Thorpe, 1979, Chapter 13, Theorem 1 and its proof).

It remains to consider the case $\eta \le 1$. Because of the continuity of the η-exponential superprediction surface in η we can and will assume, without loss of generality, that $\eta < 1$.

Let us first check that the smallest principal curvature $k_1 = k_1(u_1,\dots,u_{n-1},\eta)$ of the η-exponential superprediction surface is always positive (among the arguments of k_1 we list not only the coordinates $u_1,\dots,u_{n-1}$ of a point on the surface (5) but also the learning rate $\eta \in (0,1)$). At least at some $(u_1,\dots,u_{n-1},\eta)$ the value of $k_1(u_1,\dots,u_{n-1},\eta)$ is positive: take a sufficiently small η and the point on the surface (5) with coordinates $u_1 = \cdots = u_{n-1} = 1/n$; a simple calculation shows that this point will be a point of local maximum for $x_1+\cdots+x_n$. 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

$$\left\{(u_1,\dots,u_{n-1},\eta) \,|\, u_1 \in (0,1),\dots,u_{n-1} \in (0,1), u_1+\cdots+u_{n-1} < 1, \eta \in (0,1)\right\}, \tag{14}$$

we could connect these points by a continuous curve lying completely inside (14); 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 some point between A and B will be negative (remember that the curve is oriented by directing the normal vector field towards the origin), contradicting the positivity of k_1 at that point.

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 $\eta \leq 1$, 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 $\eta \leq 1$), we are only required to find a permitted prediction

$$\begin{pmatrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_n\end{pmatrix} = \begin{pmatrix}(u_1-1)^2+u_2^2+\cdots+u_n^2\\ u_1^2+(u_2-1)^2+\cdots+u_n^2\\ \vdots\\ u_1^2+u_2^2+\cdots+(u_n-1)^2\end{pmatrix} \tag{15}$$

(cf. (5)) satisfying

$$\lambda_1 \leq l_1,\dots,\lambda_n \leq l_n. \tag{16}$$

Now suppose we are given a generalized prediction $(L_1,\dots,L_n)^{\mathrm{T}}$ computed by the aggregating pseudo-algorithm 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 (15) satisfying (16) 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 (15) satisfying

$$\lambda_1 \leq L_1-t,\dots,\lambda_n \leq L_n-t. \tag{17}$$

Since $t \geq c$, it is clear that $(\lambda_1,\dots,\lambda_n)^{\mathrm{T}}$ will also satisfy the required (16).

Proposition 2 *Define $s \in \mathbb{R}$ by the requirement*

$$\sum_{i=1}^n (s-L_i)^+ = 2. \tag{18}$$

The unique solution to the optimization problem $t \to \max$ under the constraints (17) with $\lambda_1,\dots,\lambda_n$ as in (15) will be

$$u_i = \frac{(s-L_i)^+}{2}, \quad i=1,\dots,n, \tag{19}$$

$$t = s-1-u_1^2-\cdots-u_n^2. \tag{20}$$

There exists a unique s satisfying (18) since the left-hand side of (18) is a continuous, increasing (strictly increasing when positive) and unbounded above function of s. The substitution function is given by (19).

Proof of Proposition 2 Let us denote the u_i and t defined by (19) and (20) as $\overline{u}_i$ and $\overline{t}$, respectively. To see that they satisfy the constraints (17), notice that the ith constraint can be spelt out as

$$\overline{u}_1^2 + \cdots + \overline{u}_n^2 - 2\overline{u}_i + 1 \le L_i - \overline{t},$$

which immediately follows from (19) and (20). As a by-product, we can see that the inequality becomes an equality, that is,

$$\overline{t} = L_i - 1 + 2\overline{u}_i - \overline{u}_1^2 - \cdots - \overline{u}_n^2, \tag{21}$$

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

We can rewrite (17) as

$$\begin{cases} t \le L_1 - 1 + 2u_1 - u_1^2 - \cdots - u_n^2, \\ \vdots \\ t \le L_n - 1 + 2u_n - u_1^2 - \cdots - u_n^2, \end{cases} \tag{22}$$

and our goal is to prove that these inequalities imply $t < \overline{t}$ (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 $\varepsilon_i := \overline{u}_i - u_i$ is maximal. Then every value of t satisfying (22) will also satisfy

$$\begin{aligned} t &\le L_i - 1 + 2u_i - \sum_{j=1}^n u_j^2 \\ &= L_i - 1 + 2\overline{u}_i - 2\varepsilon_i - \sum_{j=1}^n \overline{u}_j^2 + 2\sum_{j=1}^n \varepsilon_j \overline{u}_j - \sum_{j=1}^n \varepsilon_j^2 \\ &\le L_i - 1 + 2\overline{u}_i - \sum_{j=1}^n \overline{u}_j^2 - \sum_{j=1}^n \varepsilon_j^2 \le \overline{t}. \end{aligned} \tag{23}$$

The penultimate $\le$ in (23) follows from

$$-\varepsilon_i + \sum_{j=1}^n \varepsilon_j \overline{u}_j = \sum_{j=1}^n (\varepsilon_j - \varepsilon_i)\overline{u}_j \le 0.$$

The last $\le$ in (23) follows from (21) and becomes $<$ 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 and 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

Victor Khutsishvili has been extremely generous in sharing his ideas with us. Comments by Alexey Chernov, Yuri Kalnishkan, Alex Gammerman, Bob Vickers, and the anonymous referees for the conference and journal versions have helped us improve the presentation. The referees for the conference version also suggested comparing our results with the Weighted Average Algorithm and the Hedge algorithm. We are grateful to Football-Data and Tennis-Data for providing access to the data used in this paper. This work was supported in part by EPSRC (grant EP/F002998/1).

