Prediction with expert advice for the Brier game

Vladimir Vovk and Fedor Zhdanov

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England

June 27, 2008

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., [5], [11], [2]) and has been applied to various loss functions; see [3] 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 [8, 14] that the optimal additive constant is attained by the "strong aggregating algorithm" proposed in [13] (we use the adjective "strong" to distinguish it from the "weak aggregating algorithm" of [9]).

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 [13]. 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., [4]).

The mixability of the square loss function in the case of observations belonging to a bounded interval of real numbers was demonstrated in [8]; Haussler et al.'s algorithm was simplified in [16]. 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 [17] 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* [1]: 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 $\gamma\{3\} = 1/4$, $\lambda(\omega, \gamma) = (1/2-1)^2 + (1/4-0)^2 + (1/4-0)^2 = 3/8$.)

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

$L_0 := 0.$
$L_0^k := 0,\ k = 1, \ldots, K.$
for $N = 1, 2, \ldots$ **do**
    Expert k announces $\gamma_N^k \in \Gamma,\ k = 1, \ldots, K.$
    Learner announces $\gamma_N \in \Gamma.$
    Reality announces $\omega_N \in \Omega.$
    $L_N := L_{N-1} + \lambda(\omega_N, \gamma_N).$
    $L_N^k := L_{N-1}^k + \lambda(\omega_N, \gamma_N^k),\ k = 1, \ldots, K.$
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. Its description uses the notation $t^+ := \max(t, 0)$.

Algorithm 1 Strong aggregating algorithm for the Brier game

$w_0^k := 1$, $k = 1, \ldots, K$.
for $N = 1, 2, \ldots$ **do**
   Read the Experts' predictions γ_N^k, $k = 1, \ldots, K$.
   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 ω_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*

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

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

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

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

The second part of this theorem follows from its special case with $|\Omega| = 2$ (the binary case). However, we are not aware of a proof of this result in the binary case, and we will not use this reduction.

3 Experimental results

In our first empirical study of Algorithm 1 we use historical data about 6473 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 three seasons, 2005/2006, 2006/2007, and 2007/2008. (The 2007/2008 season ended in May shortly after the ICML 2008 submission deadline, and so the data set used in the conference version [17] of this paper covered only part of that season, with 6416 matches in total.) 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 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, $k = 1, \ldots, 8$, 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 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 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. In Figure 1 one bookmaker clearly performs worse than the others. His poor performance may be explained by his mean overround being about 0.13, near the top end of the distribution in Figure 2. (On one hand, a high overround diminishes the need for accurate probability forecasts, and on the other, our estimates (3) of the probabilities implicit in the announced odds also become less precise.)
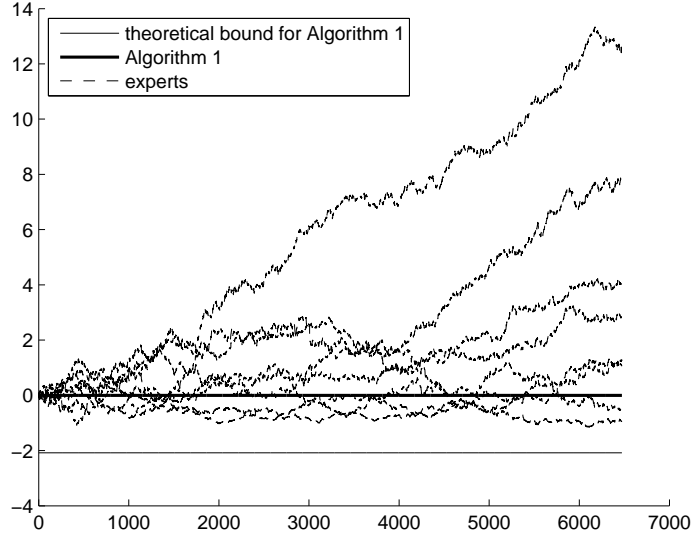
Figure 1: The difference between the cumulative loss of each of the 8 bookmakers (experts) and of Algorithm 1 on the football data. The theoretical lower bound $-\ln 8$ from Theorem 1 is also shown.

Figure 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, e.g., 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, 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 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 [13] could have also been used (and would have given similar
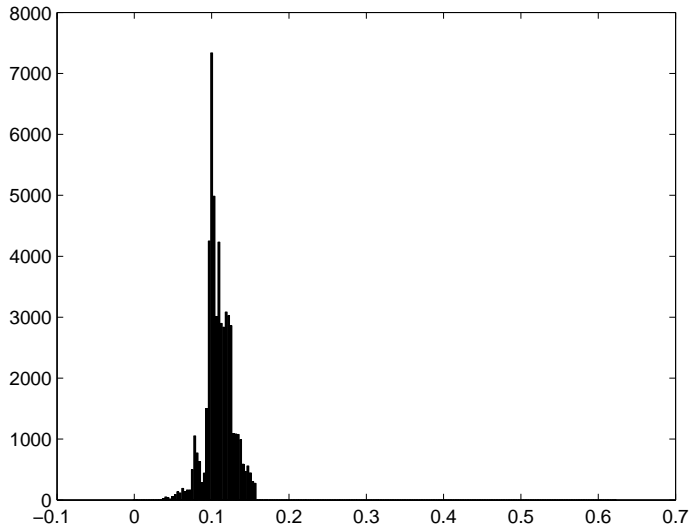
Figure 2: The overround distribution histogram for the football data, with 200 bins of equal size between the minimum and maximum values of the overround.

results). We included it since we are not aware of any empirical studies even for the binary case.

For comparison with several other popular prediction algorithms, see Appendix B. The data used for producing all the figures and tables in this section and in Appendix B 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., [13] and [8].) All definitions that we will need can be found in, e.g., [12].

