Jeffreys's law for general games of prediction: in search of a theory

A. P. Dawid and V. G. Vovk

December 22, 2009

Abstract

We are interested in the following version of Jeffreys's law: if two predictors are predicting the same sequence of events and either is doing a satisfactory job, they will make similar predictions in the long run. We give a classification of instances of Jeffreys's law, illustrated with examples.

1 Introduction

In this paper we are interested in games of prediction for which Jeffreys's law, as stated in the abstract, holds. Specific true instances of Jeffreys's law will be referred to as *Jeffreys theorems*.

In Section 2 we define several popular games of prediction and state Jeffreys theorems for the absolute-loss, square-loss, and bounded square-loss games. These results serve as illustrations for our taxonomy of Jeffreys theorems; namely, we distinguish between Jeffreys theorems of level 1 (weakest), level 2 (intermediate), and level 3 (strongest).

In Section 3 we show that in the case of so-called perfectly mixable games there is no difference between the three levels of Jeffreys theorems. Perfectly mixable games include, in particular, log-loss games and the bounded square-loss game.

In the next section, Section 4, we state level 2 Jeffreys theorems, which cover the log-loss and square-loss games (not necessarily bounded). In combination with the results of Section 3 this provides us with examples of level 3 Jeffreys theorems. Some of the results in Section 4 are explicit inequalities, not just statements of convergence.

The simple method of Section 4 does not work for the absolute-loss game. In Section 5 we will see that it is still possible to prove a Jeffreys theorem for this game, albeit only a level 1 one.

Perhaps the first instance of Jeffreys's law was proved by Blackwell and Dubins [2]; a pointwise version of their result was established in [3]. Results similar to ours but stated in terms of the algorithmic theory of randomness were earlier obtained in [9] (developing [6]) and [5] in the case of the log-loss

game, and in [11] (in essence developing [8]) in the case of the bounded square-loss game.

2 Taxonomy and examples of Jeffreys theorems

A *game of prediction* is a triple (Ω, Γ, ℓ), where Ω and Γ are arbitrary sets, called the *outcome space* and *prediction space*, respectively, and $\ell : \Omega \times \Gamma \to \mathbb{R}$ is called the *loss function*. The game is played according to the following perfect-information protocol.

COMPETITIVE PREDICTION PROTOCOL
Players: Nature, Predictor 1, Predictor 2, Sceptic
Protocol:
FOR $n = 1, 2, \ldots$:
    Predictor 1 and Predictor 2 announce $\gamma_n^{[1]} \in \Gamma$ and $\gamma_n^{[2]} \in \Gamma$.
    Sceptic announces $\tilde{\gamma}_n \in \Gamma$.
    Nature announces $\omega_n \in \Omega$.
END FOR

Three of the players, two Predictors and one Sceptic, are trying to predict the outcome ω_n to be announced by Nature. Sceptic is just like another Predictor, but he will be playing a special role in our story. At step n, Predictor 1 and Predictor 2 issue predictions $\gamma_n^{[1]}$ and $\gamma_n^{[2]}$, respectively. The Predictors can consult each other when making the predictions, and the pair $(\gamma_n^{[1]}, \gamma_n^{[2]})$ can be regarded as their joint prediction. After the two Predictors have announced, Sceptic issues his own prediction $\tilde{\gamma}_n$. Then Nature produces ω_n. Let $L_N^{[k]} := \sum_{n=1}^{N} \ell(\omega_n, \gamma_n^{[k]})$ be the cumulative loss to time N of Predictor k, $k = 1, 2$, and similarly $\tilde{L}_N$ for Sceptic.

The *absolute-loss game* is $(\mathbb{R}, \mathbb{R}, \ell)$ where $\ell(\omega, \gamma) := |\omega - \gamma|$. The next proposition states our first Jeffreys theorem.

Proposition 1. *Sceptic has a strategy in the absolute-loss game that guarantees*

$$\lim_{N \to \infty} \max \left(\frac{1}{\left| \gamma_N^{[1]} - \gamma_N^{[2]} \right|}, L_N^{[1]} - \tilde{L}_N, L_N^{[2]} - \tilde{L}_N \right) = \infty. \tag{1}$$

As usual, we set $1/0 := \infty$ in (1). For the proof of Proposition 1, see Section 5.