Appendix A. Watkins's Theorem

Watkins's theorem is stated in Vovk (1999, Theorem 8) not in sufficient generality: it presupposes that the loss function is mixable. The proof, however, shows that this assumption is irrelevant (it can be made part of the conclusion), and the goal of this appendix is to give a self-contained statement of a suitable version of the theorem. (The reader will notice that the generality of the new version is essential only for our discussion in Section 4, not for Theorem 1 itself.)

In this appendix we will use a slightly more general notion of a game of prediction (Ω,Γ,λ): namely, the loss function $\lambda : \Omega\times\Gamma \to \overline{\mathbb{R}}$ is now allowed to take values in the extended real line $\overline{\mathbb{R}} := \mathbb{R}\cup\{-\infty,\infty\}$ (although the value $-\infty$ will be later disallowed).

Partly following Vovk (1998), for each $K = 1,2,\dots$ and each $a > 0$ we consider the following perfect-information game $\mathcal{G}_K(a)$ (the "global game") between two players, Learner and Environment. Environment is a team of $K+1$ players called Expert 1 to Expert K and Reality, who play with Learner according to Protocol 1. Learner wins if, for all $N = 1,2,\dots$ and all $k \in \{1,\dots,K\}$,

$$L_N \le L_N^k + a; \tag{24}$$

otherwise, Environment wins. It is possible that $L_N = \infty$ or $L_N^k = \infty$ in (24); the interpretation of inequalities involving infinities is natural.

For each K we will be interested in the set of those $a > 0$ for which Learner has a winning strategy in the game $\mathcal{G}_K(a)$ (we will denote this by $\mathrm{L} \smile \mathcal{G}_K(a)$). It is obvious that

$$\mathrm{L} \smile \mathcal{G}_K(a) \;\&\; a' > a \Longrightarrow \mathrm{L} \smile \mathcal{G}_K(a');$$

therefore, for each K there exists a unique *borderline value* a_K such that $\mathrm{L} \smile \mathcal{G}_K(a)$ holds when $a > a_K$ and fails when $a < a_K$. It is possible that $a_K = \infty$ (but remember that we are only interested in finite values of a).

These are our assumptions about the game of prediction (similar to those in Vovk, 1998):

- Γ is a compact topological space;
- for each $\omega \in \Omega$, the function $\gamma \in \Gamma \mapsto \lambda(\omega,\gamma)$ is continuous ($\overline{\mathbb{R}}$ is equipped with the standard topology);
- there exists $\gamma \in \Gamma$ such that, for all $\omega \in \Omega$, $\lambda(\omega,\gamma) < \infty$;
- the function λ is bounded below.

We say that the game of prediction (Ω,Γ,λ) is η-*mixable*, where $\eta > 0$, if

$$\forall\gamma_1 \in \Gamma, \gamma_2 \in \Gamma, \alpha \in [0,1]\ \exists\delta \in \Gamma\ \forall\omega \in \Omega\colon e^{-\eta\lambda(\omega,\delta)} \geq \alpha e^{-\eta\lambda(\omega,\gamma_1)} + (1-\alpha)e^{-\eta\lambda(\omega,\gamma_2)}. \qquad (25)$$

In the case of finite Ω, this condition says that the image of the superprediction set under the mapping Φ_η (see (4)) is convex. The game of prediction is *mixable* if it is η-mixable for some $\eta > 0$.

It follows from Hardy et al. (1952, Theorem 92, applied to the means $\mathfrak{M}_\phi$ with $\phi(x) = e^{-\eta x}$) that if the prediction game is η-mixable it will remain η'-mixable for any positive $\eta' < \eta$. (For another proof, see the end of the proof of Lemma 9 in Vovk, 1998.) Let η^* be the supremum of the η for which the prediction game is η-mixable (with $\eta^* := 0$ when the game is not mixable). The compactness of Γ implies that the prediction game is η^*-mixable.

Theorem 3 (Chris Watkins) *For any $K \in \{2,3,\ldots\}$,*

$$a_K = \frac{\ln K}{\eta^*}.$$

In particular, $a_K < \infty$ if and only if the game is mixable.

The theorem does not say explicitly, but it is easy to check, that L $\smile$ $\mathcal{G}_K(a_K)$: this follows both from general considerations (cf. Lemma 3 in Vovk, 1998) and from the fact that the strong aggregating algorithm wins $\mathcal{G}_K(a_K) = \mathcal{G}_K(\ln K/\eta^*)$.

Proof of Theorem 3 The proof will use some notions and notation used in the statement and proof of Theorem 1 of Vovk (1998). Without loss of generality we can, and will, assume that the loss function satisfies $\lambda > 1$ (add a suitable constant to λ if needed). Therefore, Assumption 4 of Vovk (1998) (the only assumption in that paper not directly made here) is satisfied. In view of the fact that L $\smile$ $\mathcal{G}_K(\ln K/\eta^*)$, we only need to show that L $\smile$ $\mathcal{G}_K(a)$ does not hold for $a < \ln K/\eta^*$. Fix $a < \ln K/\eta^*$.

The separation curve consists of the points $(c(\beta), c(\beta)/\eta) \in [0,\infty)^2$, where $\beta := e^{-\eta}$ and η ranges over $[0,\infty]$ (see Vovk, 1998, Theorem 1). Since the two-fold convex mixture in (25) can be replaced by any finite convex mixture (apply two-fold mixtures repeatedly), setting $\eta := \eta^*$ shows that the point $(1, 1/\eta^*)$ is Northeast of (actually belongs to) the separation curve. On the other hand, the point $(1, a/\ln K)$ is Southwest and outside of the separation curve (use Lemmas 8–12 of Vovk, 1998). Therefore, E (i.e., Environment) has a winning strategy in the game $\mathcal{G}(1, a/\ln K)$. It is easy to see from the proof of Theorem 1 in Vovk (1998) that the definition of the game $\mathcal{G}$ can be modified, without changing the conclusion about $\mathcal{G}(1, a/\ln K)$, by replacing the line

E chooses $n \geq 1$ {size of the pool}

in the protocol on p. 153 of Vovk (1998) by

E chooses $n^* \geq 1$ {lower bound on the size of the pool}
L chooses $n \geq n^*$ {size of the pool}

(indeed, the proof in Section 6 of Vovk, 1998, only requires that there should be sufficiently many experts). Let n^* be the first move by Environment according to her winning strategy.

