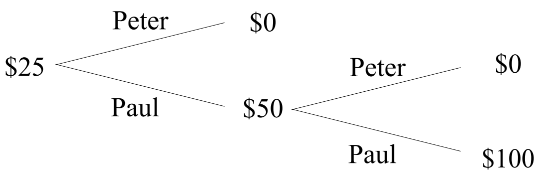
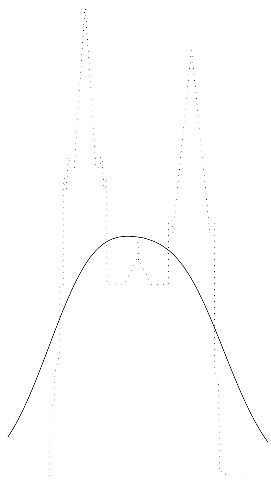
Defensive Forecasting

Vladimir Vovk   Akimichi Takemura   Glenn Shafer





The Game-Theoretic Probability and Finance Project

Working Paper #8

First posted September 29, 2004. Last revised January 22, 2005.

Project web site:
http://www.probabilityandfinance.com

Abstract

We consider how to make probability forecasts of binary labels. Our main mathematical result is that for any continuous (in particular, any computable)gambling strategy used for detecting disagreement between the forecasts and the actual labels, there exists a forecasting strategy whose forecasts are ideal as far as this gambling strategy is concerned. A forecasting strategy obtained in this way from a gambling strategy demonstrating a strong law of large numbers is simplified and studied empirically. This working paper is the full version of a paper to be published in the AI & Statistics 2005 proceedings.

Contents

1 Introduction

Probability forecasting can be thought of as a game between two players, Forecaster and Reality:

Basic Binary Forecasting Protocol
Players: Reality, Forecaster
FOR $n = 1, 2, \dots$:
  Reality announces $x_n \in \mathbf{X}$.
  Forecaster announces $p_n \in [0, 1]$.
  Reality announces $y_n \in \{0, 1\}$.

On each round, Forecaster predicts Reality's move y_n chosen from the *label space*, always taken to be $\{0, 1\}$ in this paper. His move, the *probability forecast* p_n, can be interpreted as the probability he attaches to the event $y_n = 1$. To help Forecaster, Reality presents him with an *object* x_n at the beginning of the round; x_n are chosen from an *object space* $\mathbf{X}$.

Forecaster's goal is to produce p_n that agree with the observed y_n. Various results of probability theory, in particular limit theorems (such as the weak and strong laws of large numbers, the law of the iterated logarithm, and the central limit theorem) and large-deviation inequalities (such as Hoeffding's inequality), describe different aspects of agreement between p_n and y_n. For example, according to the strong law of large numbers, we expect that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) = 0. \quad (1)$$

Such results will be called *laws of probability* and the existing body of laws of probability will be called *classical probability theory.* Historically, laws of probability form the core of probability theory.

In §2, following [18], we formalize Forecaster's goal by adding a third player, Skeptic, who is allowed to gamble at the odds given by Forecaster's probabilities. We state a result from [22] and [18] suggesting that Skeptic's gambling strategies can be used as tests of agreement between p_n and y_n and that all tests of agreement between p_n and y_n can be expressed as Skeptic's gambling strategies. Therefore, the forecasting protocol with Skeptic provides an alternative way of stating laws of probability.

As demonstrated in [18], many standard proof techniques developed in classical probability theory can be translated into computable strategies for Skeptic; all such strategies are continuous. In §3 we show that for any continuous strategy $\mathcal{S}$ for Skeptic there exists a strategy $\mathcal{F}$ for Forecaster such that $\mathcal{S}$ does not detect any disagreement between the y_n and the p_n produced by $\mathcal{F}$. This result is a "meta-theorem" that allows one to move from laws of probability to forecasting algorithms: as soon as a law of probability is expressed as a continuous strategy for Skeptic, we have a forecasting algorithm that guarantees that this law will hold; there are no assumptions about Reality, who may play adversarially.

Our meta-theorem is of any interest only if one can find sufficiently interesting laws of probability (expressed as gambling strategies) that can serve as its input. In §4 we apply it to the important properties of unbiasedness in the large and small of the forecasts p_n ((1) is an asymptotic version of the former). The resulting forecasting strategy is automatically unbiased, no matter what data $x_1, y_1, x_2, y_2, \ldots$ is observed.

In §5 we simplify the algorithm obtained in §4 and demonstrate its performance on some artificially generated data sets.

2 The gambling framework for testing probability forecasts

Skeptic is allowed to bet at the odds defined by Forecaster's probabilities, and he refutes the probabilities if he multiplies his capital manyfold. This is formalized as a perfect-information game in which Skeptic plays against a team composed of Forecaster and Reality:

Binary Forecasting Game I
Players: Reality, Forecaster, Skeptic
Protocol:
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Forecaster announces $p_n \in [0, 1]$.
  Skeptic announces $s_n \in \mathbb{R}$.
  Reality announces $y_n \in \{0, 1\}$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p_n)$.
Restriction on Skeptic: Skeptic must choose the s_n so that his capital is always nonnegative ($\mathcal{K}_n \geq 0$ for all n) no matter how the other players move.

This is a perfect-information protocol; the players move in the order indicated, and each player sees the other player's moves as they are made. It specifies both an initial value for Skeptic's capital ($\mathcal{K}_0 = 1$) and a lower bound on its subsequent values ($\mathcal{K}_n \geq 0$).

Our interpretation, which will be called the *testing interpretation*, of Binary Forecasting Game I is that $\mathcal{K}_n$ measures the degree to which Skeptic has shown Forecaster to do a bad job of predicting y_i, $i = 1, \ldots, n$.

2.1 Validity and universality of the testing interpretation

As explained in [18], the testing interpretation is valid and universal in an important sense. Let us assume, for simplicity, that objects are absent (formally, that $|\mathbf{X}| = 1$). In the case where Forecaster starts from a probability measure P on $\{0, 1\}^\infty$ and obtains his forecasts $p_n \in [0, 1]$ as conditional probabilities under P that $y_n = 1$ given $y_1, \ldots, y_{n-1}$, we have a standard way of testing P and, therefore, p_n: choose an event $A \subseteq \{0, 1\}^\infty$ (the *critical region*) with a small $P(A)$ and reject P if A happens. The testing interpretation satisfies the following two properties:

Validity Suppose Skeptic's strategy is measurable and p_n are obtained from P; $\mathcal{K}_n$ then form a nonnegative martingale w.r. to P. According to Doob's inequality [22, 6], for any positive constant C, $\sup_n \mathcal{K}_n \geq C$ with P-probability at most $1/C$. (If Forecaster is doing a bad job according to the testing interpretation, he is also doing a bad job from the standard point of view.)

Universality According to Ville's theorem ([18], §8.5), for any positive constant ϵ and any event $A \subseteq \{0,1\}^\infty$ such that $P(A) < \epsilon$, Skeptic has a measurable strategy that ensures $\liminf_{n\to\infty} \mathcal{K}_n > 1/\epsilon$ whenever A happens, provided p_n are computed from P. (If Forecaster is doing a bad job according to the standard point of view, he is also doing a bad job according to the testing interpretation.) In the case $P(A) = 0$, Skeptic actually has a measurable strategy that ensures $\lim_{n\to\infty} \mathcal{K}_n = \infty$ on A.

The universality of the gambling scenario of Binary Forecasting Game I is its most important advantage over von Mises's gambling scenario based on subsequence selection; it was discovered by Ville [22].

2.2 Constructiveness of the gambling framework

In [18] we constructed Skeptic's strategies that made him rich when the statement of any of several key laws of probability theory was violated. The constructions were explicit and lead to computable gambling strategies. We conjecture that every natural result of classical probability theory leads to a computable strategy for Skeptic.