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

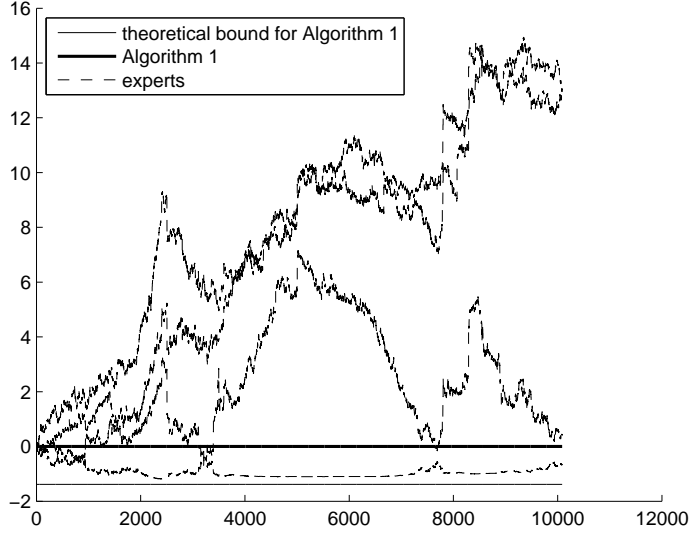$$L_N \leq \min_{k=1,\ldots,K} L_N^k + \frac{\ln K}{\eta}, \quad N = 1, 2, \ldots,$$

6

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 $-\ln 4$.

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 [14]; 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

- $\Phi_\eta(\Sigma)$ 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 [12], 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
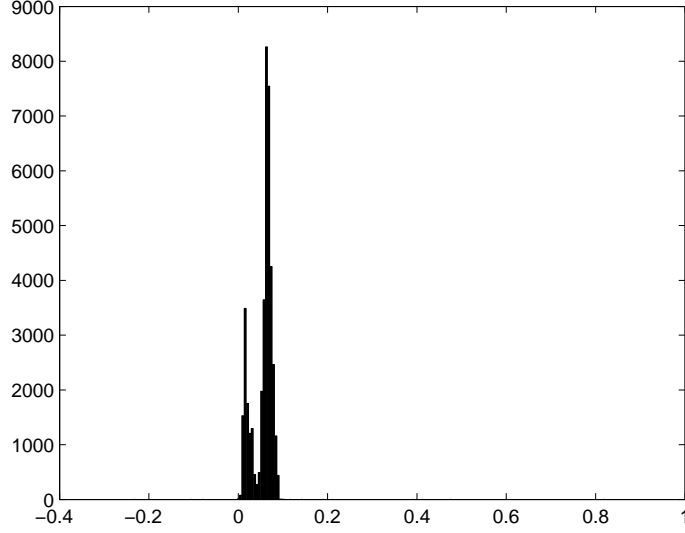
7

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

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 $\Omega = \{1, \ldots, n\}$.

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

$$
\begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^{n-1} \\ x^n \end{pmatrix} = \begin{pmatrix} e^{-\eta((u^1-1)^2+(u^2)^2+\cdots+(u^n)^2)} \\ e^{-\eta((u^1)^2+(u^2-1)^2+\cdots+(u^n)^2)} \\ \vdots \\ e^{-\eta((u^1)^2+\cdots+(u^{n-1}-1)^2+(u^n)^2)} \\ e^{-\eta((u^1)^2+\cdots+(u^{n-1})^2+(u^n-1)^2)} \end{pmatrix}, \tag{5}
$$

where $u^1, \ldots, u^{n-1}$ are the coordinates on the surface, $u^1, \ldots, u^{n-1} \in (0,1)$ 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

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

8

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

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.

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 $\mathbb{R}^n$. The coefficient in front of e_1 is the $(n-1) \times (n-1)$ determinant

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

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

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

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

$$= e^{-2\eta u^1} \left((-1)^n (u^n - u^1 - 1) + (-1)^{n+1}(u^1 - u^2) \right.$$
$$\left. + (-1)^{n+1}(u^1 - u^3) + \cdots + (-1)^{n+1}(u^1 - u^{n-1}) \right)$$
$$= e^{-2\eta u^1}(-1)^n \left((u^2 + u^3 + \cdots + u^n) - (n-1)u^1 - 1 \right)$$
$$= -e^{-2\eta u^1}(-1)^n n u^1 \propto u^1 e^{-2\eta u^1} \quad (6)$$

(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

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

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

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

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

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

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

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

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

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

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

$$
= u^1(1 - 2\eta u^2)(1 - 2\eta u^3) \cdots (1 - 2\eta u^n)
$$
$$
+ u^2(1 - 2\eta u^1)(1 - 2\eta u^3) \cdots (1 - 2\eta u^n) + \cdots
$$
$$
+ u^n(1 - 2\eta u^1)(1 - 2\eta u^2) \cdots (1 - 2\eta u^{n-1}) \tag{8}
$$

(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

$$
-1 < 1 - 2\eta < t_i < 1, \quad i = 1, \ldots, n,
$$
$$
t_1 + \cdots + t_n = n - 2\eta > n - 2. \tag{9}
$$

Our goal is to prove

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

i.e.,

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

This reduces to

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} > n \tag{11}$$

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

$$\frac{1}{t_1} + \cdots + \frac{1}{t_n} < n \tag{12}$$

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 $\eta \leq 1$, 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 $(u^1, \ldots, u^{n-1})$. 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 ([12], Chapter 13, Theorem 1 and its proof).

It remains to consider the case $\eta \leq 1$. 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

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

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 (5) 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 (5) 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 $(x^1, \ldots, x^n) = (1, \ldots, 1)$ when $\eta = 0$). Therefore, for all $(u^1, \ldots, u^{n-1}, \eta)$ the value of $k_1(u^1, \ldots, u^{n-1}, \eta)$ is positive: if k_1 had different signs at two points in the set

$$\big\{ (u^1, \ldots, u^{n-1}, \eta) \,|\, u^1 \in (0, 1), \ldots, u^{n-1} \in (0, 1),$$
$$u^1 + \cdots + u^{n-1} < 1, \eta \in (0, 1) \big\}, \quad (14)$$

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., [16], 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 [16]. 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., [16], Appendix A.1).