Now suppose L $\smile$ $\mathcal{G}_K(a)$. From the fact that there exists Learner's strategy $\mathcal{L}_1$ winning $\mathcal{G}_K(a)$ we can deduce: there exists Learner's strategy $\mathcal{L}_2$ winning $\mathcal{G}_{K^2}(2a)$ (we can split the K^2 experts into K groups of K, merge the experts' decisions in each group with $\mathcal{L}_1$, and finally merge the groups' decisions with $\mathcal{L}_1$); there exists Learner's strategy $\mathcal{L}_3$ winning $\mathcal{G}_{K^3}(3a)$ (we can split the K^3 experts

Loss resulting from (3)	Loss resulting from (26)	Difference
5585.69	5588.20	2.52
5585.94	5586.67	0.72
5586.60	5587.37	0.77
5588.47	5590.65	2.18
5588.61	5589.92	1.31
5591.97	5593.48	1.52
5596.01	5601.85	5.84
5596.56	5598.02	1.46

Table 1: The bookmakers' cumulative Brier losses over the football data set when their probability forecasts are computed using formula (3) and formula (26).

into K groups of K^2, merge the experts' decisions in each group with $\mathcal{L}_2$, and finally merge the groups' decisions with $\mathcal{L}_1$); and so on. When the number K^m of experts exceeds n^*, we obtain a contradiction: Learner can guarantee

$$L_N \leq L_N^k + ma$$

for all N and all K^m experts k, and Environment can guarantee that

$$L_N > L_N^k + \frac{a}{\ln K} \ln(K^m) = L_N^k + ma$$

for some N and k. ■

Appendix B. Khutsishvili's Theory

In the conference version of this paper (Vovk and Zhdanov, 2008a) we used

$$p_i := \frac{1/a_i}{1/a_1 + 1/a_2 + 1/a_3}, \quad i = 1, 2, 3, \tag{26}$$

in place of (3). A natural way to compare formulas (3) and (26) is to compare the losses of the probability forecasts found from the bookmakers' betting odds using those formulas. Using Khutsishvili's formula (3) consistently leads to smaller losses as measured by the Brier loss function: see Tables 1 and 2. The improvement of each bookmaker's total loss over the football data set is in the range 0.72–5.84; over the tennis data set the difference is in the range 1.27–11.64. These differences are of the order of the differences in cumulative loss between different bookmakers, and so the improvement is significant.

The goal of this appendix is to present, in a rudimentary form, Khutsishvili's theory behind (3). The theory is based on a very idealized model of a bookmaker, who is assumed to compute the betting odds a for an event of probability p using a function f,

$$a := f(p).$$

Loss resulting from (3)	Loss resulting from (26)	Difference
3935.32	3944.02	8.69
3943.83	3945.10	1.27
3945.70	3957.33	11.64
3953.83	3957.75	3.92

Table 2: The bookmakers' cumulative Brier losses over the tennis data set when their probability forecasts are computed using formula (3) and formula (26).

Different bookmakers (and the same bookmaker at different times) can use different functions f. Therefore, different bookmakers may quote different odds because they may use different f and because they may assign different probabilities to the same event.

The following simple corollary of Darboux's theorem describes the set of possible functions f; its interpretation will be discussed straight after the proof.

Theorem 4 (Victor Khutsishvili) *Suppose a function* $f:(0,1)\to(1,\infty)$ *satisfies the condition*

$$f(pq)=f(p)f(q) \tag{27}$$

for all $p,q\in(0,1)$. *There exists* $c>0$ *such that* $f(p)=p^{-c}$ *for all* $p\in(0,1)$.

Proof Equation (27) is one of the four fundamental Cauchy equations, which can be easily reduced to each other. For example, introducing a new function $g:(0,\infty)\to(0,\infty)$ by $g(u):=\ln f(e^{-u})$ and new variables $x,y\in(0,\infty)$ by $x:=-\ln p$ and $y:=-\ln q$, we transform (27) to the most standard Cauchy equation $g(x+y)=g(x)+g(y)$. By Darboux's theorem (see, e.g., Aczél, 1966, Section 2.1, Theorem 1(b)), $g(x)=cx$ for all $x>0$, that is, $f(p)=p^{-c}$ for all $p\in(0,1)$. ■

The function f is defined on $(0,1)$ since we assume that in real life no bookmaker will assign a subjective probability of exactly 0 or 1 to an event on which he accepts bets. It would be irrational for the bookmaker to have $f(p)\le 1$ for some p, so $f:(0,1)\to(1,\infty)$. To justify the requirement (27), we assume that the bookmaker offers not only "single" but also "double" bets (Wikipedia, 2009). If there are two events with quoted odds a and b that the bookmaker considers independent, his quoted odds on the conjunction of the two events will be ab. If the probabilities of the two events are p and q, respectively, the probability of their conjunction will be pq. Therefore, we have (27).

Theorem 4 provides a justification of Khutsishvili's formula (3): we just assume that the bookmaker applies the same function f to all three probabilities p_1, p_2, and p_3. If $f(p)=p^{-c}$, we have $p_i=a_i^{-\gamma}$, where $\gamma=1/c$ and $i=1,2,3$, and γ can be found from the requirement $p_1+p_2+p_3=1$.