Since Brouwer's work on intuitionist mathematics it is widely accepted that only continuous functions can be computable (this is Brouwer's *continuity principle* [3]; for a modern statement, see [11], §22). There are also idealized definitions of computability in terms of computational models able to perform operation with real numbers with infinite accuracy in unit time (such as [1]), and the functions computed in such weaker senses need not be computable. Our conjecture asserts computability in the stronger Brouwer's sense.

Remark According to [19], Brouwer introduced his continuity principle in 1916. The following quote from Borel's 1912 paper [2] is popular (see, e.g., [23], p. 64): "a function cannot be calculable unless it is continuous"; Borel, however, was interested only in the values that computable functions take on the computable values of the argument.

3 Defeating Skeptic

In this section we prove the main (albeit very simple) mathematical result of this paper: for any continuous strategy for Skeptic there exists a strategy for

Forecaster that does not allow Skeptic's capital to grow, regardless of what Reality is doing. Actually, our result will be even stronger: we will have Skeptic announce his strategy for each round before Forecaster's move on that round rather than making him announce his full strategy at the beginning of the game, and we will drop the restriction on Skeptic. Therefore, we consider the following perfect-information game that pits Forecaster against the two other players:

Binary Forecasting Game II
Players: Reality, Forecaster, Skeptic
Protocol:
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \dots$:
  Reality announces $x_n \in \mathbf{X}$.
  Skeptic announces continuous $S_n : [0,1] \to \mathbb{R}$.
  Forecaster announces $p_n \in [0,1]$.
  Reality announces $y_n \in \{0,1\}$.
  $\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n)$.

Theorem 1 *Forecaster has a strategy in Binary Forecasting Game II that ensures* $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \geq \cdots$.

Proof Forecaster can use the following strategy to ensure $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \cdots$:

- if the function $S_n(p)$ takes the value 0, choose p_n so that $S_n(p_n) = 0$;
- if S_n is always positive, take $p_n := 1$;
- if S_n is always negative, take $p_n := 0$. ∎

Remark In this paper computability serves only to motivate the assumption of continuity of S_n; only in this remark we briefly discuss computability issues. The proof of Theorem 1 shows that for computability of the constructed strategy for Forecaster we need more than the computability of S_n; namely, we need an oracle that for each point p tells us the sign of $S_n(p)$. Without such an oracle we can only claim that, for an arbitrary accuracy ϵ, either we can find, with accuracy ϵ, p_n ensuring $\mathcal{K}_n \leq \mathcal{K}_{n-1}$ or we can find p_n ensuring $\mathcal{K}_n \leq \mathcal{K}_{n-1} + \epsilon$. For an example showing that such an oracle is necessary, see [11] (Figure 5).

4 Examples of gambling strategies

In this section we discuss strategies for Forecaster obtained by Theorem 1 from different strategies for Skeptic; the former will be called *defensive forecasting strategies*. There are many results of classical probability theory that we could use, but we will concentrate on the simple strategy described in [18], p. 69, for proving the strong law of large numbers.

If $S_n(p) = S_n$ does not depend on p, the strategy from the proof of Theorem 1 makes Forecaster choose

$$p_n := \begin{cases} 0 & \text{if } S_n < 0 \\ 1 & \text{if } S_n > 0 \\ 0 \text{ or } 1 & \text{if } S_n = 0. \end{cases}$$

The basic procedure described in [18] (p. 69) is as follows. Let $\epsilon \in (0, 0.5]$ be a small number (expressing our tolerance to violations of the strong law of large numbers). In Binary Forecasting Game I, Skeptic can ensure that

$$\sup_n \mathcal{K}_n < \infty \implies \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) \le \epsilon \tag{2}$$

using the strategy $s_n = s_n^\epsilon := \epsilon \mathcal{K}_{n-1}$. Indeed, since

$$\mathcal{K}_n = \prod_{i=1}^{n} (1 + \epsilon(y_i - p_i)),$$

on the paths where $\mathcal{K}_n$ is bounded we have

$$\prod_{i=1}^{n} (1 + \epsilon(y_i - p_i)) \le C,$$

$$\sum_{i=1}^{n} \ln(1 + \epsilon(y_i - p_i)) \le \ln C,$$

$$\epsilon \sum_{i=1}^{n} (y_i - p_i) - \epsilon^2 \sum_{i=1}^{n} (y_i - p_i)^2 \le \ln C,$$

$$\epsilon \sum_{i=1}^{n} (y_i - p_i) \le \ln C + \epsilon^2 n,$$

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) \le \frac{\ln C}{\epsilon n} + \epsilon$$

(we have used the fact that $\ln(1 + t) \ge t - t^2$ when $|t| \le 0.5$). If Skeptic wants to ensure

$$\sup_n \mathcal{K}_n < \infty \implies -\epsilon \le \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) \le \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) \le \epsilon,$$

he can use the strategy $s_n := (s_n^\epsilon + s_n^{-\epsilon})/2$, and if he wants to ensure

$$\sup_n \mathcal{K}_n < \infty \implies \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) = 0, \tag{3}$$

he can use a convex mixture of $(s_n^{\epsilon} + s_n^{-\epsilon})/2$ over a sequence of ϵ converging to zero. There are also standard ways of strengthening (3) to

$$\liminf_{n\to\infty} \mathcal{K}_n < \infty \implies \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i) = 0;$$

for details, see [18].

In the rest of this section we will draw on the excellent survey [5]. We will see how Forecaster defeats increasingly sophisticated strategies for Skeptic.

4.1 Unbiasedness in the large

Following Murphy and Epstein [12], we say that Forecaster is *unbiased in the large* if (1) holds. Let us first consider the one-sided relaxed version of this property

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i) \le \epsilon. \tag{4}$$

The strategy for Skeptic described above, $S_n(p) := \epsilon\mathcal{K}_n$, leads to Forecaster always choosing $p_n := 1$; (4) is then satisfied in a trivial way.

Forecaster's strategy corresponding to the two-sided version

$$-\epsilon \le \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i) \le \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(y_i - p_i) \le \epsilon \tag{5}$$

is not much more reasonable. Indeed, it can be represented as follows. The initial capital 1 is split evenly between two accounts, and Skeptic gambles with the two accounts separately. If at the outset of round n the capital on the first account is $\mathcal{K}_{n-1}^1$ and the capital on the second account is $\mathcal{K}_{n-1}^2$, Skeptic plays $s_n^1 := \epsilon\mathcal{K}_{n-1}^1$ with the first account and $s_n^2 := -\epsilon\mathcal{K}_{n-1}^2$ with the second account; his total move is

$$S_n(p) := \epsilon\mathcal{K}_{n-1}^1 - \epsilon\mathcal{K}_{n-1}^2 = \epsilon\left(\prod_{i=1}^{n-1}(1+\epsilon(y_i - p_i)) - \prod_{i=1}^{n-1}(1+\epsilon(p_i - y_i))\right).$$

Therefore, Forecaster's move is $p_n := 1$ if

$$\sum_{i=1}^{n-1}\ln(1+\epsilon(y_i - p_i)) > \sum_{i=1}^{n-1}\ln(1+\epsilon(p_i - y_i)),$$

$p_n := 0$ if

$$\sum_{i=1}^{n-1}\ln(1+\epsilon(y_i - p_i)) < \sum_{i=1}^{n-1}\ln(1+\epsilon(p_i - y_i)),$$

and p_n can be chosen arbitrarily in the case of equality. The limiting form of this strategy as $\epsilon \to 0$ is: Forecaster's move is $p_n := 1$ if

$$\sum_{i=1}^{n-1}(y_i - p_i) > 0,$$

$p_n := 0$ if

$$\sum_{i=1}^{n-1}(y_i - p_i) < 0,$$

and p_n can be chosen arbitrarily in the case of equality.