We call (1), perhaps with $|\gamma_N^{[1]} - \gamma_N^{[2]}|$ replaced by a different distance, a *level 1 Jeffreys theorem*. It says that for a sufficiently distant outcome ω_N, $N \gg 1$, at least one of the following three things happen: the two Predictors' predictions $\gamma_N^{[1]}$ and $\gamma_N^{[2]}$ are close to each other; Sceptic greatly outperforms Predictor 1 by time N; Sceptic greatly outperforms Predictor 2 by time N. The weakness of this statement is that no "stabilization" is guaranteed along a given infinite sequence of outcomes $\omega_1 \omega_2 \ldots$: it is possible that each one of the three terms of the disjunction will be violated infinitely often.

A stronger Jeffreys theorem, which we call a *level 2 Jeffreys theorem*, would say that

$$\lim_{N\to\infty}\left|\gamma_N^{[1]}-\gamma_N^{[2]}\right|=0 \text{ or } \lim_{N\to\infty}\max\left(L_N^{[1]}-\tilde{L}_N, L_N^{[2]}-\tilde{L}_N\right)=\infty. \tag{2}$$

An even stronger statement, which we call a *level 3 Jeffreys theorem*, would be

$$\lim_{N\to\infty}\left|\gamma_N^{[1]}-\gamma_N^{[2]}\right|=0 \text{ or } \lim_{N\to\infty}\left(L_N^{[1]}-\tilde{L}_N\right)=\infty \text{ or } \lim_{N\to\infty}\left(L_N^{[2]}-\tilde{L}_N\right)=\infty. \tag{3}$$

The following two propositions give examples of level 2 and level 3 Jeffreys theorems. The *square-loss game* is $(\mathbb{R},\mathbb{R},\ell)$ where $\ell(\omega,\gamma):=(\omega-\gamma)^2$.

Proposition 2. *Sceptic has a strategy in the square-loss game that guarantees (2).*

The *bounded square-loss game* is $([0,1],[0,1],\ell)$ where $\ell(\omega,\gamma):=(\omega-\gamma)^2$. (We fix specific bounds, 0 and 1, for outcomes and predictions, but our results generalize in a straightforward manner to any other bounds.)

Proposition 3. *Sceptic has a strategy in the bounded square-loss game that guarantees (3).*

Proposition 2 will be proved in Section 4, and it will imply Proposition 3 in combination with results of Section 3.

Counterexample

The *bounded absolute-loss game* is $([0,1],[0,1],\ell)$ where $\ell(\omega,\gamma):=|\omega-\gamma|$. The level 3 Jeffreys theorem does not hold for the bounded absolute-loss game:

Proposition 4. *Sceptic does not have a strategy that guarantees (3) in the bounded absolute-loss game.*

Proof. Suppose Sceptic has such a strategy and is playing it. Let Nature produce 0 and 1 independently with probability 1/2 each. Predictor 1 always predicts 0 and Predictor 2 always predicts 1. The restriction of Sceptic's strategy to $\omega_n\in\{0,1\}$ and $\gamma_n^{[1]},\gamma_n^{[2]}\in\{0,1\}$ is automatically measurable. We can see that $L_n^{[1]}-\tilde{L}_n$ and $L_n^{[2]}-\tilde{L}_n$ are martingales with bounded increments, and so tend to ∞ with probability zero (see [7], Theorem VII.5.1 and its corollary). Therefore, (3) happens with probability zero. □

The proof shows that Proposition 4 remains true for the restricted game $(\{0,1\},[0,1],\ell)$, $\ell(\omega,\gamma):=|\omega-\gamma|$.

3 Reductions between Jeffreys theorems

It appears that the main factor that determines which Jeffreys theorems hold for a particular game of prediction is the degree of convexity of the game. We might define a game to be convex if its prediction set Γ is a convex set in a linear space and its loss function $\ell(\omega, \gamma)$ is convex in $\gamma \in \Gamma$. However, this definition would be too narrow, since the predictions γ are usually just arbitrary labels. We start from introducing a much less arbitrary representation of games of prediction.