An important advantage of (3) over (26) is that (3) does not impose any upper limits on the overround that the bookmaker may charge (Khutsishvili, 2009). If the game has n possible outcomes ($n=3$ for football and $n=2$ for tennis) and the bookmaker uses $f(p)=p^{-c}$, the overround is

$$\sum_{i=1}^{n}a_i^{-1}-1=\sum_{i=1}^{n}p_i^{c}-1$$

and so continuously changes between -1 and $n-1$ as c ranges over $(0,\infty)$ (in practice, the overround is usually positive, and so $c \in (0,1)$). Even for $n=2$, the upper bound of 1 is too large to be considered a limitation. The situation with (26) is very different: upper bounding the numerator of (26) by 1 and replacing the denominator by $1+o$, where o is the overround, we obtain $p_i < \frac{1}{1+o}$ for all i, and so $o < \min_i p_i^{-1} - 1$; this limitation on o is restrictive when one of the p_i is close to 1.

An interesting phenomenon in racetrack betting, known since Griffith (1949), is that favourites are usually underbet while longshots are overbet (see, e.g., Snowberg and Wolfers, 2007, for a recent survey and analysis). Khutsishvili's formula (3) can be regarded as a way of correcting this "favourite-longshot bias": when a_i is large (the outcome i is a longshot), (3) slashes $1/a_i$ when computing p_i more than (26) does.

Appendix C. Comparison with Other Prediction Algorithms

Other popular algorithms for prediction with expert advice that could be used instead of Algorithm 1 in our empirical studies reported in Section 3 are, among others, the Weighted Average Algorithm (WdAA, proposed by Kivinen and Warmuth, 1999), the weak aggregating algorithm (WkAA, proposed independently by Kalnishkan and Vyugin, 2008, and Cesa-Bianchi and Lugosi, 2006, Theorem 2.3; we are using Kalnishkan and Vyugin's name), and the Hedge algorithm (HA, proposed by Freund and Schapire, 1997). In this appendix we pay most attention to the WdAA since neither WkAA nor HA satisfy bounds of the form (2). (The reader can consult Vovk and Zhdanov, 2008b, for details of experiments with the latter two algorithms and formula (26) used for extracting probabilities from the quoted betting odds.) We also briefly discuss three more naive algorithms.

The Weighted Average Algorithm is very similar to the strong aggregating algorithm (SAA) used in this paper: the WdAA maintains the same weights for the experts as the SAA, and the only difference is that the WdAA merges the experts' predictions by averaging them according to their weights, whereas the SAA uses a more complicated "minimax optimal" merging scheme (given by (19) for the Brier game). The performance guarantee for the WdAA applied to the Brier game is weaker than the optimal (1), but of course this does not mean that its empirical performance is necessarily worse than that of the SAA (i.e., Algorithm 1). Figures 5 and 6 show the performance of this algorithm, in the same format as before (see Figures 1 and 3). We can see that for the football data the maximal difference between the cumulative loss of the WdAA and the cumulative loss of the best expert is slightly larger than that for Algorithm 1 but still well within the optimal bound $\ln K$ given by (1). For the tennis data the maximal difference is almost twice as large as for Algorithm 1, violating the optimal bound $\ln K$.

In its most basic form (Kivinen and Warmuth, 1999, the beginning of Section 6), the WdAA works in the following protocol. At each step each expert, Learner, and Reality choose an element of the unit ball in $\mathbb{R}^n$, and the loss function is the squared distance between the decision (Learner's or an expert's move) and the observation (Reality's move). This covers the Brier game with $\Omega = \{1,\dots,n\}$, each observation $\omega \in \Omega$ represented as the vector $(\delta_\omega\{1\},\dots,\delta_\omega\{n\})$, and each decision $\gamma \in \mathcal{P}(\Omega)$ represented as the vector $(\gamma\{1\},\dots,\gamma\{n\})$. However, in the Brier game the decision makers' moves are known to belong to the simplex $\{(u_1,\dots,u_n) \in [0,\infty)^n \mid \sum_{i=1}^n u_i = 1\}$, and Reality's move is known to be one of the vertices of this simplex. Therefore, we can optimize the ball radius by considering the smallest ball containing the simplex rather than the unit ball. This is what we did for the results reported here (although the results reported in the conference version of this paper, Vovk and Zhdanov, 2008a, are for the WdAA applied to the unit ball in $\mathbb{R}^n$). The

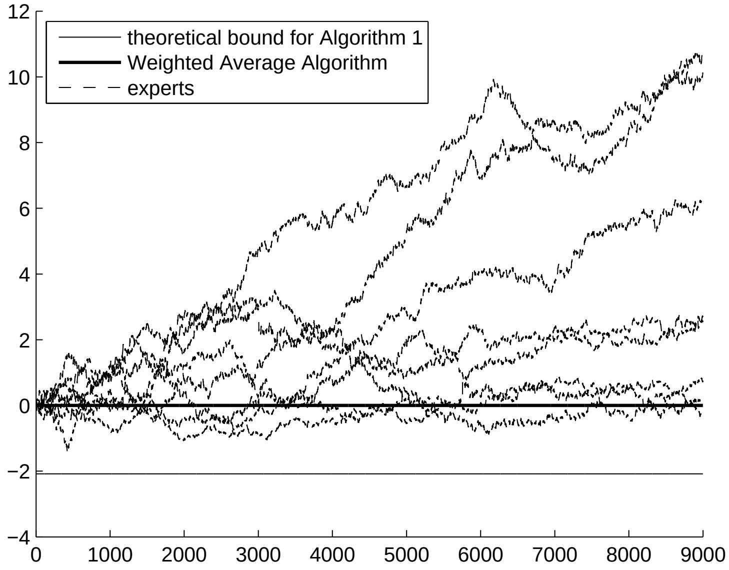

Figure 5: The difference between the cumulative loss of each of the 8 bookmakers and of the Weighted Average Algorithm (WdAA) on the football data. The chosen value of the parameter $c = 1/\eta$ for the WdAA, $c := 16/3$, minimizes its theoretical loss bound. The theoretical lower bound $-\ln 8 \approx -2.0794$ for Algorithm 1 is also shown (the theoretical lower bound for the WdAA, -11.0904, can be extracted from Table 3 below).