We can see that unbiasedness in the large does not lead to interesting forecasts: Forecaster fulfils his task too well. In the one-sided case (4), he always chooses $p_n := 1$ making

$$\sum_{i=1}^{n}(y_i - p_i)$$

as small as possible. In the two-sided case (5) with $\epsilon \to 0$, he manages to guarantee that

$$\left|\sum_{i=1}^{n}(y_i - p_i)\right| \le 1. \tag{6}$$

His goals are achieved with categorical forecasts, $p_n \in \{0, 1\}$.

In the rest of this section we consider the more interesting case where $S_n(p)$ depends on p.

4.2 Unbiasedness in the small

We now consider a subtler requirement that forecasts should satisfy, which we introduce informally. We say that the forecasts p_n are *unbiased in the small* (or reliable, or valid, or well calibrated) if, for any $p^* \in [0, 1]$,

$$\frac{\sum_{i=1,\dots,n:p_i \approx p^*} y_i}{\sum_{i=1,\dots,n:p_i \approx p^*} 1} \approx p^* \tag{7}$$

provided $\sum_{i=1,\dots,n:p_i \approx p^*} 1$ is not too small.

Let us first consider just one value for p^*. Instead of the "crisp" point p^* we will consider a "fuzzy point" $I : [0, 1] \to [0, 1]$ such that $I(p^*) = 1$ and $I(p) = 0$ for all p outside a small neighborhood of p^*. A standard choice would be something like $I := \mathbb{I}_{[p_-,p_+]}$, where $[p_-, p_+]$ is a short interval containing p^* and $\mathbb{I}_{[p_-,p_+]}$ is its indicator function, but we will want I to be continuous (it can, however, be arbitrarily close to $\mathbb{I}_{[p_-,p_+]}$).

The strategy for Skeptic ensuring (2) can be modified as follows. Let $\epsilon \in (0, 0.5]$ be again a small number. Now we consider the strategy $S_n(p) = S_n^{\epsilon,I}(p) := \epsilon I(p)\mathcal{K}_{n-1}$. Since

$$\mathcal{K}_n = \prod_{i=1}^{n}(1 + \epsilon I(p_i)(y_i - p_i)),$$

on the paths where $\mathcal{K}_n$ is bounded we have

$$\prod_{i=1}^{n}(1+\epsilon I(p_i)(y_i-p_i)) \le C,$$

$$\sum_{i=1}^{n}\ln(1+\epsilon I(p_i)(y_i-p_i)) \le \ln C,$$

$$\epsilon\sum_{i=1}^{n} I(p_i)(y_i-p_i) - \epsilon^2\sum_{i=1}^{n} I^2(p_i)(y_i-p_i)^2 \le \ln C,$$

$$\epsilon\sum_{i=1}^{n} I(p_i)(y_i-p_i) - \epsilon^2\sum_{i=1}^{n} I(p_i) \le \ln C$$

(the last step involves replacing $I^2(p_i)$ with $I(p_i)$; the loss of precision is not great if I is close to $\mathbb{I}_{[p_-,p_+]}$),

$$\epsilon\sum_{i=1}^{n} I(p_i)(y_i-p_i) \le \ln C + \epsilon^2\sum_{i=1}^{n} I(p_i),$$

$$\frac{\sum_{i=1}^{n} I(p_i)(y_i-p_i)}{\sum_{i=1}^{n} I(p_i)} \le \frac{\ln C}{\epsilon\sum_{i=1}^{n} I(p_i)} + \epsilon.$$

The last inequality shows that the mean of y_i for p_i close to p^* is close to p^* provided we have observed sufficiently many such p_i; its interpretation is especially simple when I is close to $\mathbb{I}_{[p_-,p_+]}$.

In general, we may consider a mixture of $S_n^{\epsilon,I}(p)$ and $S_n^{-\epsilon,I}(p)$ for different values of ϵ and for different I covering all $p^* \in [0,1]$. If we make sure that the mixture is continuous (which is always the case for continuous I and finitely many ϵ and I), Theorem 1 provides us with forecasts that are unbiased in the small.

4.3 Using the objects

Unbiasedness, even in the small, is only a necessary but far from sufficient condition for good forecasts: for example, a forecaster who ignores the objects x_n can be perfectly calibrated, no matter how much useful information x_n contain. (Cf. the discussion of resolution in [5]; we prefer not to use the term "resolution", which is too closely connected with the very special way of probability forecasting based on sorting and labeling.) It is easy to make the algorithm of the previous subsection take the objects into account: we can allow the test functions I to depend not only on p but also on the current object x_n; $S_n(p)$ then becomes a mixture of

$$S_n^{\epsilon,I}(p) := \epsilon I(p,x_n)\prod_{i=1}^{n-1}(1+\epsilon I(p_i,x_i)(y_i-p_i))$$

and $S_n^{-\epsilon,I}(p)$ (defined analogously) over ϵ and I.

Remark Another popular example is: if the observed sequence of labels is $1, 0, 1, 0, \dots$, the sequences of forecasts $0.5, 0.5, \dots$ and $1, 0, 1, 0, \dots$ are both unbiased in the small, but the first of them is not as good as the second. For this example to be covered by our discussion, the labels $y_n = n \mod 2$ should be complemented by objects, $x_n := n$.

4.4 Relation to a standard counter-example

Suppose, for simplicity, that objects are absent ($|\mathbf{X}| = 1$). The standard construction from Dawid [4] showing that no forecasting strategy produces forecasts p_n that are unbiased in the small for all sequences is as follows. Define an infinite sequence $y_1, y_2, \dots$ recursively by

$$y_n := \begin{cases} 1 & \text{if } p_n < 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where p_n is the forecast produced by the forecasting strategy after seeing $y_1, \dots, y_{n-1}$. For the forecasts $p_n < 0.5$ we always have $y_n = 1$ and for the forecasts $p_n \geq 0.5$ we always have $y_n = 0$; obviously, we do not have unbiasedness in the small.

Let us see what Dawid's construction gives when applied to the defensive forecasting strategy constructed from the mixture of $S_n^{\epsilon,I}(p)$ and $S_n^{-\epsilon,I}(p)$, as described above, over different ϵ and different I; we will assume not only that the test functions I cover all $[0, 1]$ but also that each point $p \in [0, 1]$ is covered by arbitrarily narrow (concentrated in a small neighborhood of p) test functions. It is clear that we will inevitably have $p_n \to 0.5$ if p_n are produced by the defensive forecasting strategy and y_n are produced by Dawid's construction. On the other hand, since all test functions I are continuous and so cannot sharply distinguish between the cases $p_n < 0.5$ and $p_n \geq 0.5$, we do not have any contradiction: neither the test functions nor any observer who can only measure the p_n with a finite precision can detect the lack of unbiasedness in the small.

In this paper we are only interested in unbiasedness in the small when the test functions I are required to be continuous. Dawid's construction shows that unbiasedness in the small is impossible to achieve if I are allowed to be indicator functions of intervals (such as $[0, 0.5)$ and $[0.5, 1]$). To achieve unbiasedness in the small in this stronger sense, randomization appears necessary (see, e.g., [27]). It is interesting that already a little bit of randomization suffices, as explained in [8].