Suppose that we are given a generalized prediction $(l_1, \ldots, l_n)^{\mathrm{T}}$ computed by the aggregating pseudo-algorithm from a normalized distribution on the experts. Since $(l_1, \ldots, l_n)^{\mathrm{T}}$ 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

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

Now suppose we are given a generalized prediction $(L_1, \ldots, L_n)^{\mathrm{T}}$ 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 $(L_1 - t, \ldots, L_n - t)^{\mathrm{T}}$ is still a superprediction and then find (15) satisfying

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

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

Proposition 1. *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 $\lambda_1, \ldots, \lambda_n$ as in (15) will be

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

$$t = s - 1 - (u^1)^2 - \cdots - (u^n)^2. \tag{20}$$

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 1. 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, i.e.,

$$\overline{t} = L_i - 1 + 2\overline{u}^i - (\overline{u}^1)^2 - \cdots - (\overline{u}^n)^2, \tag{21}$$

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

We can rewrite (17) as

$$\begin{cases} t \le L_1 - 1 + 2u^1 - (u^1)^2 - \cdots - (u^n)^2, \\ \qquad\qquad \vdots \\ t \le L_n - 1 + 2u^n - (u^1)^2 - \cdots - (u^n)^2, \end{cases} \tag{22}$$

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 $\epsilon_i := \overline{u}^i - u^i$ is maximal. Then every value of t satisfying (22) will also satisfy

$$t \le L_i - 1 + 2u^i - \sum_{j=1}^n (u^j)^2$$

$$= L_i - 1 + 2\overline{u}^i - 2\epsilon_i - \sum_{j=1}^n (\overline{u}^j)^2 + 2\sum_{j=1}^n \epsilon_j \overline{u}^j - \sum_{j=1}^n \epsilon_j^2$$

$$\le L_i - 1 + 2\overline{u}^i - \sum_{j=1}^n (\overline{u}^j)^2 - \sum_{j=1}^n \epsilon_j^2 \le \overline{t},$$

with the last $\le$ following from (21) and becoming $<$ when not all u^j coincide with $\overline{u}^j$. $\qquad\square$

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., [3], 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

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] G. H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, Cambridge, England, second edition, 1952.

[8] 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.

[9] Yuri Kalnishkan and Michael V. Vyugin. The Weak Aggregating Algorithm and weak mixability. In Peter Auer and Ron Meir, editors, *Proceedings of the Eighteenth Annual Conference on Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 188–203, Berlin, 2005. Springer.

[10] 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.

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

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

[13] 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.

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

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

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

[17] 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*, 2008.

A   Watkins's theorem

Watkins's theorem is stated in [15] (Theorem 8) not in sufficient generality: it presupposes that the loss function is perfectly 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.

In this appendix we will use a slightly more general notion of a game of prediction $(\Omega, \Gamma, \lambda)$: 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 [14], for each $K = 1, 2, \ldots$ 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 $N = 1, 2, \ldots$ and all $k \in \{1, \ldots, K\}$,

$$L_N \le L_N^k + a; \tag{23}$$

otherwise, Environment wins. It is possible that $L_N = \infty$ or $L_N^k = \infty$ in (23); 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 [14]):

- 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 $(\Omega, \Gamma, \lambda)$ is *η-mixable*, where $\eta > 0$, if

$$\forall \gamma_1 \in \Gamma, \gamma_2 \in \Gamma, \alpha \in [0,1] \; \exists \delta \in \Gamma \; \forall \omega \in \Omega :$$
$$e^{-\eta\lambda(\omega,\delta)} \geq \alpha e^{-\eta\lambda(\omega,\gamma_1)} + (1-\alpha)e^{-\eta\lambda(\omega,\gamma_2)}. \quad (24)$$

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 *perfectly mixable* if it is η-mixable for some $\eta > 0$.

It follows from [7] (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 [14].) Let η^* be the supremum of the η for which the prediction game is η-mixable (with $\eta^* := 0$ when the game is not perfectly mixable). The compactness of Γ implies that the prediction game is η^*-mixable.

Theorem 2 (Chris Watkins). *For any $K \in \{1, 2, \ldots\}$,*

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

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

The theorem does not say explicitly, but it is easy to check, that $\mathrm{L} \smile \mathcal{G}_K(a_K)$: this follows both from general considerations (cf. Lemma 3 in [14]) and from the fact that the SAA wins $\mathcal{G}_K(a_K) = \mathcal{G}_K(\ln K/\eta^*)$.

Proof of Theorem 2. The proof will use some notions and notation used in the statement and proof of Theorem 1 of [14]. 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 [14] (the only assumption in [14] not directly made in this paper) is satisfied. In view of the fact that $\mathrm{L} \smile \mathcal{G}_K(\ln K/\eta^*)$, we only need to show that $\mathrm{L} \smile \mathcal{G}_K(a)$ does not hold for $a < \ln K/\eta^*$. Fix $a < \ln K/\eta^*$.

The separation curve, as defined in [14], consists of the points $(c(\beta), c(\beta)/\eta) \in [0, \infty)^2$, where $\beta := e^{-\eta}$ and η ranges over $[0, \infty]$ (see [14], Theorem 1). Since the two-fold convex mixture in (24) 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 [14]). Therefore, E (=Environment) has a winning strategy in the game $\mathcal{G}(1, a/\ln K)$, as defined in [14]. It is easy to see from the proof of Theorem 1 in [14] that the definition of the game $\mathcal{G}$ in [14] 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 [14] 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 [14] 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 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. □

