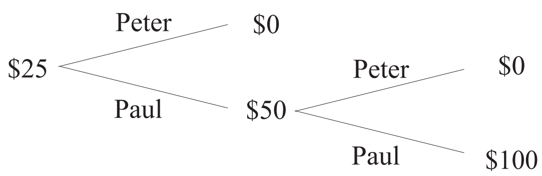
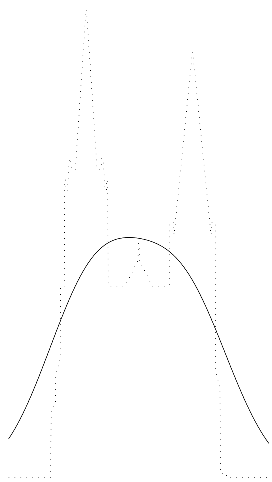
Game-theoretic probability and its uses, especially defensive forecasting

Glenn Shafer[1]





The Game-Theoretic Probability and Finance Project

Working Paper #22

August 24, 2007

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

[1]The author thanks Volodya Vovk for his advice on the exposition as well as for his central role in the work reviewed. He also thanks Hans-Joachim Lenz and Giacomo Della Riccia for their invitation to present this work at the 2006 ISSEK Invitational Workshop on Preferences and Similarities in Udine, Italy.

Abstract

This expository article reviews the game-theoretic framework for probability and the method of defensive forecasting that derives from it.

The game-theoretic framework, introduced by Vladimir Vovk and myself in *Probability and Finance: It's Only a Game!* (Wiley 2001 [51]), can replace measure theory as a mathematical framework for classical probability theory, discrete and continuous. Classical theorems are proven by betting strategies that multiply a player's stake by a large factor if the theorem's prediction fails. These strategies can be specified explicitly, giving probability a constructive character appropriate for applications.

Defensive forecasting, introduced by Vovk, Takemura, and myself in 2005 [77], is one of the most interesting applications. It identifies a betting strategy that succeeds if probabilistic forecasts are inaccurate, and it makes probabilistic forecasts that will defeat this betting strategy. The fact that this is possible provides new insight into the meaning of probability.

Contents

1 Introduction

Game-theoretic probability begins with the idea that probabilities are tested by betting: a forecaster's probabilities are refuted when an opponent betting at the odds defined by the probabilities multiplies his stake by a large factor. Defensive forecasting is a method for giving probability forecasts that cannot be refuted in this way.

1.1 Dawid's counterexample

At first glance, it seems implausible that we can give probability forecasts that cannot possibly be defeated by a betting opponent. The intuition that this is impossible can be formalized using the following betting protocol, in which Reality successively decides the outcomes of a sequence of events. Just before Reality announces the outcome of the nth event ($y_n = 1$ if the event happens, $y_n = 0$ if it fails), Forecaster gives a probability p_n for its happening, and his betting opponent, Skeptic, sets the stakes s_n for a bet for or against its happening. We assume that this is a perfect-information protocol: each player sees the other player's moves as they are made.

PROTOCOL 1. PROBABILITY FORECASTING

FOR $n = 1, 2, \dots$:

Forecaster announces $p_n \in [0, 1]$.

Skeptic announces $s_n \in \mathbb{R}$.

Reality announces $y_n \in \{0, 1\}$.

Skeptic's profit $:= s_n(y_n - p_n)$.

If $s_n > 0$, Skeptic is betting on the event; he gets $s_n(1 - p_n)$ if it happens and loses $s_n p_n$ if it fails. If $s_n < 0$, he is betting against the event; he gets $-s_n p_n$ if it fails and loses $-s_n(1 - p_n)$ if it happens.