5 Simplified algorithm

Let us assume first that objects are absent, $|\mathbf{X}| = 1$. It was observed empirically that the performance of defensive forecasting strategies with a fixed ϵ does not depend on ϵ much (provided it is not too large; e.g., in the above calculations we assumed $\epsilon \leq 0.5$). This suggests letting $\epsilon \to 0$ (in particular, we will assume that $\epsilon \ll n^{-2}$). As the test functions I we will take Gaussian bells I_j with standard deviation $\sigma > 0$ located densely and uniformly in the interval $[0,1]$. Letting $\approx$ stand for approximate equality and using the shorthand $\sum_{\pm} f(\pm) := f(+) + f(-)$, we obtain:

$$
\begin{aligned}
S_n(p) &= \sum_{\pm} \sum_j (\pm\epsilon) I_j(p) \prod_{i=1}^{n-1} (1 \pm \epsilon I_j(p_i)(y_i - p_i)) \\
&= \sum_{\pm} \sum_j (\pm\epsilon) I_j(p) \exp\left(\sum_{i=1}^{n-1} \ln(1 \pm \epsilon I_j(p_i)(y_i - p_i)) \right) \\
&\approx \sum_{\pm} \sum_j (\pm\epsilon) I_j(p) \exp\left(\pm\epsilon \sum_{i=1}^{n-1} I_j(p_i)(y_i - p_i) \right) \\
&\approx \sum_{\pm} \sum_j (\pm\epsilon) I_j(p) \left(1 \pm \epsilon \sum_{i=1}^{n-1} I_j(p_i)(y_i - p_i) \right) \\
&= \sum_{\pm} \sum_j (\pm\epsilon) I_j(p) \left(\pm\epsilon \sum_{i=1}^{n-1} I_j(p_i)(y_i - p_i) \right) \\
&\propto \sum_j I_j(p) \sum_{i=1}^{n-1} I_j(p_i)(y_i - p_i) \\
&= \sum_{i=1}^{n-1} K(p, p_i)(y_i - p_i), \qquad (8)
\end{aligned}
$$

where $K(p, p_i)$ is the Mercer kernel

$$K(p, p_i) := \sum_j I_j(p) I_j(p_i).$$

This Mercer kernel can be approximated by

$$\int_0^1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-p)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-p_i)^2}{2\sigma^2}\right) dt$$

$$
\begin{aligned}
&\propto \int_0^1 \exp\left(-\frac{(t-p)^2 + (t-p_i)^2}{2\sigma^2}\right) dt \\
&\approx \int_{-\infty}^{\infty} \exp\left(-\frac{(t-p)^2 + (t-p_i)^2}{2\sigma^2}\right) dt.
\end{aligned}
$$

As a function of p, the last expression is proportional to the density of the sum of two Gaussian random variables of variance σ^2; therefore, it is proportional to

$$\exp\left(-\frac{(p-p_i)^2}{4\sigma^2}\right).$$

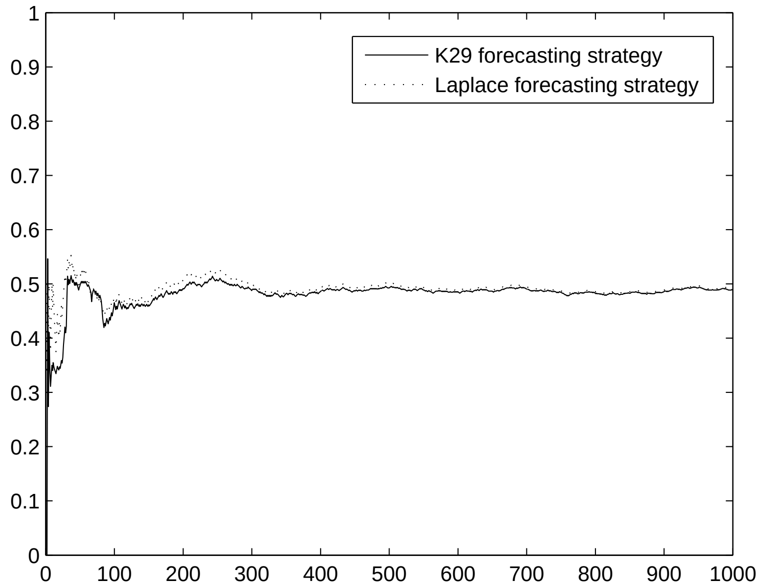


Figure 1: The First 1000 Probabilities Output by the K29 ($\sigma = 0.01$) and Laplace Forecasting Strategies on a Randomly Generated Bit Sequence

To get an idea of the properties of this forecasting strategy, which we call the *K29* strategy (or algorithm), we run it and the Laplace forecasting strategy ($p_n := (k+1)/(n+1)$, where k is the number of 1s observed so far) on a randomly generated bit sequence of length 1000 (with the probability of 1 equal to 0.5). A zero point p_n of S_n was found using the simple bisection procedure (see, e.g., [15], §§9.2–9.4, for more sophisticated methods): (a) start with the interval $[0,1]$; (b) let p be the mid-point of the current interval; (c) if $S_n(p) > 0$, remove the left half of the current interval; otherwise, remove its right half; (d) go to (b). We did 10 iterations, after which the mid-point of the remaining interval was output as p_n. Notice that the values $S_n(0)$ and $S_n(1)$ did not have to be tested. Our program was written in MATLAB, Version 7, and the initial state of the random number generator was set to 0.

Figure 1 shows that the probabilities output by the K29 ($\sigma = 0.01$) and Laplace forecasting strategies are almost indistinguishable. To see that these two forecasting strategies can behave very differently, we complemented the 1000 bits generated as described above with 1000 0s followed by 1000 1s. The result is shown in Figure 2. The K29 strategy detects that the probability p of 1 changes after the 1000th round, and fairly quickly moves down. When the probability changes again after the 2000th round, K29 starts moving toward $p = 1$, but interestingly, hesitates around the line $p = 0.5$, as if expecting the process to reverse to the original probability of 1.

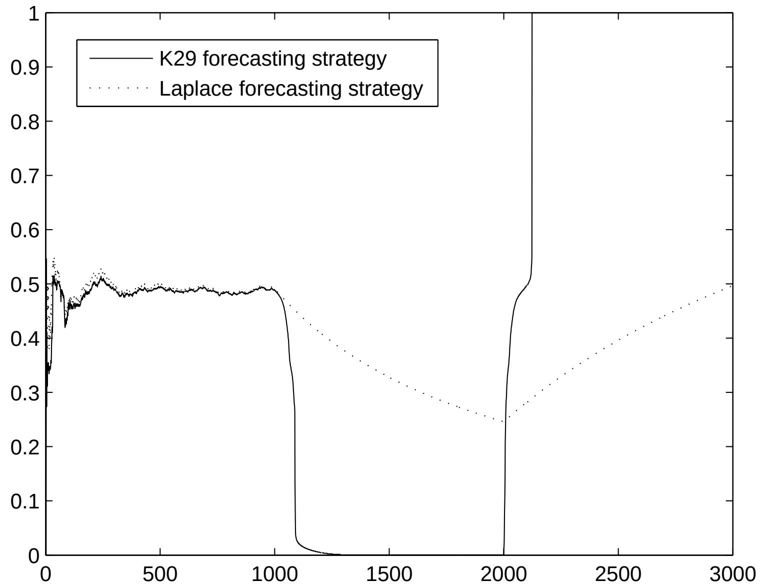


Figure 2: The Probabilities Output by the K29 ($\sigma = 0.01$) and Laplace Forecasting Strategies on a Randomly Generated Sequence of 1000 Bits Followed by 1000 0s and 1000 1s

The Mercer kernel

$$K(p, p_i) = \exp\left(-\frac{(p - p_i)^2}{4\sigma^2}\right)$$

used in these experiments is known in machine learning as the Gaussian kernel (in the usual parameterization $4\sigma^2$ is replaced by $2\sigma^2$ or c); however, many other Mercer kernels also give reasonable results.

If we start from test functions I depending on the object, instead of (8) we will arrive at the expression

$$S_n(p) = \sum_{i=1}^{n-1} K((p, x_n), (p_i, x_i))(y_i - p_i), \tag{9}$$

where K is a Mercer kernel on the squared product $([0, 1] \times \mathbf{X})^2$. There are standard ways of constructing such Mercer kernels from Mercer kernels on $[0, 1]^2$ and $\mathbf{X}^2$ (see, e.g., the description of tensor products and direct sums in [21, 17]). For S_n to be continuous, we have to require that K be *forecast-continuous* in the following sense: for all $x \in \mathbf{X}$ and all $(p', x') \in [0, 1] \times \mathbf{X}$, $K((p, x), (p', x'))$ is continuous as a function of p. The overall procedure can be summarized as follows.