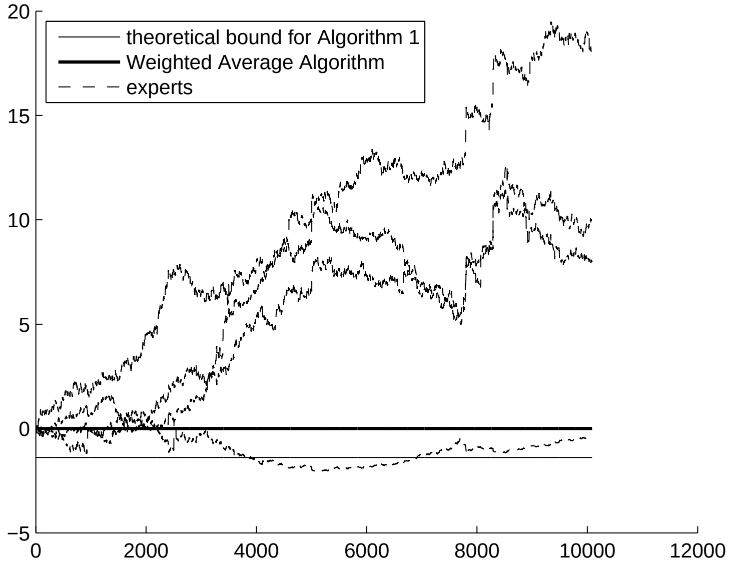

Figure 6: The difference between the cumulative loss of each of the 4 bookmakers and of the WdAA for $c := 4$ on the tennis data.

Algorithm	Maximal difference	Theoretical bound
Algorithm 1	1.2318	2.0794
WdAA ($c = 16/3$)	1.4076	11.0904
WdAA ($c = 1$)	1.2255	none

Table 3: The maximal difference between the loss of each algorithm in the selected set and the loss of the best expert for the football data (second column); the theoretical upper bound on this difference (third column).

radius of the smallest ball is

$$R := \sqrt{1-\frac{1}{n}} \approx \begin{cases} 0.8165 & \text{if } n = 3 \\ 0.7071 & \text{if } n = 2 \\ 1 & \text{if } n \text{ is large.} \end{cases}$$

As described in Kivinen and Warmuth (1999), the WdAA is parameterized by $c := 1/\eta$ instead of η, and the optimal value of c is $c = 8R^2$, leading to the guaranteed loss bound

$$L_N \leq \min_{k=1,\dots,K} L_N^k + 8R^2 \ln K$$

for all $N = 1, 2, \dots$ (see Kivinen and Warmuth, 1999, Section 6). This is significantly looser than the bound (1) for Algorithm 1.

The values $c = 16/3$ and $c = 4$ used in Figures 5 and 6, respectively, are obtained by minimizing the WdAA's performance guarantee, but minimizing a loose bound might not be such a good idea. Figure 7 shows the maximal difference

$$\max_{N=1,\dots,8999} \left(L_N(c) - \min_{k=1,\dots,8} L_N^k \right), \tag{28}$$

where $L_N(c)$ is the loss of the WdAA with parameter c on the football data over the first N steps and L_N^k is the analogous loss of the kth expert, as a function of c. Similarly, Figure 8 shows the maximal difference

$$\max_{N=1,\dots,10087} \left(L_N(c) - \min_{k=1,\dots,4} L_N^k \right) \tag{29}$$

for the tennis data. And indeed, in both cases the value of c minimizing the empirical loss is far from the value minimizing the bound; as could be expected, the empirical optimal value for the WdAA is not so different from the optimal value for Algorithm 1. The following two figures, 9 and 10, demonstrate that there is no such anomaly for Algorithm 1.

Figures 11 and 12 show the behaviour of the WdAA for the value of parameter $c = 1$, that is, $\eta = 1$, that is optimal for Algorithm 1. They look remarkably similar to Figures 1 and 3, respectively.

Precise numbers associated with the figures referred to above are given in Tables 3 and 4: the second column gives the maximal differences (28) and (29), respectively. The third column gives the theoretical upper bound on the maximal difference (i.e., the optimal value of A in (2), if available).

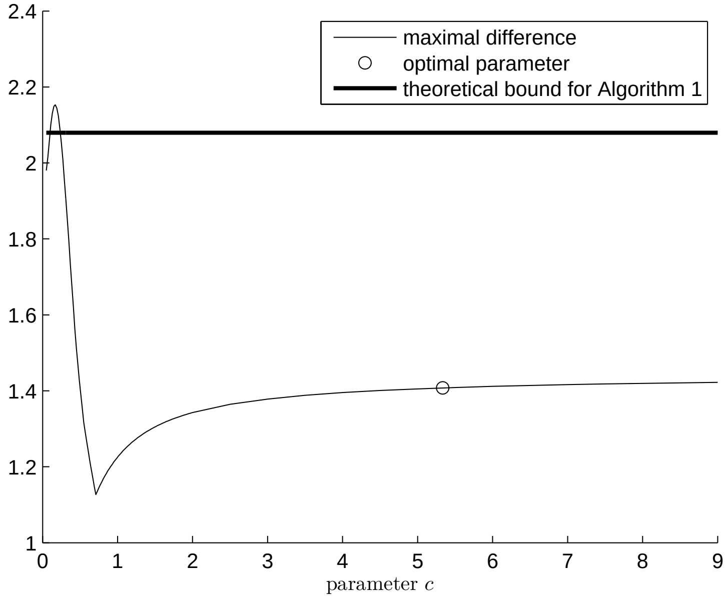

Figure 7: The maximal difference (28) for the WdAA as function of the parameter *c* on the football data. The theoretical guarantee ln 8 for the maximal difference for Algorithm 1 is also shown (the theoretical guarantee for the WdAA, 11.0904, is given in Table 3).