B   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, Kivinen and Warmuth's [10] Weighted Average Algorithm (WdAA), Kalnishkan and Vyugin's [9] Weak Aggregating Algorithm (WkAA), and Freund and Schapire's [6] Hedge algorithm (HA). In this appendix we consider these three algorithms and three more naive algorithms (which, nevertheless, perform surprisingly well).
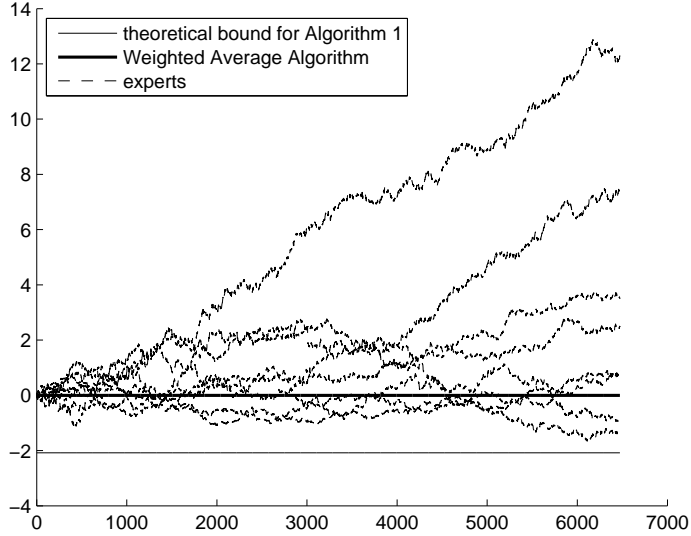
Figure 5: The difference between the cumulative loss of each of the 8 bookmakers and of the Weighted Average Algorithm (WdAA) on the football data. The chosen value of the parameter $c = 1/\eta$ for the WdAA, $c := 16/3$, minimizes its theoretical loss bound. The theoretical lower bound $-\ln 8 \approx -2.0794$ for Algorithm 1 is also shown (the theoretical lower bound for the Weighted Average Algorithm, -11.0904, can be extracted from Table 1 below).

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 larger that for Algorithm 1 but still well within the optimal bound $\ln K$ given by (1). For the tennis data the maximal difference is about twice as large as for Algorithm 1, violating the optimal bound $\ln K$.

In its most basic form ([10], 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 dis-
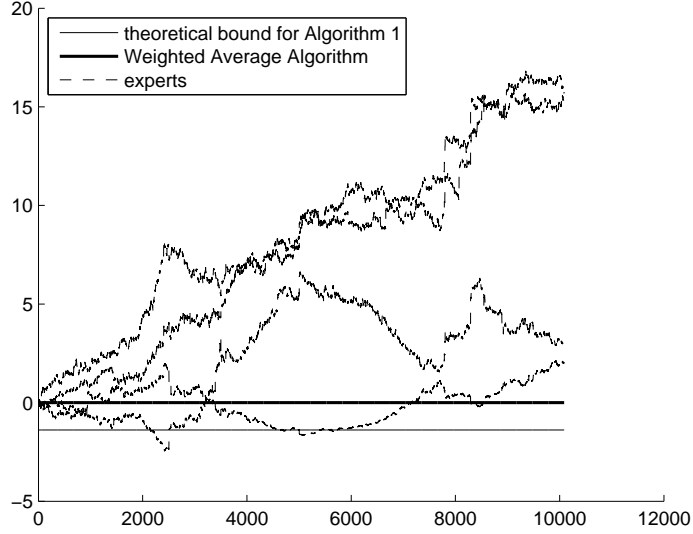
20

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.

tance between the decision (Learner's or an expert's move) and the observation (Reality's move). This covers the Brier game with $\Omega = \{1, \ldots, n\}$, each observation $\omega \in \Omega$ represented as the vector $(\delta_\omega\{1\}, \ldots, \delta_\omega\{n\})$, and each decision $\gamma \in \mathcal{P}(\Omega)$ represented as the vector $(\gamma\{1\}, \ldots, \gamma\{n\})$. However, in the Brier game the decision makers' moves are known to belong to the simplex $\{(u^1, \ldots, u^n) \in [0, \infty)^n \mid \sum_{i=1}^{n} u^i = 1\}$, 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 [17] are for the WdAA applied to the unit cube in $\mathbb{R}^n$). The 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 [10], 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

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

for all $N = 1, 2, \ldots$ (see [10], Section 6). This is significantly looser than the bound (1) for Algorithm 1.

21
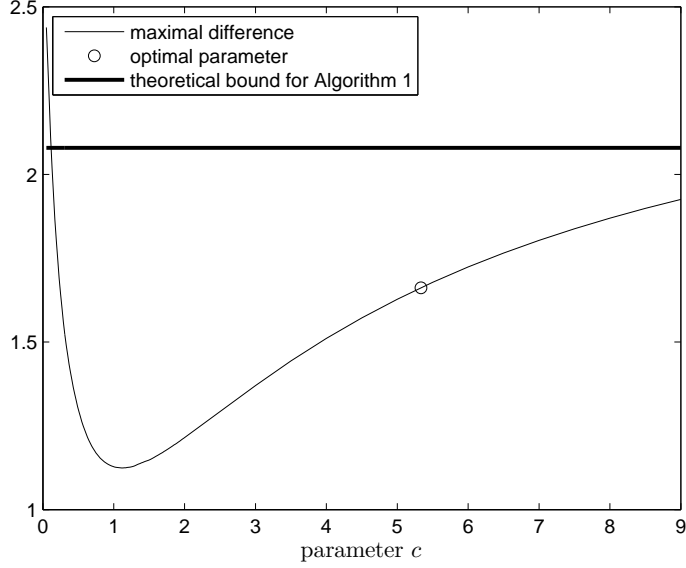
Figure 7: The maximal difference (25) for the WdAA as function of the parameter c on the football data. The theoretical guarantee $\ln 8$ for the maximal difference for Algorithm 1 is also shown (the theoretical guarantee for the WdAA, 11.0904, is given in Table 1).

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

$$\max_{N=1,\ldots,6473} \left(L_N(c) - \min_{k=1,\ldots,8} L_N^k \right), \tag{25}$$

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

$$\max_{N=1,\ldots,10087} \left(L_N(c) - \min_{k=1,\ldots,4} L_N^k \right) \tag{26}$$

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 $c = 1$, i.e., $\eta = 1$, that is optimal for Algorithm 1. They look remarkably