K29 Algorithm

Parameter: forecast-continuous Mercer kernel K on $([0,1] \times \mathbf{X})^2$
FOR $n = 1, 2, \ldots$:
  Read $x_n \in \mathbf{X}$.
  Define $S_n(p)$ as per (9).
  Output any root p of $S_n(p) = 0$ as p_n;
  if there are no roots, $p_n := (1 + \text{sign}(S_n))/2$.
  Read $y_n \in \{0, 1\}$.

Computer experiments reported in [25] show that the K29 algorithm performs well on a standard benchmark data set. For a theoretical discussion of the K29 algorithm, see the appendix and [26].

6 Related work and directions of further research

This paper's methods connect two areas that have been developing independently so far: probability forecasting and classical probability theory. It appears that, when properly developed, these methods can benefit both areas:

- the powerful machinery of classical probability theory can be used for probability forecasting;
- practical problems of probability forecasting may suggest new laws of probability.

Classical probability theory started from Bernoulli's weak law of large numbers (1713) and is the subject of countless monographs and textbooks. The original statements of most of its results were for independent random variables, but they were later extended to the martingale framework; the latter was reduced to its game-theoretic core in [18]. The proof of the strong law of large numbers used in this paper was extracted from Ville's [22] martingale proof of the law of the iterated logarithm (upper half).

The theory of probability forecasting was a topic of intensive research in meteorology in the 1960s and 1970s; this research is summarized in [5]. Machine learning is still mainly concerned with categorical prediction, but the situation appears to be changing. Probability forecasting using Bayesian networks is a mature field; the literature devoted to probability forecasting using decision trees and to calibrating other algorithms is also fairly rich. So far, however, the field of probability forecasting has been developing without any explicit connections with classical probability theory.

Defensive forecasting is indirectly related, in a sense dual, to prediction with expert advice (reviewed in [24], §4) and its special case, Bayesian prediction. In prediction with expert advice one starts with a given loss function and tries to make predictions that lead to a small loss as measured by that loss function. In defensive forecasting, one starts with a law of probability and then makes predictions such that this law of probability is satisfied. So the choice of the law of probability when designing the forecasting strategy plays a role analogous to the choice of the loss function in prediction with expert advice.

In prediction with expert advice one combines a pool of potentially promising forecasting strategies to obtain a forecasting strategy that performs not much worse than the best strategies in the pool. In defensive forecasting one combines strategies for Skeptic (such as the strategies corresponding to different test functions I and different $\pm\epsilon$ in §4) to obtain one strategy achieving an interesting goal (such as unbiasedness in the small); a strategy for Forecaster is then obtained using Theorem 1. The possibility of mixing strategies for Skeptic is as fundamental in defensive forecasting as the possibility of mixing strategies for Forecaster in prediction with expert advice.

This paper continues the work started by Foster and Vohra [7] and later developed in, e.g., [10, 16, 27] (the last paper replaces the von Mises–style framework of the previous papers with a martingale framework, as in this paper). The approach of this paper is similar to that of the recent paper [8], which also considers deterministic forecasting strategies and continuous test functions for unbiasedness in the small.

The main difference of this paper's approach from the bulk of work in learning theory is that we do not make any assumptions about Reality's strategy.

The following directions of further research appear to us most important:

- extending Theorem 1 to other forecasting protocols (such as multi-label classification) and designing efficient algorithms for finding the corresponding p_n;
- exploring forecasting strategies corresponding to: (a) Hoeffding's inequality, (b) the central limit theorem, (c) the law of the iterated logarithm (all we did in this paper was to slightly extend the strong law of large numbers and then use it for probability forecasting).

Acknowledgments

We are grateful to the participants of the PASCAL workshop "Notions of complexity: information-theoretic, computational and statistical approaches" (October 2004, EURANDOM) who commented on this work and to the anonymous referees for useful suggestions. This work was partially supported by BBSRC (grant 111/BIO14428), EPSRC (grant GR/R46670/01), MRC (grant S505/65), Royal Society, and, especially, the Superrobust Computation Project (Graduate School of Information Science and Technology, University of Tokyo).

References

[1] Lenore Blum, Michael Shub, and Steve Smale. On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the American Mathematical Society*, 21:1–46, 1989.

[2] Émile Borel. Le calcul des intégrales définies. *Journal de Mathématiques pures et appliquées, Séries 6*, 8:159–210, 1912.

[3] L. E. J. Brouwer. Begründung der Mengenlehre unabhängig vom logischen Satz vom ausgeschlossenen Dritte. Erster Teil. Allgemeine Mengelehre. *Koninklijke Nederlandse Akademie van Wetenschschappen Verhandelingen*, 5:1–43, 1918.

[4] A. Philip Dawid. Self-calibrating priors do not exist: Comment. *Journal of the American Statistical Association*, 80:340–341, 1985. This is a contribution to the discussion in [13].

[5] 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.

[6] Joseph L. Doob. *Stochastic Processes.* Wiley, New York, 1953.

[7] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

[8] Sham M. Kakade and Dean P. Foster. Deterministic calibration and Nash equilibrium. In John Shawe-Taylor and Yoram Singer, editors, *Proceedings of the Seventeenth Annual Conference on Learning Theory*, volume 3120 of *Lecture Notes in Computer Science*, pages 33–48, Heidelberg, 2004. Springer.

[9] Andrei N. Kolmogorov. Sur la loi des grands nombres. *Atti della Reale Accademia Nazionale dei Lincei. Classe di scienze fisiche, matematiche, e naturali. Rendiconti Serie VI*, 185:917–919, 1929.

[10] Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 69:1333–1347, 2001.

[11] Per Martin-Löf. *Notes on Constructive Mathematics.* Almqvist & Wiksell, Stockholm, 1970.

[12] Allan H. Murphy and Edward S. Epstein. Verification of probabilistic predictions: a brief review. *Journal of Applied Meteorology*, 6:748–755, 1967.

[13] David Oakes. Self-calibrating priors do not exist (with discussion). *Journal of the American Statistical Association*, 80:339–342, 1985.

[14] K. R. Parthasarathy and Klaus Schmidt. *Positive definite kernels, continuous tensor products, and central limit theorems of probability theory*, volume 272 of *Lecture Notes in Mathematics.* Springer, Berlin, 1972.

[15] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C.* Cambridge University Press, Cambridge, second edition, 1992.

[16] Alvaro Sandroni, Rann Smorodinsky, and Rakesh V. Vohra. Calibration with many checking rules. *Mathematics of Operations Research*, 28:141–153, 2003.

[17] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels.* MIT Press, Cambridge, MA, 2002.

[18] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[19] Mark van Atten and Dirk van Dalen. Arguments for the continuity principle. *Bulletin of Symbolic Logic*, 8:329–347, 2002.

[20] Dirk van Dalen. *Mystic, Geometer, and Intuitionist. The Life of L. E. J. Brower*, volume 1. The Dawning Revolution. Clarendon Press, Oxford, 1999.

[21] Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[22] Jean Ville. *Etude critique de la notion de collectif.* Gauthier-Villars, Paris, 1939.

[23] Jan von Plato. Review of [20]. *Bulletin of Symbolic Logic*, 7:62–65, 2001.

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

[25] Vladimir Vovk. Defensive forecasting for a benchmark data set, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #9, September 2004.

[26] Vladimir Vovk. Non-asymptotic calibration and resolution, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #11, November 2004.

[27] Vladimir Vovk and Glenn Shafer. Good randomized sequential probability forecasting is always possible, The Game-Theoretic Probability and Finance project, `http://probabilityandfinance.com`, Working Paper #7, June 2003 (revised September 2004).

Appendix: Geometry of the K29 algorithm and weak probability theory

In this paper the K29 algorithm was derived, somewhat informally, from the requirement that the forecasts should be unbiased in the small. However, since in this derivation we assumed that ϵ was very small, we cannot longer assert that the algorithm obtained (K29) is unbiased in the small. In this appendix we will give a direct argument showing that the K29 algorithm can be expected to be unbiased in the small. The figures in this appendix will be given in color (color is especially important for Figure 8).