A *canonical prediction* is a function $\lambda : \Omega \to \mathbb{R}$ such that

$$\exists \gamma \in \Gamma \; \forall \omega \in \Omega : \lambda(\omega) = \ell(\omega, \gamma).$$

The *canonical representation* of the game (Ω, Γ, ℓ) is the pair (Ω, Λ) where Λ, called the *canonical prediction set*, is the set of all canonical predictions. We will not always distinguish between the game and its canonical representation and will usually consider games that are non-redundant in the sense that

$$(\lambda_1, \lambda_2 \in \Lambda \;\&\; \lambda_1 \le \lambda_2) \Longrightarrow \lambda_1 = \lambda_2. \tag{4}$$

A *superprediction* (resp. *subprediction*) is a function $\lambda : \Omega \to \mathbb{R}$ such that $\lambda \ge \lambda'$ (resp. $\lambda \le \lambda'$) for some canonical prediction λ'. The set of all superpredictions (resp. subpredictions) will be denoted $\overline{\Lambda}$ (resp. $\underline{\Lambda}$) and called the *superprediction set* (resp. *subprediction set*).

We will be interested in three notions of convexity for games of prediction:

- a game is *convex* if its superprediction set $\overline{\Lambda}$ is convex (equivalently, if a convex mixture of two canonical predictions is always a superprediction); this condition is always satisfied if Γ is a convex set and the loss function $\ell(\omega, \gamma)$ is convex in $\gamma \in \Gamma$;
- a game is *strictly convex* if a non-degenerate convex mixture of two canonical predictions is always an interior point of $\overline{\Lambda}$ (in the topology of uniform convergence);
- a game is *perfectly mixable* if, for some $\eta > 0$, the set $e^{-\eta\overline{\Lambda}}$ is convex.

For illustrative purposes it is convenient to consider the case where the game (Ω, Γ, ℓ) is *binary*, in the sense $\Omega = \{0, 1\}$. In this case Λ can be represented as the subset of $\mathbb{R}^2$ consisting of the points $(x, y) = (\lambda(0), \lambda(1))$ where λ ranges over Λ. An example is given as the curved line in Figure 1 below; the superpredictions are the points North-East of the line, and the subpredictions are the points South-West of the line.

It is easy to see that for perfectly mixable prediction games there is no real difference between the three levels of Jeffreys theorems:

Proposition 5. *Suppose Sceptic can guarantee (1) in the competitive prediction protocol for a perfectly mixable game. Then he can also guarantee (3) (and,* a fortiori*, (2)).*

Proof. Consider the generalization of the competitive prediction protocol in which there are infinitely many Predictors (called Experts and numbered by $k = 1, 2, \ldots$) instead of just two. Using the Aggregating Algorithm (see, e.g., [10], Subsection 2.1), for any sequence $p_1, p_2, \ldots$ of positive weights summing to 1 Sceptic can guarantee that his loss satisfies

$$\tilde{L}_N \le L_N^{[k]} + C \ln \frac{1}{p_k} \tag{5}$$

for all $N = 1, 2, \ldots$ and $k = 1, 2, \ldots$, where C is a constant depending on the prediction game.

Let Sceptic play a strategy that guarantees (1). We will construct a new strategy for Sceptic that guarantees (3). Consider the following doubly infinite set of experts:

- Expert $(k, 1)$, $k = 1, 2, \ldots$, plays as Sceptic until the difference $L_n^{[1]} - \tilde{L}_n$ exceeds 2^k; as soon as this happens (if it ever happens), he starts playing as Predictor 1;
- Expert $(k, 2)$ plays as Sceptic until the difference $L_n^{[2]} - \tilde{L}_n$ exceeds 2^k; as soon as this happens, he starts playing as Predictor 2.

The weights $p_{k,1}$ and $p_{k,2}$ assigned to these experts are $p_{k,1} = p_{k,2} = 2^{-k-1}$. Applied to these experts, the Aggregating Algorithm provides a new strategy for Sceptic that guarantees (3). Indeed, suppose the first of the three terms in (3) is false. Then, by (1), either the second or the third term in (3) becomes true when lim is replaced by lim sup. Suppose, for concreteness, it is the second term. For each k, Expert $(k, 1)$'s loss satisfies $L_N^{[k,1]} < L_N^{[1]} - 2^k$ from some N on, and so (5) implies that the Aggregating Algorithm's loss L_N satisfies

$$L_N \le L_N^{[k,1]} + C \ln \frac{1}{p_{k,1}} < L_N^{[1]} - 2^k + (C \ln 2)(k + 1)$$

for all k and from some N on. Letting $k \to \infty$, we can see that the second term of (3), with L_N in place of $\tilde{L}_N$, is true. □

Of course, Proposition 5 will continue to hold if the Euclidean distance in (1), (2), and (3) is replaced by any other distance.