Working together, Skeptic and Reality can refute Forecaster spectacularly. When Forecaster gives a low probability, Skeptic bets on the event and Reality makes it happen; when Forecaster gives a high probability, Skeptic bets against the event and Reality makes it fail. This idea was formalized in 1985 by A. P. Dawid, whose studies of probability forecasting were seminal for game-theoretic probability [17], in a discussion of an article by David Oakes [16, 41]. As Dawid explained, if Reality follows the strategy

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

and Skeptic follows the strategy

$$s_n := \begin{cases} 1 & \text{if } p_n < 0.5 \\ -1 & \text{if } p_n \geq 0.5, \end{cases}$$

then Skeptic makes a profit of at least 0.5 on every round: $1-p_n$ when $p_n < 0.5$ and p_n when $p_n \geq 0.5$.

Surprisingly, Dawid's counterexample to the possibility of good probability forecasting is not as watertight as it first appears. The weak point of Skeptic's and Reality's strategy is that it requires them to be able to tell whether p_n is exactly equal to 0.5 and to switch their moves with infinite abruptness (discontinuously) as p_n shifts from 0.5 to something ever so slightly less. Because the idea of an infinitely abrupt shift lives in a idealized mathematical world, we should be wary about drawing practical conclusions from it.

To show that Dawid's counterexample depends on the artificiality of the mathematical idealization, it suffices to show that it disappears when we hobble Skeptic and Reality in small ways that are reasonable or inconsequential in practice. This has been done in two different ways:

1. We can hobble Skeptic a bit by requiring that he follow a strategy that makes his bet a continuous function of the forecast p_n. This is not a practical impediment, because a continuous function can change with any degree of abruptness short of infinite. It can easily be shown—I reproduce in §4.2 below the proof given in [77]—that Forecaster can beat any continuous strategy. And as I explain in §4.3, there are continuous strategies for Skeptic that make a lot money unless Forecaster's probabilities perform well. In particular, Skeptic has continuous strategies that make a lot of money unless Forecaster is well calibrated, in the sense that about 30% of the events for which p_n is near 0.3 will happen, about 40% of the events for which p_n is near 0.4 will happen, etc. By playing against and beating a continuous strategy for Skeptic that makes money if Forecaster does not perform well, Forecaster is sure to perform well no matter what Reality does. This may seem surprising, but Forecaster's play turns out to be quite natural. To guarantee calibration, for example, he always chooses a value of p_n where he has been fairly well calibrated so far, so that Reality cannot make his calibration much worse, no matter whether she puts y_n equal to 1 or 0; see §4.3.3.

2. Alternatively, we can hobble Skeptic and Reality by allowing Forecaster to make the exact value of p_n slightly indeterminate. Forecaster announces a probability distribution for p_n rather than the exact value of p_n, leaving the exact value to be determined by a random drawing from the probability distribution after Skeptic and Reality make their moves. In this case, Forecaster can beat any strategy for Skeptic, even it if is not continuous. Vovk and I initially gave a direct proof of this [76], building on earlier work by Foster and Vohra and others [22], but the result actually follows easily from the result concerning continuous strategies for Skeptic [66].

The idea of getting good forecasts by playing against strategies for Skeptic is called *defensive forecasting*. It turns out to be very productive. It works not only when we give probabilities for a sequence of events but also when we merely give point forecasts for a sequence of quantities [72, 74].

1.2 The significance of defensive forecasting

In order to understand the practical and philosophical significance of defensive forecasting, we need to understand the following claims:

- The empirical meaning of a theory that gives probabilities for a sequence of events lies in the predictions it makes very confidently. In other words, such a theory's meaning is captured by statistical tests, which reject the theory when something to which it gives very small probability happens.
- Statistical testing can alternatively be understood in betting terms, because an event with very small probability happening is equivalent to a betting strategy multiplying the capital it risks by a large factor.
- This game-theoretic notion of testing generalizes from the case where full probability distributions are given to the case where only some quantities are predicted.
- In many cases, it is possible to a specify a single betting strategy that is quasi-universal, in the sense that it makes a lot of money (multiplies the capital risked by a large factor) whenever our probabilistic predictions fail in ways that are important to us.