What K29 accomplishes

In this subsection we reproduce a result from [26] (Theorem 3 below coincides with Theorem 1 of [26]) and complement it by related results showing connections with classical probability theory.

Following the K29 algorithm Forecaster ensures that Skeptic will never increase his capital with the strategy

$$s_n := \sum_{i=1}^{n-1} K\left((p_n, x_n), (p_i, x_i)\right)(y_i - p_i). \tag{10}$$

(This strategy is not necessarily valid in the sense of guaranteeing Skeptic's solvency; we will take care of the latter later on.) The increase in Skeptic's capital when he follows (10) is

$$\begin{aligned}\mathcal{K}_N - \mathcal{K}_0 &= \sum_{n=1}^{N} s_n(y_n - p_n)\\ &= \sum_{n=1}^{N}\sum_{i=1}^{n-1} K\left((p_n, x_n), (p_i, x_i)\right)(y_n - p_n)(y_i - p_i)\\ &= \frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{N} K\left((p_n, x_n), (p_i, x_i)\right)(y_n - p_n)(y_i - p_i)\\ &\qquad - \frac{1}{2}\sum_{n=1}^{N} K\left((p_n, x_n), (p_n, x_n)\right)(y_n - p_n)^2 \end{aligned}\tag{11}$$

(we generalize slightly the protocol of §2 allowing initial values $\mathcal{K}_0$ of Skeptic's capital different from 1). According to Mercer's theorem (a very simple proof of a suitable version can be found in [6], Theorem II.3.1), there exists a function $\Phi : [0,1] \times \mathbf{X} \to H$ (a *feature mapping* taking values in a Hilbert space H) such that

$$K(a, b) = \Phi(a) \cdot \Phi(b), \ \ \forall a, b \in [0,1] \times \mathbf{X} \tag{12}$$

($\cdot$ standing for the dot product in H). Therefore, we can rewrite (11) as

$$\mathcal{K}_N - \mathcal{K}_0 = \frac{1}{2}\left\|\sum_{n=1}^{N}(y_n - p_n)\Phi(p_n, x_n)\right\|^2 - \frac{1}{2}\sum_{n=1}^{N}\left\|(y_n - p_n)\Phi(p_n, x_n)\right\|^2. \tag{13}$$

To make sure that Skeptic never goes bankrupt, let us consider a finite-horizon game with N the horizon (i.e., "FOR $n = 1, 2, \ldots$" in the protocol of §2 is replaced by "FOR $n = 1, \ldots, N$"), assume that

$$C := \sup_{p,x} \|\Phi(p, x)\| < \infty \tag{14}$$

(it is often a good idea to use Mercer kernels with $C = 1$), and set

$$\mathcal{K}_0 := \frac{1}{2} N C^2$$

(the latter ensuring $\mathcal{K}_n \ge 0$, $\forall n$). With game-theoretic lower probability at least $1-\delta$ we will have $\mathcal{K}_N < \frac{1}{\delta}\mathcal{K}_0$, which, in combination with (13), implies

$$\frac{1}{2}\left\|\sum_{n=1}^{N}(y_n - p_n)\Phi(p_n, x_n)\right\|^2 \le \mathcal{K}_N < \frac{1}{\delta}\mathcal{K}_0 = \frac{1}{2\delta}NC^2,$$

i.e.,

$$\left\|\sum_{n=1}^{N}(y_n - p_n)\Phi(p_n, x_n)\right\|^2 < \frac{NC^2}{\delta},$$

$$\left\|\frac{1}{N}\sum_{n=1}^{N}(y_n - p_n)\Phi(p_n, x_n)\right\| < \frac{C}{\sqrt{N\delta}}.$$

Therefore, we have proved:

Theorem 2 *Let $N \in \{1, 2, \dots\}$, $\delta > 0$, $\Phi : [0,1] \times \mathbf{X} \to H$ for a Hilbert space H, and C be defined by (14). Then*

$$\underline{\mathbb{P}}\left\{\left\|\frac{1}{N}\sum_{n=1}^{N}(y_n - p_n)\Phi(p_n, x_n)\right\| < \frac{C}{\sqrt{N\delta}}\right\} \ge 1-\delta \tag{15}$$

in Binary Forecasting Game I with horizon N.

It is clear that the K29 algorithm ensures the inequality within the curly braces in (15) for any $\delta > 0$; this is stated in the following theorem, in which we also remove the assumption of a finite horizon.

Theorem 3 *The K29 algorithm with parameter K ensures*

$$\sup_{N \in \{1,2,\dots\}}\left\|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(y_n - p_n)\Phi(p_n, x_n)\right\| \le C \tag{16}$$

in Binary Forecasting Game I, where Φ is a mapping taking values in a Hilbert space and satisfying (12).

In this theorem we can still observe the phenomenon we saw earlier (cf. (6)): the forecasts are calibrated better than in the case of genuine randomness. Let us take, for simplicity, $\Phi \equiv 1$. According to the law of the iterated logarithm, we would expect

$$\limsup_{N\to\infty}\left|\frac{1}{\sqrt{2A_N \ln\ln A_N}}\sum_{n=1}^{N}(y_n - p_n)\right| = 1,$$

where

$$A_N := \sum_{n=1}^{N} p_n(1-p_n),$$

and so

$$\sup_{N\in\{1,2,\dots\}} \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n, x_n) \right\|$$

to be infinite for p_n not consistently very close to 0 or 1.

Remark The parameter K of the K29 algorithm is required to be forecast-continuous. If K satisfies (12) with $\Phi(p,x)$ continuous in p (for any x), K is forecast-continuous; moreover, in this case $K((p,x),(p',x'))$ is continuous in (p,p') (for any (x,x')). On the other hand, the last property implies that the mapping Φ in the representation (12) can be chosen continuous ([14]; [17], Proposition 2.14 on p. 41).

Connection with §§4.2–4.3

Let (p^*,x^*) be a point in $[0,1]\times\mathbf{X}$; we would like the average of y_n, $n=1,\dots,N$, such that (p_n,x_n) is close to (p^*,x^*) to be close to p^*. (Cf. (7) and the discussion in §4.3.) Fix a forecast-continuous Mercer kernel $K : ([0,1]\times\mathbf{X})^2 \to \mathbb{R}$ and consider the "soft neighborhood"

$$I_{(p^*,x^*)}(p,x) := K((p^*,x^*),(p,x)) \tag{17}$$

of the point (p^*,x^*). The following is an easy corollary of Theorem 3 (we refrain from stating the analogous corollary for Theorem 2).

Corollary 1 *In Binary Forecasting Game I with horizon N, the K29 algorithm with parameter $K \geq 0$ ensures*

$$\left| \frac{\sum_{n=1}^{N}(y_n - p_n) I_{(p^*,x^*)}(p_n,x_n)}{\sum_{n=1}^{N} I_{(p^*,x^*)}(p_n,x_n)} \right| \leq \frac{C^2\sqrt{N}}{\sum_{n=1}^{N} I_{(p^*,x^*)}(p_n,x_n)} \tag{18}$$

for each point $(p^*,x^*) \in ([0,1]\times\mathbf{X})$, where I is defined by (17).