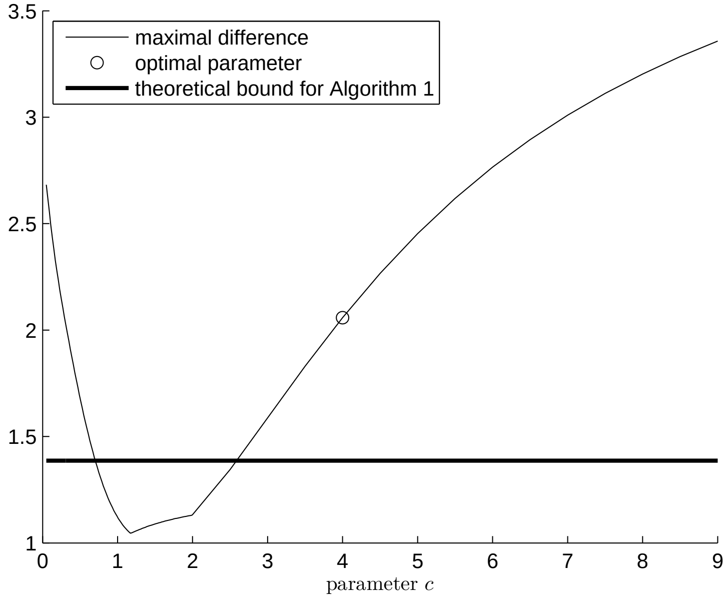

Figure 8: The maximal difference (29) for the WdAA as function of the parameter *c* on the tennis data. The theoretical bound for the WdAA is 5.5452 (see Table 4).

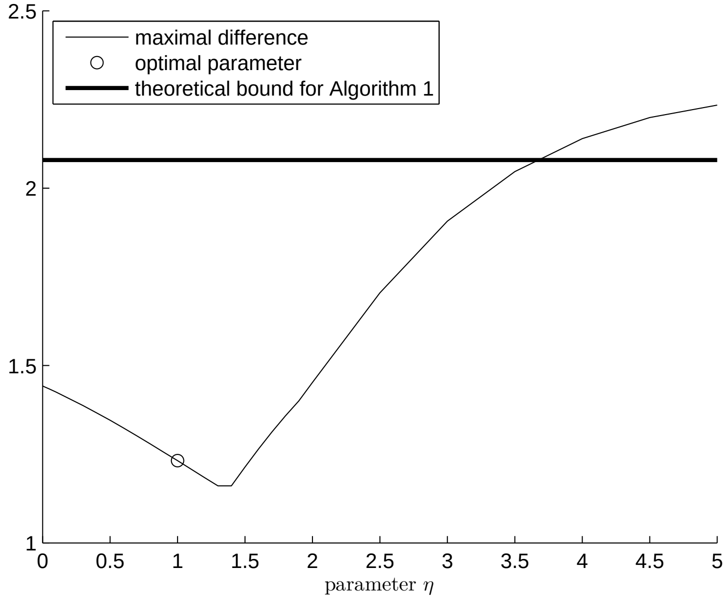

Figure 9: The maximal difference ((28) with η in place of *c*) for Algorithm 1 as function of the parameter η on the football data.

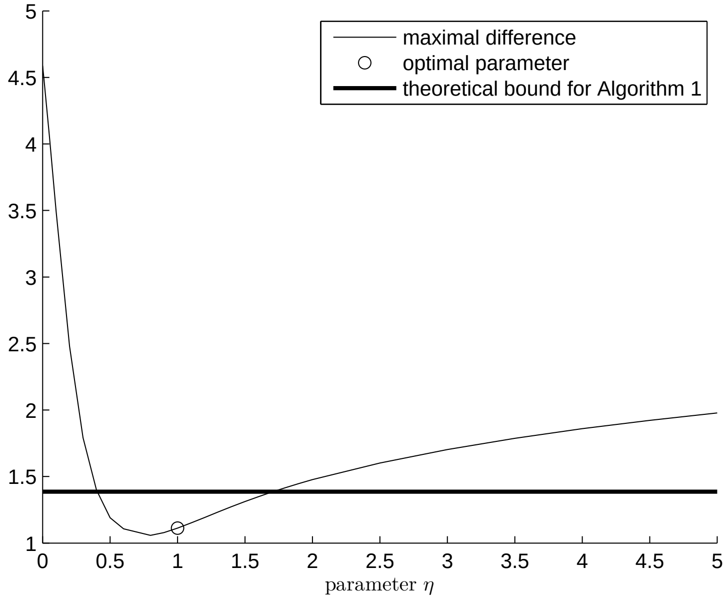

Figure 10: The maximal difference ((29) with η in place of *c*) for Algorithm 1 as function of the parameter η on the tennis data.

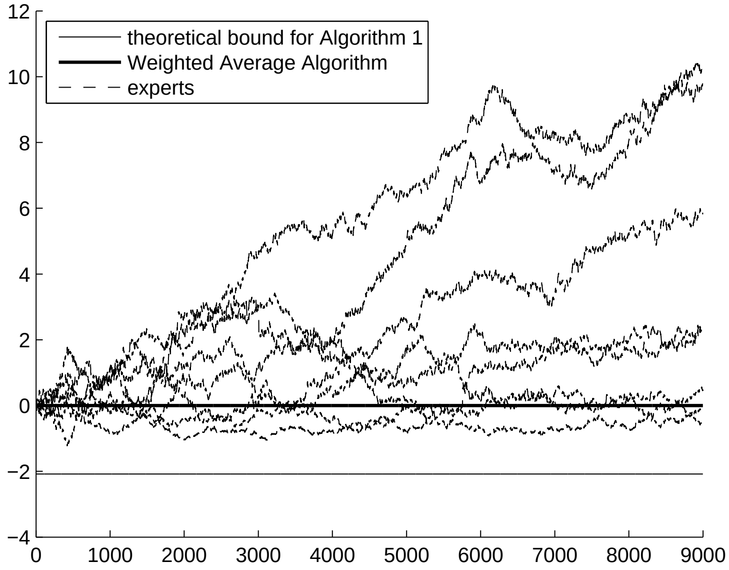

Figure 11: The difference between the cumulative loss of each of the 8 bookmakers and of the WdAA on the football data for $c = 1$ (the value of parameter minimizing the theoretical performance guarantee for Algorithm 1).