Examples of perfectly mixable games

The bounded square-loss game is perfectly mixable ([10], Subsection 2.4).

Perhaps the most fundamental class of games of prediction is that of log-loss games. If (Ω, Γ, ℓ) is a *log-loss game*, Ω is a measurable space with a fixed σ-finite measure μ (more generally, $\mu = \mu_n$ may depend on n and be announced by a player, say Nature, at the beginning of step n of the game), Γ is the set of all measurable functions $\gamma : \Omega \to [0, \infty)$ satisfying $\int \gamma d\mu = 1$, and $\ell(\omega, \gamma) = -\ln \gamma(\omega)$. For log-loss games the loss function is allowed to take value

∞ ($-\ln 0 := \infty$). A simple and instructive special case to keep in mind is where μ is the counting measure on a countable Ω. The perfect mixability of log-loss games is a well-known fact, and the Aggregating Algorithm for them reduces to the Bayes rule (for details see, e.g., [10], Subsection 2.2).

For other examples of perfectly mixable games (such as the Kullback–Leibler game and Cover's game), see [10], Subsection 2.5.

4 Level 2 Jeffreys theorems

If λ_1 and λ_2 are canonical predictions and $\alpha \in (-1, 1)$, we set

$$\underline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2) := \frac{4}{1-\alpha^2} \sup \left\{ t \in \mathbb{R} : \frac{1-\alpha}{2}\lambda_1 + \frac{1+\alpha}{2}\lambda_2 - t \in \overline{\Lambda} \right\} \tag{6}$$

(the *lower α-divergence between λ_1 and λ_2*) and

$$\overline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2) := \frac{4}{1-\alpha^2} \inf \left\{ t \in \mathbb{R} : \frac{1-\alpha}{2}\lambda_1 + \frac{1+\alpha}{2}\lambda_2 - t \in \underline{\Lambda} \right\}$$

(the *upper α-divergence between λ_1 and λ_2*). The lower and upper divergence make take values $-\infty$ or ∞. We will be mostly interested in lower divergences (which for many interesting games coincides with upper divergences). In the case of binary (Ω, Γ, ℓ) this definition is illustrated in Figure 1 (notice that the difference between lower and upper α-divergences disappears for convex binary games; in such cases, we will sometimes write $D^{[\alpha]}(\lambda_1 \,\|\, \lambda_2)$ for the common value of $\underline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2)$ and $\overline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2)$ and omit the adjectives "lower" and "upper"). We will also write $\underline{D}^{[\alpha]}(\gamma_1 \,\|\, \gamma_2)$ and $\overline{D}^{[\alpha]}(\gamma_1 \,\|\, \gamma_2)$ for $\gamma_1, \gamma_2 \in \Gamma$, in the obvious sense.

Notice that, for strictly convex and non-redundant (in the sense of (4)) games,

$$\overline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2) \geq \underline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2) > 0,$$

for all $\lambda_1, \lambda_2 \in \Lambda$. For $\alpha = 0$ the lower (resp. upper) α-divergence is called the *lower* (resp. *upper*) *Hellinger distance*; the word "distance" is partly explained by its symmetry (although simplest examples show that there is no continuous function f such that $f(\underline{D}^{[0]})$ or $f(\overline{D}^{[0]})$ is a metric for every strictly convex game).

The values of lower and upper α-divergences for $\alpha = \pm 1$ are defined as their limits as $\alpha \to \pm 1$ when those limits exist. The lower (resp. upper) -1-divergence is called the *lower* (resp. *upper*) *Kullback–Leibler divergence* and is especially important.

Remark. It is not difficult to see that upper divergences can be very different from the corresponding lower divergences even for "nice" (in particular, strictly convex) games. For example, for the game $([-1, 1], [-1, 1], (\omega - \gamma)^4)$ the lower and upper Hellinger distances between the predictions -1 and 1 are different, 1 and 7. (Cf. [4], Lemma 3.)