This corollary implies that we can expect unbiasedness in the "soft neighborhood" of (p^*,x^*) when

$$\sum_{n=1}^{N} I_{(p^*,x^*)}(p_n,x_n) \gg \sqrt{N}.$$

Proof of Corollary 1 Let $\Phi : [0,1]\times\mathbf{X} \to H$ be a function taking values in a Hilbert space H and satisfying (12). Theorem 3 and the Cauchy–Schwarz inequality imply

$$\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(y_n-p_n)I_{(p^*,x^*)}(p_n,x_n)\right|$$
$$=\left|\left(\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(y_n-p_n)\Phi(p_n,x_n)\right)\cdot\Phi(p^*,x^*)\right|$$
$$\leq\left\|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(y_n-p_n)\Phi(p_n,x_n)\right\|\|\Phi(p^*,x^*)\|\leq C^2;$$

the inequality between the extreme terms of this chain is equivalent to (18). ▮

Connection with the weak law of large numbers

Despite the fact that K29 was derived from a proof of the strong law of large numbers, it is closely connected with the weak law: indeed, if $\Phi(p,x)\equiv 1$, (15) is a form of Bernoulli's theorem. To see this, rewrite (15) as

$$\overline{\mathbb{P}}\left\{\left\|\frac{1}{N}\sum_{n=1}^{N}(y_n-p_n)\Phi(p_n,x_n)\right\|\geq\epsilon\right\}\leq\frac{C^2}{N\epsilon^2},$$

substitute 1 for Φ and 1 for C (see (14)), and compare

$$\overline{\mathbb{P}}\left\{\left|\frac{1}{N}\sum_{n=1}^{N}(y_n-p_n)\right|\geq\epsilon\right\}\leq\frac{1}{N\epsilon^2}\tag{19}$$

with the first displayed equation on p. 126 of [18].

Remark The measure-theoretic counterpart of (19) was first proven by Kolmogorov in 1929; this is the origin of the current (provisional) name of the K29 algorithm. The combination of equations (3) and (7) in [9] gives, essentially,

$$\mathbb{P}\left\{\left|\sum_{n=1}^{N}Z_n\right|\geq\epsilon\right\}\leq\frac{1}{\epsilon^2}\sum_{n=1}^{N}\mathbb{E}\left(Z_n^2\mid\mathcal{F}_{n-1}\right),\tag{20}$$

where $Z_1,\ldots,Z_n$ is a martingale difference w.r. to the filtration $(\mathcal{F}_i)_{i=1}^n$ and $\mathcal{F}_0$ is the trivial σ-algebra; (20) is more general than the measure-theoretic counterpart of (19). (Of course, Kolmogorov did not use this terminology; the modern notion of martingale was introduced by Ville [22].)

Examples

We can see that, for $\Phi\equiv 1$, (15) is a formalization of unbiasedness in the large. It is intuitively clear that (15) may express unbiasedness in the small with good resolution if Φ is a sufficiently twisted surface (assuming, for simplicity, that **X**

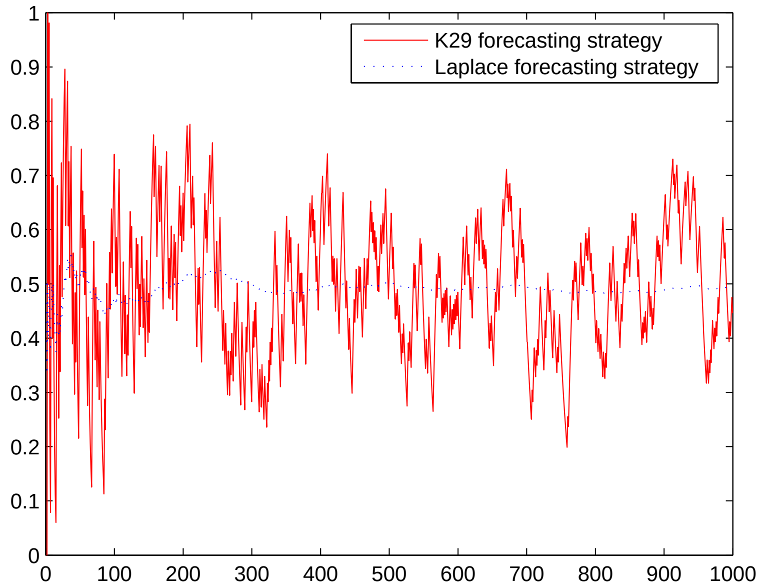


Figure 3: Analogue of Figure 1 for the Mercer kernel $K(p, p') := \cos(p - p')$

is a continuous space) in H: if pairwise distant points (p_i, x_i), $i = 1, \ldots, m$, in $[0, 1] \times \mathbf{X}$ are mapped by Φ to vectors $\Phi(p_i, x_i)$ that are far from being dependent (we will call this the *diversity* property of Φ), then unbiasedness is required to hold in the neighborhood of each point, not just in the large.

To consider a simple example, let us now assume that objects x_n are absent ($|\mathbf{X}| = 1$); $\Phi = \Phi(p)$ is then a path in the Hilbert space H. The path $\Phi(p) := e^{ip}$ in the complex plane satisfies, to some (rather weak) degree, the diversity property: $\Phi(p)$ and $\Phi(p')$ are not collinear for distant p and p'; the corresponding Mercer kernel is $K(p, p') = \cos(p - p')$. Figures 3 and 4 are the analogues of Figures 1 and 2 for this Mercer kernel. A possible explanation for the rugged shape of the solid line in Figure 3 is that Φ is not diverse enough: the two-dimensional complex plane simply does not have enough room for much diversity. If we take $\Phi(p) := e^{cip}$ for $c > 1$ in an attempt to increase diversity for moderately distant points, we will risk very distant points becoming collinear or nearly collinear; even the Mercer kernel based on $\Phi(p) := e^{\pi ip}$ occasionally confuses 0 and 1, which are mapped to collinear vectors (see Figures 5 and 6). The forecasts become very bad for $\Phi(p) := e^{2\pi ip}$ (Figure 7), although even in this case the performance can be surprisingly good for categorical (0 or 1) forecasts (see Figure 6; there is only one error after round 2000).

Much greater diversity is provided by the Gaussian kernel

$$K(p, p') := \exp\left(-\frac{(p - p')^2}{c}\right),$$

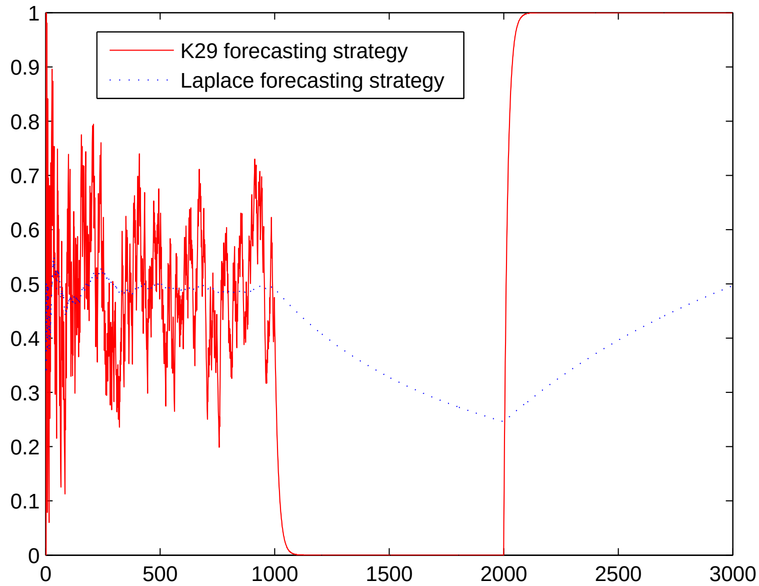


Figure 4: Analogue of Figure 2 for the Mercer kernel $K(p,p') := \cos(p-p')$

where c is a small positive number: $\Phi(p)$ and $\Phi(p')$ are nearly orthogonal for distant p and p'. It is easy to check that this kernel corresponds to the Fourier feature mapping

$$\Phi(p) : \mathbb{R} \to \mathbb{C}$$
$$\lambda \mapsto e^{i\lambda p}$$

with the following dot product in the feature space:

$$f \cdot g = \frac{1}{\sqrt{2\pi}} \int f(\lambda)\overline{g}(\lambda) e^{-c\lambda^2/4} d\lambda.$$

Weak probability theory

The research program proposed in this paper consists in using Theorem 1 to transform laws of probability into forecasting strategies. However, a closer look at Theorem 1 reveals that we need much less than a law of probability to derive a forecasting strategy. In this subsection we introduce a suitable relaxation of the game-theoretic probability theory as developed in [18]; we will assume that the reader is familiar with, or has access to, the main definitions of [18] (we will, however, use the words "weak" and "weakly" in a different sense from [18]).