Once these claims are accepted, the possibility of defensive forecasting casts a new light on the objective meaning of probability. It is often supposed that the success of probability predictions proves the existence of some sort of chance mechanism underlying the phenomenon being predicted. Something like the throwing of dice must be going on. But now we see that a sequence of probability predictions can be successful regardless of how events are actually determined.

It should be emphasized that this result holds only in the perfect-information sequential protocol that I have been discussing, where Forecaster sees all the preceding outcomes $y_1, \ldots, y_{n-1}$ before gives a probability p_n for y_n. A probabilistic theory that gives successful probabilities much farther in advance (quantum mechanics, for example) can claim more empirical content.

The remainder of this article is divided into four sections. In §2, I draw on the historical record to make a case for my claims concerning the empirical meaning of probabilistic theories—the claim that empirical meaning is captured by statistical testing and the claim that testing can be understood in terms of betting. In §3, I review the game-theoretic framework for probability and some of its accomplishments other than defensive forecasting. In §4, I review the basic theory of defensive forecasting and discuss its implications further.

2 Historical context

It is a commonplace that a probabilistic or statistical theory can be tested by checking whether an event to which it assigns very small probability happens. But during the past 50 years, few scholars have seen this as fundamental to the meaning of objective probability. Instead, proponents of the objective interpretation of probability have continued to emphasize the relation between objective probability and frequency, even while the difficulties involved in identifying probability with frequency have led others to abandon the objective interpretation in favor of a subjective one.

I begin this section, in §2.1, by reviewing difficulties with the frequency interpretation that emerged in the twentieth century. Then in §2.2, I review the understanding of objective probability based on testing, which dates back to the nineteenth century and was dominant in France during much of the twentieth century. Here I speak of *Cournot's principle*, which says that an event of very small probability singled out in advance will not happen. Finally, in §2.3, I explain Ville's theorem, which relates small probability to the success of a betting strategy and therefore allows us to recast Cournot's principle in betting terms.

2.1 The failure of frequentism

Most debate about the meaning of probability is carried on within a consensus that frequencies can be used to estimate probabilities. This consensus relies on Bernoulli's theorem, which says that when successive observations are independent, the frequency with which an event happens almost certainly approximates its probability. The classical subjectivist and objectivist positions in the debate were already familiar in Germany and Britain in the nineteenth century. According to the subjectivist position, probabilities are primarily degrees of belief. According to the objectivist position, they are merely frequencies, whose reality is independent of our knowledge. (I leave aside the secondary debate about the scope of application of Bayes's theorem, which introduces purely subjective probabilities.)

This classical picture began to fray in the middle of the twentieth century, when statisticians began to emphasize stochastic processes more than independent observations. As Jerzy Neyman explained in 1960 [40], the history of indeterminism in science had entered a new period of "dynamic indeterminism,"

> characterized by the search for evolutionary chance mechanisms capable of explaining the various frequencies observed in the development of the phenomena studied. The chance mechanism of carcinogenesis and the chance mechanism behind the varying properties of the comets in the Solar System exemplify the subjects of dynamic indeterministic studies. One might hazard the assertion that every serious contemporary study is a study of the chance mechanism behind some phenomenon. The statistical and probabilistic tool in such studies is the theory of stochastic processes, now involving many unsolved problems. In order that the applied statistician be in a position to cooperate effectively with the modern experimental scientist, the theoretical equipment of the statistician must include familiarity and capability of dealing with stochastic processes.

The shift from independent observations to stochastic processes destabilized the identification of probabilities with frequencies. We can observe only one Solar System: the stochastic process plays itself out only once. There are recurring events in the Solar System, and so there are, as Neyman says, various frequencies to explain. But the probabilities required to define so complex a stochastic process go far beyond those that can be estimated by frequencies, and the frequencies explained may go beyond those that approximate single probabilities.