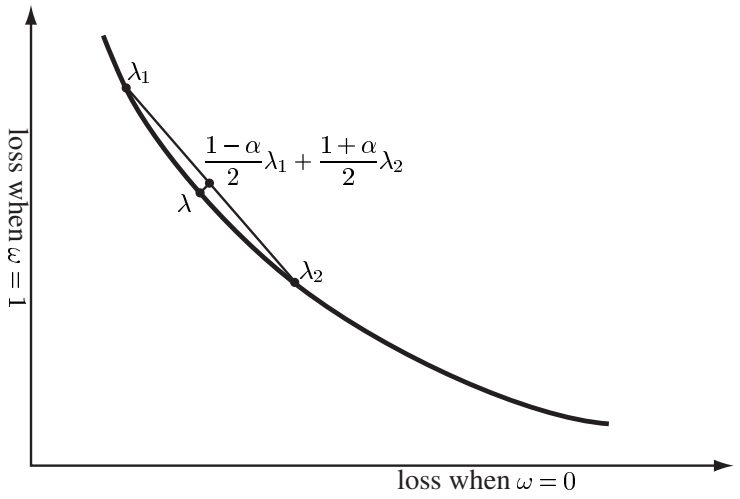

Figure 1: The interpretation of the α-divergence between canonical predictions λ_1 and λ_2 in the binary case: find the mean $\frac{1-\alpha}{2}\lambda_1 + \frac{1+\alpha}{2}\lambda_2$ of λ_1 and λ_2; find the intersection λ of the prediction set and the slope 1 line passing through the mean; multiply the horizontal (=vertical) distance between the mean and λ by $\frac{4}{1-\alpha^2}$.

The square-loss and log-loss games

In this subsections we will compute lower and upper divergences for two popular games of prediction defined earlier.

Lemma 1. *In the square-loss game,*

$$D^{[\alpha]}(\gamma_1 \,\|\, \gamma_2) = (\gamma_1 - \gamma_2)^2 \tag{7}$$

for all $\alpha \in [-1, 1]$ *and* $\gamma_1, \gamma_2 \in \mathbb{R}$.

Proof. It suffices to consider the case $\alpha \in (-1, 1)$. The statement of the lemma will follow from the fact that, for all $\omega \in \mathbb{R}$,

$$\frac{1-\alpha}{2}(\gamma_1 - \omega)^2 + \frac{1+\alpha}{2}(\gamma_2 - \omega)^2 - \frac{1-\alpha^2}{4}(\gamma_1 - \gamma_2)^2 = \left(\frac{1-\alpha}{2}\gamma_1 + \frac{1+\alpha}{2}\gamma_2 - \omega\right)^2.$$

If we set $t_1 := \gamma_1 - \omega$ and $t_2 := \gamma_2 - \omega$, the last equality simplifies to the obvious

$$\frac{1-\alpha}{2}t_1^2 + \frac{1+\alpha}{2}t_2^2 - \frac{1-\alpha^2}{4}(t_1 - t_2)^2 = \left(\frac{1-\alpha}{2}t_1 + \frac{1+\alpha}{2}t_2\right)^2. \qquad \square$$

Lemma 2. *In any log-loss game,*

$$D^{[\alpha]}(\gamma_1 \,\|\, \gamma_2) = -\frac{4}{1-\alpha^2} \ln \int_\Omega (\gamma_1(\omega))^{\frac{1-\alpha}{2}} (\gamma_2(\omega))^{\frac{1+\alpha}{2}} \mu(d\omega) \tag{8}$$

for all $\alpha \in (-1, 1)$ *and* $\gamma_1, \gamma_2 \in \Gamma$.

Proof. The left-hand side of (8) can be written as $\frac{4}{1-\alpha^2} t$ where t is defined from the condition that, for some $\gamma \in \Gamma$ and all $\omega \in \Omega$,

$$-\frac{1-\alpha}{2} \ln \gamma_1(\omega) - \frac{1+\alpha}{2} \ln \gamma_2(\omega) - t = -\ln \gamma(\omega).$$

Deducing

$$\int_\Omega \gamma d\mu = \int_\Omega (\gamma_1(\omega))^{\frac{1-\alpha}{2}} (\gamma_2(\omega))^{\frac{1+\alpha}{2}} \mu(d\omega) e^t,$$

substituting 1 for $\int \gamma d\mu$, and solving the resulting equation for t, we obtain the statement of the lemma. □

The standard definition of the α-divergence for the log-loss game (see, e.g., [1], p. 57) is

$$D^{(\alpha)}(\gamma_1 \,\|\, \gamma_2) = \frac{4}{1-\alpha^2} \left(1 - \int_\Omega (\gamma_1(\omega))^{\frac{1-\alpha}{2}} (\gamma_2(\omega))^{\frac{1+\alpha}{2}} \mu(d\omega) \right);$$

it is clear that this will differ little from (8) when γ_1 and γ_2 are close in a suitable sense. The inequality $\ln x \le x - 1$ implies $D^{(\alpha)} \le D^{[\alpha]}$.