Let us say that Skeptic can *weakly force* an event E if he has a strategy in Binary Forecasting Game I that guarantees the disjunction

$$(\exists n : \mathcal{K}_n < \mathcal{K}_{n+1}) \text{ or } E.$$

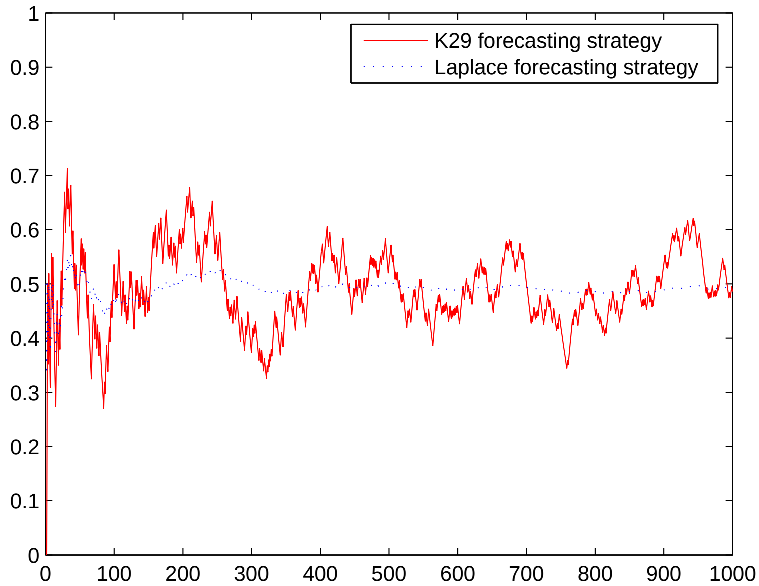


Figure 5: Analogue of Figure 1 for the Mercer kernel $K(p, p') := \cos(\pi(p - p'))$

Theorem 1 shows that if Skeptic can weakly force an event E with a continuous strategy, Forecaster has s strategy in Basic Binary Forecasting Protocol that guarantees E.

Weak probability theory corresponding to the notion of weakly forcing is radically different from the standard probability theory. To see this, remember that the usual results of probability theory (the strong law of large numbers, the law of the logarithm, etc.; cf. [18]) remain interesting even when we fix Forecaster's strategy; in fact, most of these results were first discovered for the fair-coin protocol ([18], §3.1). It is easy to see that an event E can be weakly forced when Forecaster follows a fixed strategy producing forecasts in $(0, 1)$ if and only if E is non-empty. It is clear that the situation will not change if we require Skeptic's strategy to be continuous.

All three of the following closely related properties of an event E appear to be interesting:

- E can be weakly forced;
- E can be weakly forced with a continuous strategy;
- Forecaster can guarantee E in Basic Binary Forecasting Protocol.

We believe that these properties and relations between them deserve serious study.

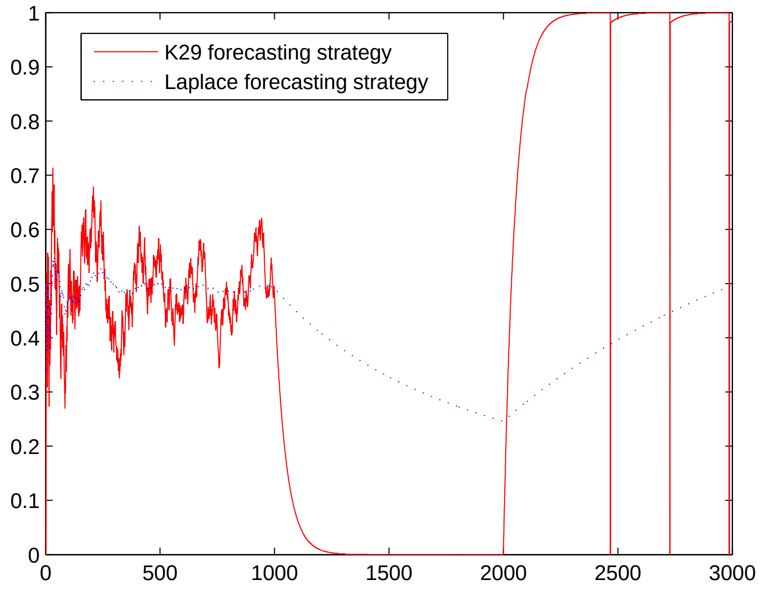


Figure 6: Analogue of Figure 2 for the Mercer kernel $K(p, p') := \cos(\pi(p - p'))$

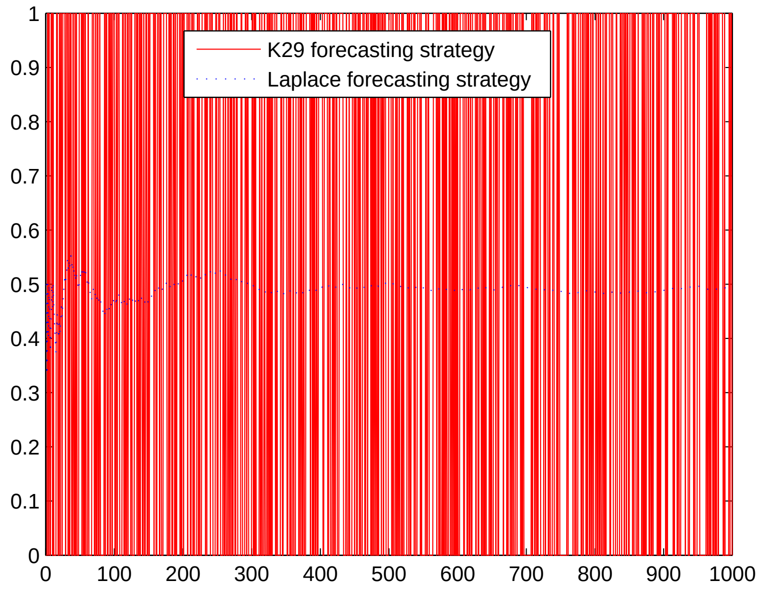


Figure 7: Analogue of Figure 1 for the Mercer kernel $K(p, p') := \cos(2\pi(p - p'))$

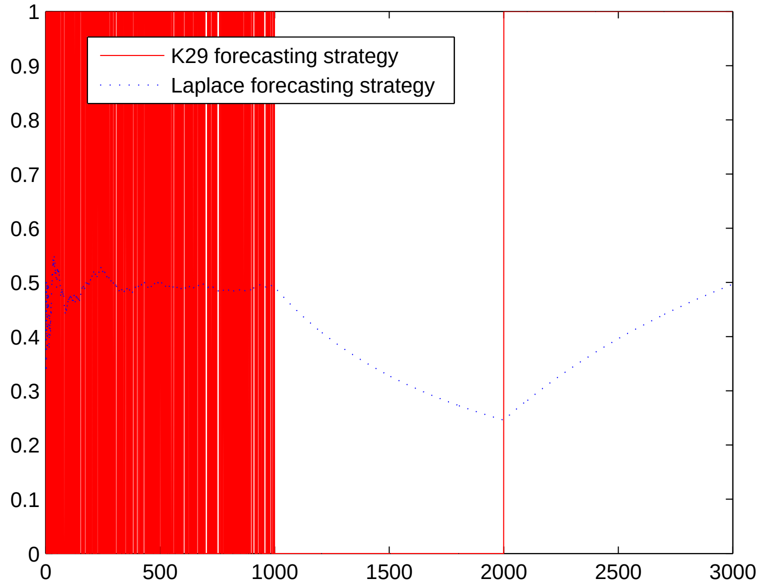


Figure 8: Analogue of Figure 2 for the Mercer kernel $K(p, p') := \cos(2\pi(p - p'))$