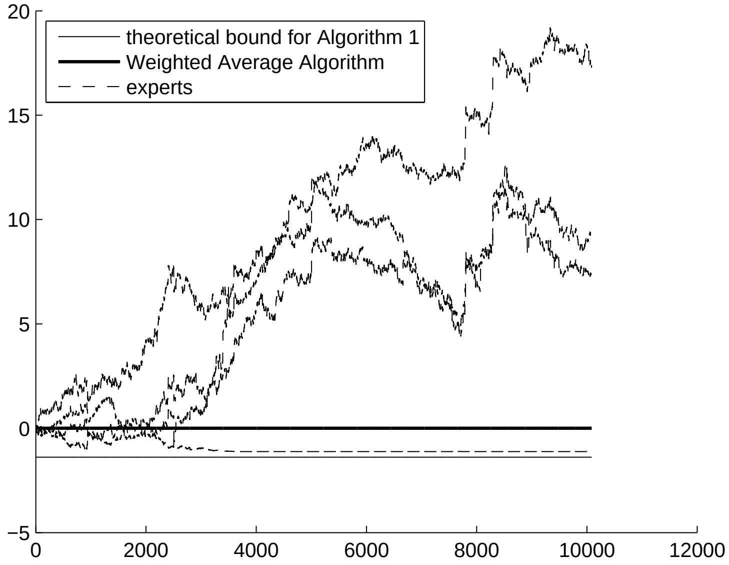

Figure 12: The difference between the cumulative loss of each of the 4 bookmakers and of the WdAA for $c = 1$ on the tennis data.

Algorithm	Maximal difference	Theoretical bound
Algorithm 1	1.1119	1.3863
WdAA ($c = 4$)	2.0583	5.5452
WdAA ($c = 1$)	1.1207	none

Table 4: The maximal difference between the loss of each algorithm in the selected set and the loss of the best expert for the tennis data (second column); the theoretical upper bound on this difference (third column).

The following two algorithms, the weak aggregating algorithm (WkAA) and the Hedge algorithm (HA), make increasingly weaker assumptions about the prediction game being played. Algorithm 1 computes the experts' weights taking full account of the degree of convexity of the loss function and uses a minimax optimal substitution function. Not surprisingly, it leads to the optimal loss bound of the form (2). The WdAA computes the experts' weights in the same way, but uses a suboptimal substitution function; this naturally leads to a suboptimal loss bound. The WkAA "does not know" that the loss function is strictly convex; it computes the experts' weights in a way that leads to decent results for all convex functions. The WkAA uses the same substitution function as the WdAA, but this appears less important than the way it computes the weights. The HA "knows" even less: it does not even know that its and the experts' performance is measured using a loss function. At each step the HA decides which expert it is going to follow, and at the end of the step it is only told the losses suffered by all experts. Both WkAA and HA depend on a parameter, which is denoted c in the case of WkAA and β in the case of HA; the ranges of the parameters are $c \in (0,\infty)$ and $\beta \in [0,1)$. The loss bounds that we give below assume that the loss function takes values in the interval $[0,L]$, in the case of the WkAA, and that the losses are chosen from $[0,L]$, in the case of HA, where L is a known constant. In the case of the Brier loss function, $L = 2$.

In the notation of (1), a simple loss bound for the WkAA is

$$L_N \leq \min_{k=1,\dots,K} L_N^k + 2L\sqrt{N \ln K} \tag{30}$$

(Kalnishkan and Vyugin, 2008, Corollary 14); this is quite different from (1) as the "regret term" $2L\sqrt{N \ln K}$ in (30) depends on N. This bound is guaranteed for $c = \sqrt{\ln K}/L$. For $c = \sqrt{8 \ln K}/L$, Cesa-Bianchi and Lugosi (2006, Theorem 2.3) prove the stronger bound

$$L_N \leq \min_{k=1,\dots,K} L_N^k + L\sqrt{2N \ln K} + L\sqrt{\frac{\ln K}{8}}.$$

The performance of the WkAA on our data sets is significantly worse than that of the WdAA with $c = 1$: the maximal difference (28)–(29) does not exceed $\ln K$ for all reasonable values of c in the case of football but only for a very narrow range of c (which is far from both Kalnishkan and Vyugin's $\sqrt{\ln K}/2$ and Cesa-Bianchi and Lugosi's $\sqrt{8 \ln K}/2$) in the case of tennis. Moreover, the WkAA violates the bound for Algorithm 1 for all reasonable values of c on some natural subsets of the football data set: for example, when prediction starts from the second (2006/2007) season. Nothing similar happens for the WdAA with $c = 1$ on our data sets.

The loss bound for the HA is

$$\mathbb{E}L_N \leq \frac{L_N^* \ln \frac{1}{\beta} + L \ln K}{1-\beta} \tag{31}$$

(Freund and Schapire, 1997, Theorem 2), where $\mathbb{E}L_N$ stands for Learner's expected loss (the HA is a randomized algorithm) and L_N^* stands for $\min_{k=1,\dots,K} L_N^k$. In the same framework, the strong aggregating algorithm attains the stronger bound

$$\mathbb{E}L_N \leq \frac{L_N^* \ln \frac{1}{\beta} + L \ln K}{K \ln \frac{K}{K+\beta-1}} \tag{32}$$

(Vovk, 1998, Example 7). Of course, the SAA applied to the HA framework (as described above, with no loss function) is very different from Algorithm 1, which is the SAA applied to the Brier game; we refer to the former algorithm as SAA-HA. Figure 13 shows the ratio of the right-hand side of (32) to the right-hand side of (31) as function of β.