22
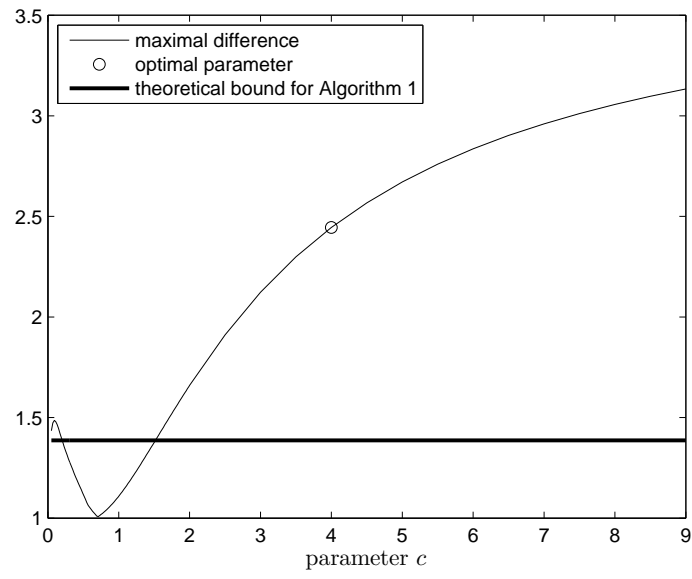
Figure 8: The maximal difference (26) for the WdAA as function of the parameter c on the tennis data. The theoretical bound for the WdAA is 5.5452 (see Table 1).
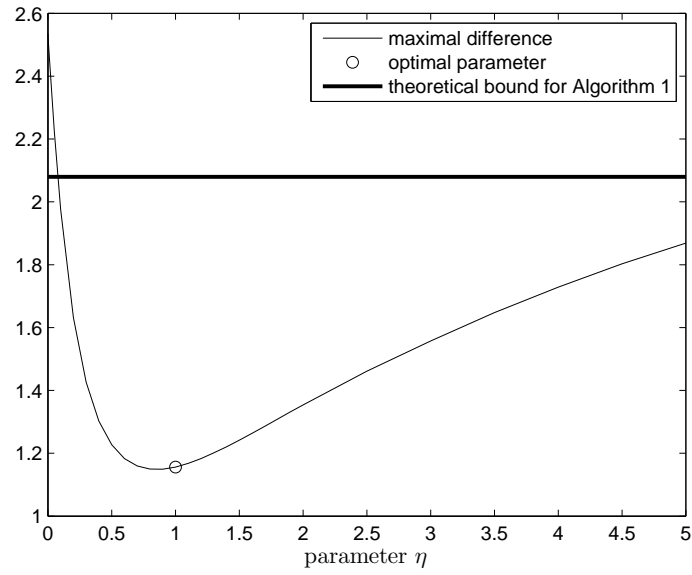
Figure 9: The maximal difference ((25) with η in place of c) for Algorithm 1 as function of the parameter η on the football data.
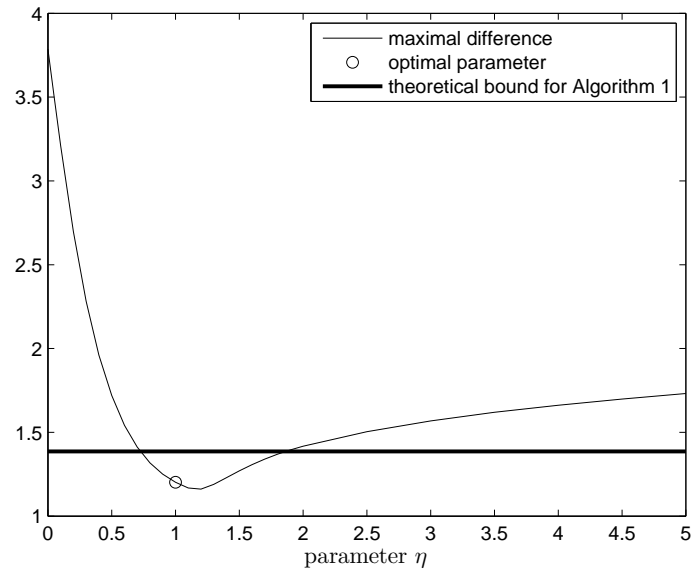
Figure 10: The maximal difference ((26) with η in place of c) for Algorithm 1 as function of the parameter η on the tennis data.
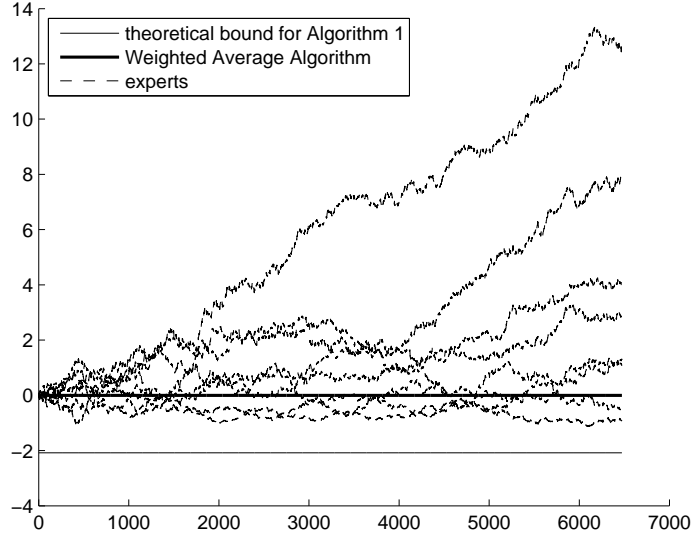
Figure 11: The difference between the cumulative loss of each of the 8 book-makers and of the WdAA on the football data for $c = 1$ (the value of parameter minimizing the theoretical performance guarantee for Algorithm 1).

similar to Figures 1 and 3, respectively.

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. Therefore, it is not surprising that the WkAA does not perform as well as Algorithm 1 and the WdAA with $c = 1$; the performance of the HA is even weaker: see Figures 13–16. The HA is a randomized algorithm, so we show the expected performance.