Level 2 and level 3 Jeffreys theorems

This is our most general level 2 Jeffreys theorem:

Proposition 6. *For each* $\alpha \in (-1, 1)$ *and* $\epsilon > 0$ *Sceptic has a strategy that guarantees*

$$\frac{1-\alpha^2}{4} \sum_{n=1}^{N} \underline{D}^{[\alpha]} \left(\gamma_n^{[1]} \,\|\, \gamma_n^{[2]} \right) \le \frac{1-\alpha}{2} L_N^{[1]} + \frac{1+\alpha}{2} L_N^{[2]} - \tilde{L}_N + \epsilon \tag{9}$$

Proof. The strategy is obvious: according to (6), at step n Sceptic can choose a canonical prediction λ satisfying

$$\lambda \le \frac{1-\alpha}{2} \lambda_1 + \frac{1+\alpha}{2} \lambda_2 - \frac{1-\alpha^2}{4} \underline{D}^{[\alpha]}(\lambda_1 \,\|\, \lambda_2) + \epsilon 2^{-n}$$

(λ_1 and λ_2 being the canonical predictions corresponding to $\gamma_n^{[1]}$ and $\gamma_n^{[2]}$). Summing over the first N steps, we obtain (9). □

Specializing (9) to the case $\alpha = 0$ and the square-loss game gives

$$\frac{1}{4}\sum_{n=1}^{N}\left(\gamma_n^{[1]} - \gamma_n^{[2]}\right)^2 \le \frac{L_N^{[1]} + L_N^{[2]}}{2} - \tilde{L}_N + \epsilon.$$

This implies a stronger version of the level 2 Jeffreys theorem (2):

$$\sum_{n=1}^{\infty}\left(\gamma_n^{[1]} - \gamma_n^{[2]}\right)^2 < \infty \text{ or } \lim_{N\to\infty} \max\left(L_N^{[1]} - \tilde{L}_N, L_N^{[2]} - \tilde{L}_N\right) = \infty.$$

In combination with the proof of Proposition 5, this implies the stronger form

$$\sum_{n=1}^{\infty}\left(\gamma_n^{[1]} - \gamma_n^{[2]}\right)^2 < \infty \text{ or } \lim_{N\to\infty}\left(L_N^{[1]} - \tilde{L}_N\right) = \infty \text{ or } \lim_{N\to\infty}\left(L_N^{[2]} - \tilde{L}_N\right) = \infty \tag{10}$$

of the level 3 Jeffreys theorem (3) for the bounded square-loss game.

For the log-loss game, we obtain (10) with the Hellinger distance $D^{[0]}(\gamma_n^{[1]} \,\|\, \gamma_n^{[2]})$, or the standard Hellinger distance $D^{(0)}(\gamma_n^{[1]} \,\|\, \gamma_n^{[2]})$, in place of $(\gamma_n^{[1]} - \gamma_n^{[2]})^2$.

5 Level 1 Jeffreys theorems

The main goal of this section is to prove Proposition 1. In the absolute-loss game, the divergence between any two predictions is 0, and so the methods of the previous section are not applicable.

First we describe a strategy for Sceptic that will later be shown to ensure (1). Let $f : [0,\infty) \to [0,1/2)$ be a strictly increasing and concave function satisfying $f(0) = 0$ and $f(\infty) < 1/2$; see Figure 2. Later it will be convenient to extend f to $(-\infty,\infty)$ by the central symmetry w.r. to the origin O (so that $f : (-\infty,\infty) \to (-1/2, 1/2)$ is an odd function).

Suppose just before step $n = 1, 2, \ldots$ of the competitive prediction protocol we have $D_{n-1} := L_{n-1}^{[1]} - L_{n-1}^{[2]} \ge 0$ (the case where $L_{n-1}^{[1]} \le L_{n-1}^{[2]}$ will later be reduced to this one). Sceptic's move can be represented as

$$\tilde{\gamma}_n := (1 - t_n)\gamma_n^{[1]} + t_n\gamma_n^{[2]},$$

where t_n will be chosen later from the interval $[0, 1/2]$. Set

$$\begin{aligned} d_n &:= \left|\gamma_n^{[1]} - \gamma_n^{[2]}\right| \in [0,1], \\ \bar{L}_n &:= \frac{L_n^{[1]} + L_n^{[2]}}{2}, \\ \bar{\ell}_n &:= \frac{\ell(\omega_n, \gamma_n^{[1]}) + \ell(\omega_n, \gamma_n^{[2]})}{2}. \end{aligned}$$