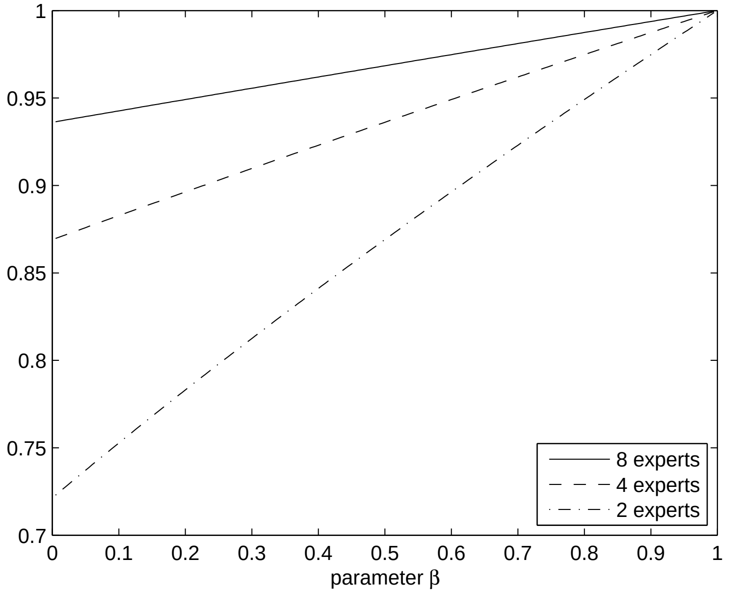

Figure 13: The relative performance of the HA and SAA-HA for various numbers of experts as function of parameter β.

The losses suffered by the HA and the SAA-HA on our data sets are very close and violate Algorithm 1's regret term $\ln K$ for all values of β. It is interesting that, for both football and tennis data, the loss of the HA is almost minimized by setting its parameter β to 0 (the qualification "almost" is necessary only in the case of the tennis data). The HA with $\beta = 0$ coincides with the Follow the Leader Algorithm (FLA), which chooses the same decision as the best (with the smallest loss up to now) expert; if there are several best experts (which almost never happens after the first step), their predictions are averaged with equal weights. Standard examples (see, e.g., Cesa-Bianchi and Lugosi, 2006, Section 4.3) show that this algorithm (unlike its version Follow the Perturbed Leader) can fail badly on some data sequences. Its empirical performance on the football data set is not so bad: it violates the loss bound for Algorithm 1 only slightly; however, on the tennis data set the bound is violated badly.

The decent performance of the Follow the Leader Algorithm on the football data set suggests checking the empirical performance of other similarly naive algorithms, such as the following two. The *Simple Average Algorithm*'s decision is defined as the arithmetic mean of the experts' decisions (with equal weights). The *Bayes Mixture Algorithm* (BMA) is the strong aggregating algorithm applied to the log loss function; this algorithm is in fact optimal, but not for the Brier loss function. The BMA has a very simple description (Cesa-Bianchi and Lugosi, 2006, Section 9.2), and was studied from the point of view of prediction with expert advice already in DeSantis et al. (1988).

We have found that none of the three naive algorithms perform consistently poorly, but they always fail badly on some natural part of our data sets. The advantage of the more sophisticated algorithms having strong performance guarantees is that there is no danger of catastrophic performance on any data set.

References

János Aczél. *Lectures on Functional Equations and their Applications*. Academic Press, New York, 1966.

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

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, England, 2006.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44:427–485, 1997.

A. Philip Dawid. Probability forecasting. In Samuel Kotz, Norman L. Johnson, and Campbell B. Read, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley, New York, 1986.

Alfredo DeSantis, George Markowsky, and Mark N. Wegman. Learning probabilistic prediction functions. In *Proceedings of the Twenty Ninth Annual IEEE Symposium on Foundations of Computer Science*, pages 110–119, Los Alamitos, CA, 1988. IEEE Computer Society.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

Richard M. Griffith. Odds adjustments by American horse-race bettors. *American Journal of Psychology*, 62:290–294, 1949.

Godfrey H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, England, second edition, 1952.

David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44:1906–1925, 1998.

Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. *Journal of Computer and System Sciences*, 74:1228–1244, 2008. Special Issue devoted to COLT 2005.

Victor Khutsishvili. Personal communication. E-mail exchanges (from 27 November 2008), 2009.

Jyrki Kivinen and Manfred K. Warmuth. Averaging expert predictions. In Paul Fischer and Hans U. Simon, editors, *Proceedings of the Fourth European Conference on Computational Learning Theory*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 153–167, Berlin, 1999. Springer.

Nick Littlestone and Manfred K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108:212–261, 1994.

Erik Snowberg and Justin Wolfers. Explaining the favorite-longshot bias: Is it risk-love or misperceptions? Available on-line at `http://bpp.wharton.upenn.edu/jwolfers/` (accessed on 2 November 2009), November 2007.

John A. Thorpe. *Elementary Topics in Differential Geometry*. Springer, New York, 1979.

Vladimir Vovk. Aggregating strategies. In Mark Fulk and John Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

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

Vladimir Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35:247–282, 1999.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

Vladimir Vovk and Fedor Zhdanov. Prediction with expert advice for the Brier game. In Andrew McCallum and Sam Roweis, editors, *Proceedings of the Twenty Fifth International Conference on Machine Learning*, pages 1104–1111, New York, 2008a. ACM.

Vladimir Vovk and Fedor Zhdanov. Prediction with expert advice for the Brier game. Technical Report `arXiv:0708.2502v2` [cs.LG], `arXiv.org` e-Print archive, June 2008b.

Wikipedia. Glossary of bets offered by UK bookmakers — Wikipedia, The Free Encyclopedia, 2009. Accessed on 2 November.