Figures 13–16 show the performance of the WdAA and the HA for all possible values of their parameters (c and β, respectively). We do not show the optimal
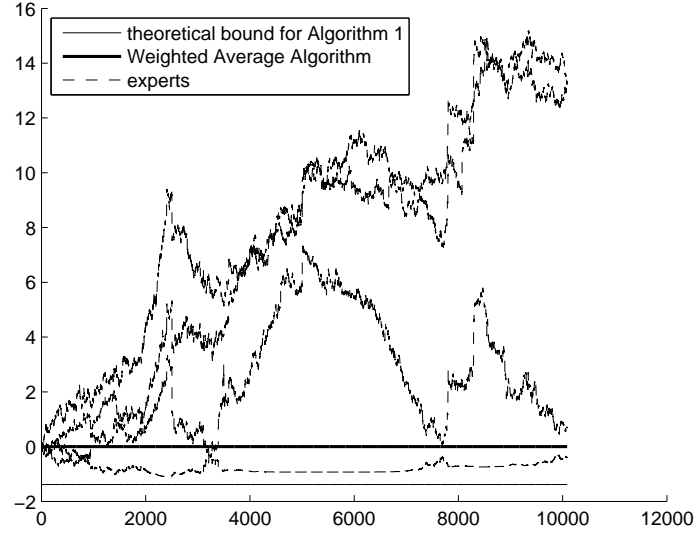
26

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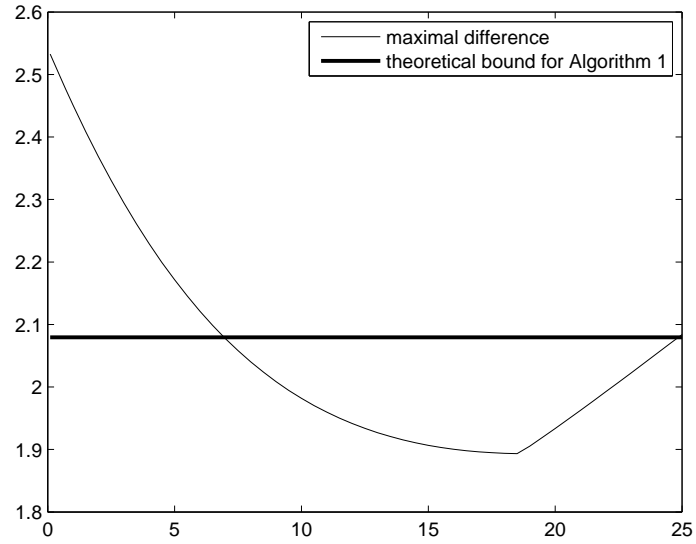

Figure 13: The maximal difference for the Weak Aggregating Algorithm (WkAA) as function of c on the football data.
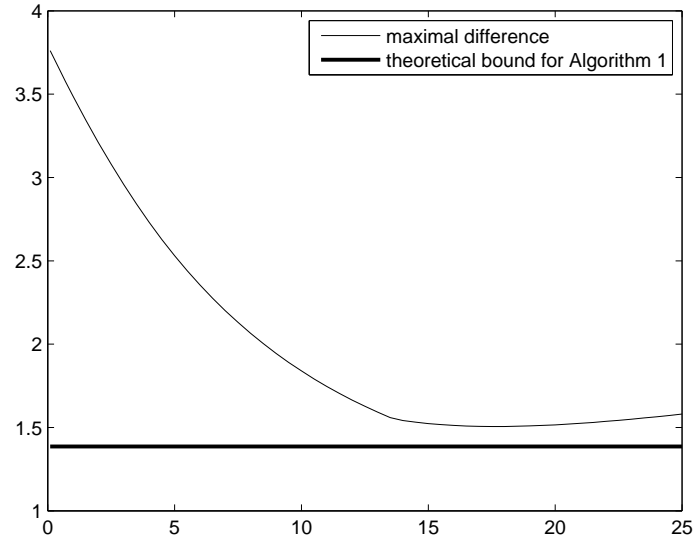
Figure 14: The maximal difference for the WkAA as function of c on the tennis data.
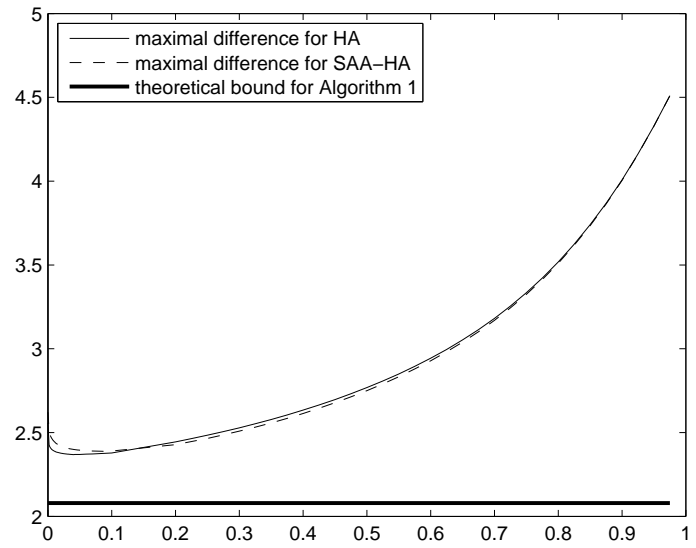

Figure 15: The expected maximal difference for the Hedge algorithm (HA) and for the SAA Hedge algorithm (SAA-HA) as a function of β on the football data.

values of parameters since neither algorithm satisfies a loss bound of the form (2) (typical loss bounds for these algorithms allow A to depend on N, and the optimal value would also depend on N).

In the case of the HA, the loss bound given in the original paper [6] was replaced, in the same framework, by a stronger bound in [14] (Example 7). The stronger bound is achieved by the SAA applied to the HA framework described above (with no loss function); this algorithm is referred to as SAA-HA in the captions. The description of the SAA-HA given in [14] admits some freedom in the choice of Learner's decision; our implementation replaces the HA's weights p^k, $k = 1, \ldots, K$, with

$$\frac{-\ln\left(1 + (\beta - 1)p^k\right)}{-\sum_{k=1}^{K} \ln\left(1 + (\beta - 1)p^k\right)}, \quad k = 1, \ldots, K.$$

The losses suffered by the HA and the SAA-HA are very close.

An interesting observation is 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 in the case of the tennis data as well: the lines of maximal difference in Figure 16 are not monotonic for β extremely close to 0). 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., [3], Section 4.3) show that this algorithm (unlike its version Follow the Perturbed Leader) can fail badly on some data sequences. However, its empirical performance (Figures 17 and 18) on our data sets is not so bad: it violates the loss bounds for Algorithm 1 only slightly.

The decent performance of the Follow the Leader Algorithm suggests checking the empirical performance of other similarly naive algorithms. The Simple Average Algorithm's decision is defined as the arithmetic mean of the experts' decisions (with equal weights). Figures 19 and 20 show the performance of this algorithm. It does violate the theoretical loss bound for Algorithm 1, but not significantly (especially in the case of football data).