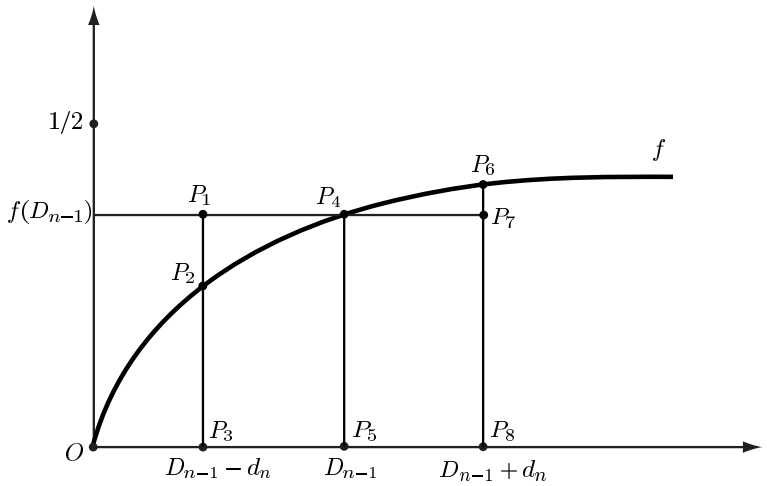

Figure 2: The function f from the proof of Proposition 1.

If the actual outcome ω_n is in favour of Predictor 1,

$$\ell(\omega_n, \gamma_n^{[1]}) \le \ell(\omega_n, \gamma_n^{[2]}),$$

the difference $L_n^{[1]} - L_n^{[2]}$ between the losses of the two Predictors will decrease to $D_n = D_{n-1} - d_n$ and the difference $\tilde{L}_n - \bar{L}_n$ will increase by

$$\ell(\omega_n, \tilde{\gamma}_n) - \bar{\ell}_n = (1 - t_n)\left(\bar{\ell}_n - \frac{d_n}{2}\right) + t_n\left(\bar{\ell}_n + \frac{d_n}{2}\right) - \bar{\ell}_n = \left(t_n - \frac{1}{2}\right) d_n.$$

So in fact it will decrease as $t_n \le 1/2$. Let us set $t_n := 1/2 - f(D_{n-1})$. The difference $\tilde{L}_n - \bar{L}_n$ will decrease by the area of the rectangle $P_3P_5P_4P_1$.

If the actual outcome ω_n is in favour of Predictor 2,

$$\ell(\omega_n, \gamma_n^{[1]}) \ge \ell(\omega_n, \gamma_n^{[2]}),$$

the difference between the losses of the two Predictors will increase to $D_n = D_{n-1} + d_n$ and the difference $\tilde{L}_n - \bar{L}_n$ will increase by

$$\begin{aligned}\ell(\omega_n, \tilde{\gamma}_n) - \bar{\ell}_n = (1 - t_n)\left(\bar{\ell}_n + \frac{d_n}{2}\right) + t_n\left(\bar{\ell}_n - \frac{d_n}{2}\right) - \bar{\ell}_n \\ = \left(\frac{1}{2} - t_n\right) d_n = f(D_{n-1}) d_n,\end{aligned}$$

i.e., by the area of the rectangle $P_5P_8P_7P_4$.

We can see that in both cases, $D_n = D_{n-1} \pm d_n$, the difference $\tilde{L}_n - \bar{L}_n$ increases by $\int_{D_{n-1}}^{D_n} f$ minus the area A_n of a curvilinear triangle ($P_1P_2P_4$ if $D_n = D_{n-1} - d_n$ and $P_4P_7P_6$ if $D_n = D_{n-1} + d_n$). Now extend f to the

whole of $(-\infty, \infty)$ as an odd function. Suppose that $D_{n-1} \le 0$ and, moreover, $D_{n-1} + d_n \le 0$. Applying the same argument as above but with the roles of Predictor 1 and Predictor 2 interchanged, we can see that the difference $\tilde{L}_n - \bar{L}_n$ again increases by $\int_{D_{n-1}}^{D_n} f$ minus the area A_n of a curvilinear triangle. It is easy to check that the difference $\tilde{L}_n - \bar{L}_n$ will change in the same way also in the case where $D_{n-1} \ge 0$ but $D_{n-1} - d_n \le 0$ and in the case where $D_{n-1} \le 0$ but $D_{n-1} + d_n \ge 0$. Since $\tilde{L}_N - \bar{L}_N$ is the cumulative increase in $\tilde{L}_n - \bar{L}_n$ over $n = 1, \dots, N$, we can see that

$$\tilde{L}_N - \bar{L}_N = \int_0^{D_N} f - \sum_{n=1}^{N} A_n.$$

It remains to consider two cases:

$\sum_{n=1}^{\infty} A_n < \infty$**:** In this case, $A_N \to 0$ and so

$$\max\left(\frac{1}{\left|\gamma_N^{[1]} - \gamma_N^{[2]}\right|}, |D_N|\right) \to \infty$$

as $N \to \infty$. The sequence $N = 1, 2, \dots$ can be split into three subsequences such that $|\gamma_N^{[1]} - \gamma_N^{[2]}| \to 0$ along the first, $D_N \to \infty$ along the second, and $D_N \to -\infty$ along the third. It suffices to show that (1) holds along the second subsequence (the case of the third subsequence is analogous, and the case of the first subsequence is trivial). Assuming $D_N > 0$, we can see that along the second subsequence:

$$\begin{aligned}\tilde{L}_N &= \bar{L}_N + \int_0^{D_N} f - \sum_{n=1}^{N} A_n \\ &\le \frac{L_N^{[1]} + L_N^{[1]} - D_N}{2} + \int_0^{D_N} f \le L_N^{[1]} + D_N\left(f(\infty) - \frac{1}{2}\right),\end{aligned}$$

and so $L_N^{[1]} - \tilde{L}_N \to \infty$.

$\sum_{n=1}^{\infty} A_n = \infty$**:** In this case we have along the subsequence of N for which $D_N \ge 0$:

$$\begin{aligned}\tilde{L}_N &= \bar{L}_N + \int_0^{D_N} f - \sum_{n=1}^{N} A_n \\ &= \frac{L_N^{[1]} + L_N^{[1]} - D_N}{2} + \int_0^{D_N} f - \sum_{n=1}^{N} A_n \le L_N^{[1]} - \sum_{n=1}^{N} A_n,\end{aligned}$$

and so $L_N^{[1]} - \tilde{L}_N \to \infty$. Similarly, $L_N^{[2]} - \tilde{L}_N \to \infty$ along the subsequence of N for which $D_N \le 0$. Therefore, (1) holds.

Convex games

It is easy to see that the proof of Proposition 1 is applicable to any convex game. For any such game Sceptic has a strategy in the competitive prediction protocol that guarantees

$$\lim_{N\to\infty} \max\left(\frac{1}{\left|\lambda\left(\omega_N, \gamma_N^{[1]}\right) - \lambda\left(\omega_N, \gamma_N^{[2]}\right)\right|}, L_N^{[1]} - \tilde{L}_N, L_N^{[2]} - \tilde{L}_N\right) = \infty.$$

Acknowledgements

We are grateful to Akio Fujiwara for a useful discussion, to Glenn Shafer for his advice, and to participants of WITMSE 2009 for their comments. This work was supported in part by EPSRC (grant EP/F002998/1).

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2000.

[2] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–886, 1962.

[3] A. P. Dawid. Calibration-based empirical probability (with discussion). *Annals of Statistics*, 13:1251–1285, 1985.

[4] A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59:77–93, 2007.

[5] A. Fujiwara. Randomness criteria in terms of α-divergences. *IEEE Transactions on Information Theory*, 54:1252–1261, 2008.

[6] Y. M. Kabanov, R. S. Liptser, and A. N. Shiryaev. To the question of absolute continuity and singularity of probability measures. *Mathematics of the USSR—Sbornik*, 33:203–221, 1977.

[7] A. N. Shiryaev. *Probability*. Springer, New York, second edition, 1996. Third Russian edition published in 2004.

[8] K. Skouras and A. P. Dawid. On efficient point prediction systems. *Journal of the Royal Statistical Society* B, 60:765–780, 1998.

[9] V. G. Vovk. On a randomness criterion. *Soviet Mathematics Doklady*, 35:656–660, 1987.

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

[11] V. G. Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261:57–79, 2001.