The last naive algorithm that we consider is in fact optimal, but for a different loss function. The *Bayes Mixture Algorithm* (BMA) is the Strong Aggregating Algorithm applied to the log loss function. This algorithm has a very simple description [13], and was studied from the point of view of prediction with expert advice already in [5]. Figures 21 and 22 show the performance of the BMA measured by the Brier loss function, as usual. The performance is excellent for the football data but much weaker for tennis.

Despite the decent performance of the three naive algorithms on our two data sets, there is always a danger of catastrophic performance on some data set: there are no performance guarantees for these algorithms whatsoever. It is an important advantage of more sophisticated algorithms that they establish some upper bound on the algorithm's regret.

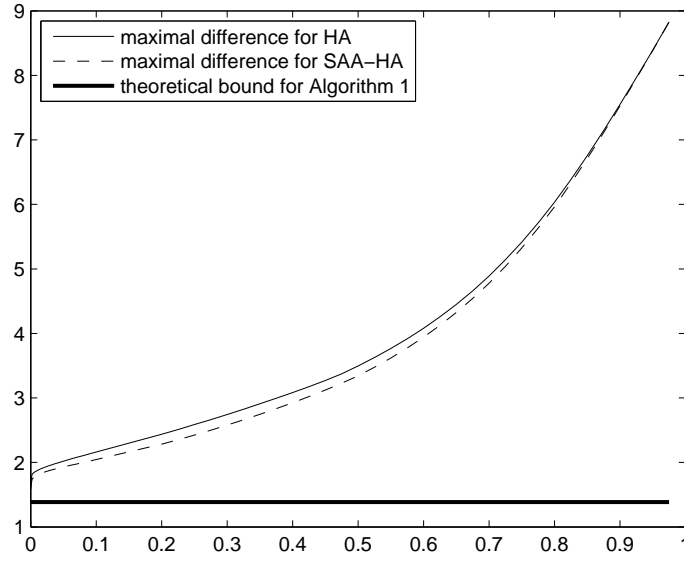Precise numbers associated with the figures referred to above are given in

Figure 16: The expected maximal difference for the HA and for the SAA-HA as a function of β on the tennis data.
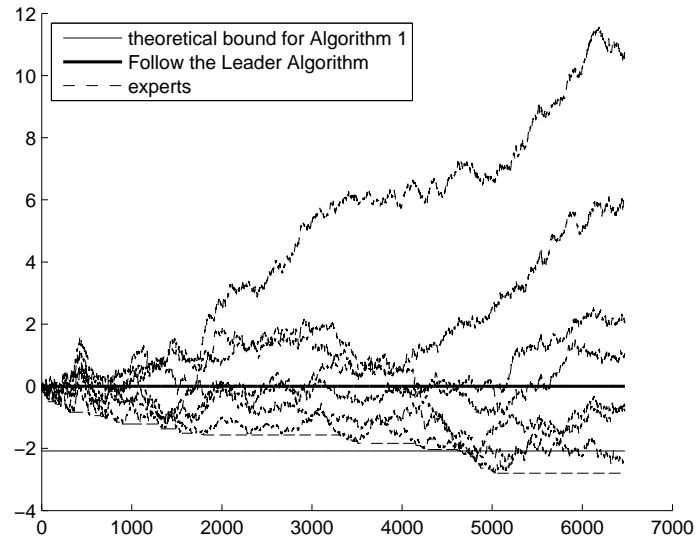

Figure 17: The difference between the cumulative loss of each of the 8 book-makers and of the Follow the Leader Algorithm on the football data.
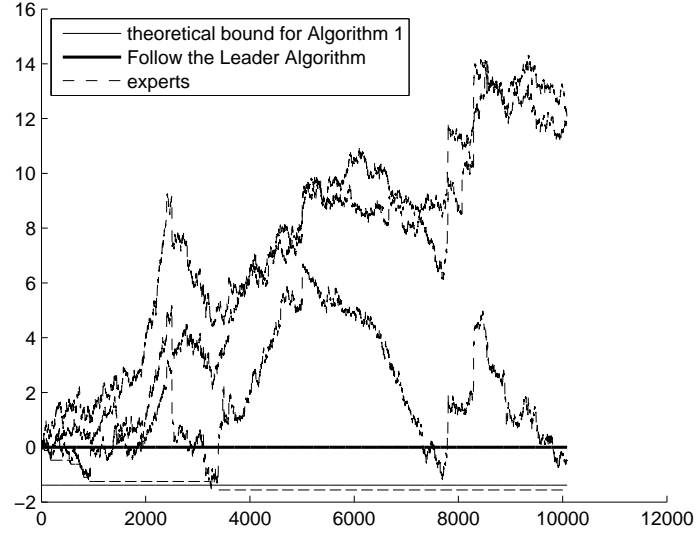
Figure 18: The difference between the cumulative loss of each of the 4 book-makers and of the Follow the Leader Algorithm on the tennis data.
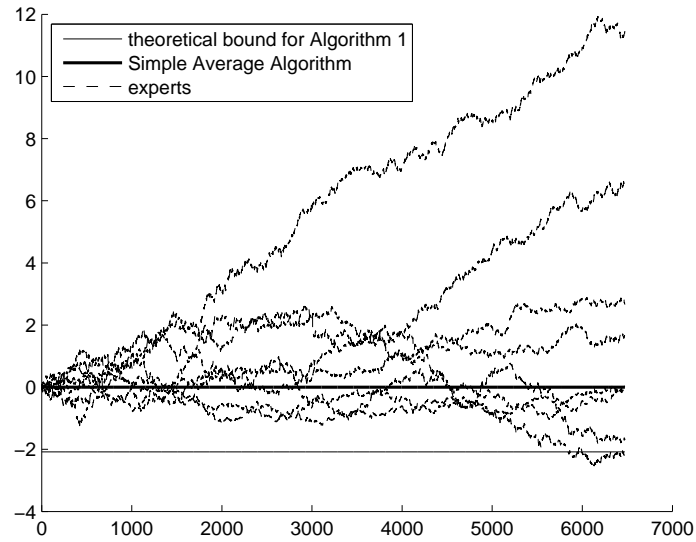

Figure 19: The difference between the cumulative loss of each of the 8 book-makers and of the Simple Average Algorithm on the football data.
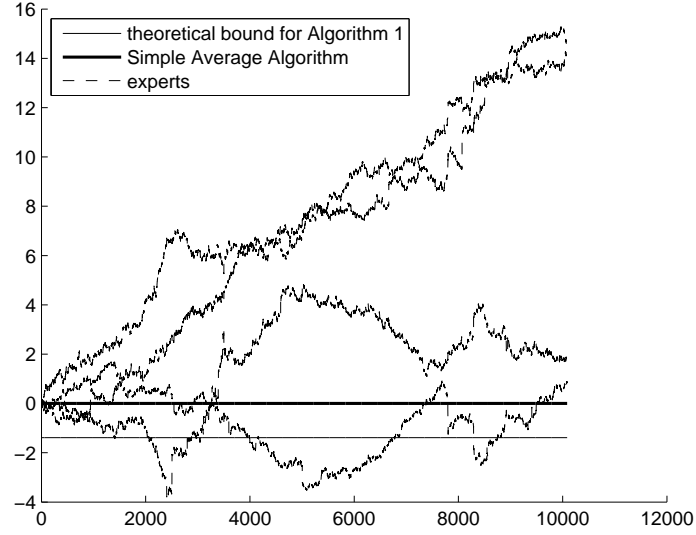
31

Figure 20: The difference between the cumulative loss of each of the 4 book-makers and of the Simple Average Algorithm on the tennis data.
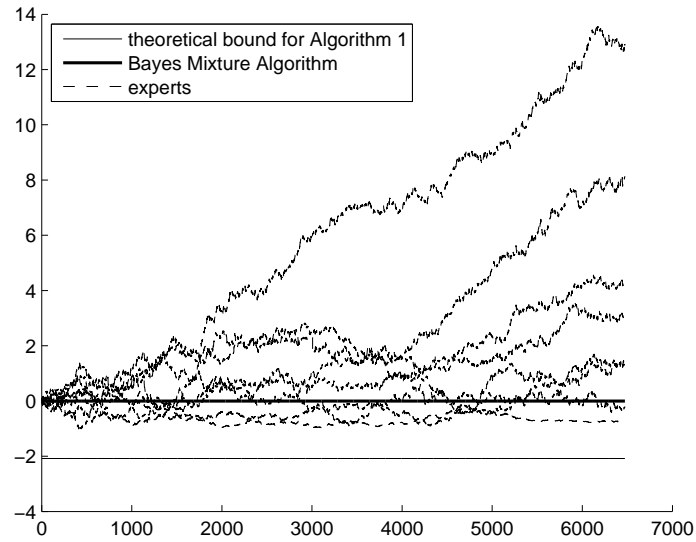

Figure 21: The difference between the cumulative loss of each of the 8 book-makers and of the Bayes Mixture Algorithm on the football data.
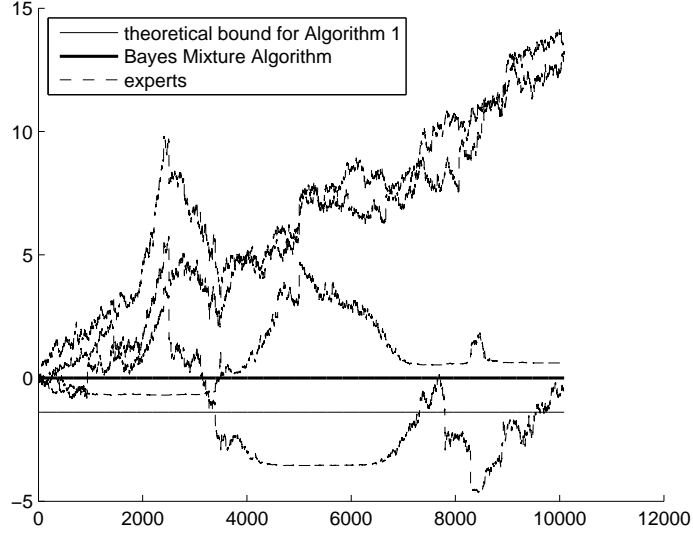
32

Figure 22: The difference between the cumulative loss of each of the 4 bookmakers and of the Bayes Mixture Algorithm on the tennis data.

Tables 1 and 2: the second column gives the maximal differences (25) and (26), respectively. The numbers preceded by "$\geq$" are the maximal differences corresponding to the best value of parameter chosen in hindsight, after seeing the data set. Therefore, the corresponding numbers involve "data snooping" and cannot serve as a fair measure of performance. The third column gives the theoretical performance guarantees (if available).

Algorithm	Maximal difference	Theoretical bound
Algorithm 1	1.1562	2.0794
WdAA ($c = 16/3$)	1.6619	11.0904
WdAA ($c = 1$)	1.1281	none of the form (2)
WkAA	≥ 1.8933	none of the form (2)
HA (expected)	≥ 2.3694	none of the form (2)
SAA-HA (expected)	≥ 2.3882	none of the form (2)
Follow the Leader Algorithm	2.7983	none
Simple Average Algorithm	2.5422	none
Bayes Mixture Algorithm	1.0602	none

Table 1: The maximal difference between the loss of each algorithm and the loss of the best expert for the football data (second column); the theoretical upper bound on this difference (third column).

Algorithm	Maximal difference	Theoretical bound
Algorithm 1	1.2021	1.3863
WdAA ($c = 4$)	2.4450	5.5452
WdAA ($c = 1$)	1.1089	none of the form (2)
WkAA	≥ 1.5059	none of the form (2)
HA (expected)	≥ 1.4153	none of the form (2)
SAA-HA (expected)	≥ 1.3909	none of the form (2)
Follow the Leader Algorithm	1.5597	none
Simple Average Algorithm	3.7928	none
Bayes Mixture Algorithm	4.6531	none

Table 2: The maximal difference between the loss of each algorithm and the loss of the best expert for the tennis data (second column); the theoretical upper bound on this difference (third column).