Examples of frequencies that are not explained by single probabilities go back to the law of large numbers formulated by Poisson in 1837 [42]. If events $E_1, \ldots, E_N$ have probabilities $p_1, \ldots, p_N$, and if the relative frequency of the E_n (the fraction that happen) approximates the average of the p_n, then the p_n explain the frequency, even though the frequency does not approximate a single probability. In the framework of stochastic processes, each p_n is a probability conditional on earlier events. Poisson's result was treated in this spirit by Kolmogorov in 1929 [27]. I will discuss its game-theoretic formulation in §3.3 and §3.4.

Neyman dealt with the failure of classical frequentism by falling back on the thought that phenomena have "chance mechanisms" behind them. Perhaps God does play dice with the universe. With this rhetorical shift, Neyman clung to a sense of objectivity, but he lost frequentism's strong empiricism. A chance mechanism is supposed to be an objective feature of the world, but we cannot thoroughly observe and test it. We cannot refute the mere hypothesis that a phenomenon is governed by some completely unknown stochastic process. This elusiveness contributed to the resurgence of subjectivism in the late twentieth century.

In more recent years, our much increased capacity to acquire, store, and analyze data has reversed somewhat the shift from the model of independent observations to stochastic processes. We now see many large datasets in which the number of variables measured greatly exceeds the number of observations [73]. We may still be interested in predicting a particular variable Y from other variables $X_1, \ldots, X_k$, as Neyman was ([40], p. 626), but when k is so extremely large, it often seems neither necessary nor practical to model the evolution of the system over time. Instead, we may simply assume that changes in the system are adequately summarized by current values of the X_i. But this does not bring back the strong empiricism of frequentism. We are still positing a stochastic model so vast, with so many unspecified probabilities, that it can be neither estimated nor refuted. Instead of proclaiming a proud empiricism as our predecessors in the mid-twentieth century did, we now mumble that no models are strictly true; at best they are somehow useful.

2.2 Cournot's principle

A model is useful when it makes predictions, and predictions can be tested. In the British tradition, in which probability is either belief or frequency, significance testing is a topic separate from the meaning of probability. But another early-twentieth-century tradition made testing basic to the meaning of probability. Largely French, this tradition was a casualty of the eclipse of the Parisian school of mathematical probability by the American and Russian schools after World War II. But it retains its intellectual power, and it is an essential step towards game-theoretic probability.

2.2.1 Rise

An event with very small probability is *morally impossible*; it will not happen. Equivalently, an event with very high probability is *morally certain*; it will happen. This principle was first used within mathematical probability by Jacob Bernoulli. In his *Ars Conjectandi*, published posthumously in 1713, Bernoulli proved that in a sufficiently long sequence of independent trials of an event, there is a very high probability that the frequency with which the event happens will be close to its probability. Bernoulli explained that we can treat the very high probability as moral certainty and so use the frequency of the event as an estimate of its probability.

Augustin Cournot, a mathematician now remembered as an economist and a philosopher of science [35], gave the discussion a nineteenth-century cast in his 1843 treatise on probability [13]. Because he was familiar with geometric probability, Cournot could talk about probabilities that are vanishingly small. It may be mathematically possible for a heavy cone to stand in equilibrium on its vertex, he argued, but physically impossible. The event's probability is vanishingly small. Similarly, it is physically impossible for the frequency of an event in a long sequence of trials to differ substantially from the event's probability [13, pp. 57 and 106].

At the turn of the twentieth century, it was a commonplace among statisticians that one must decide what level of probability will count as practical certainty in order to apply probability theory. We find this stated explicitly in 1901, for example, in the articles by Georg Bohlmann and Ladislaus von Bortkiewicz in the section on probability in the *Encyklopädie der mathematischen Wissenschaften* [59, p. 825] [5, p. 861]. Aleksandr Chuprov, professor of statistics in Petersburg, was the champion in Russia of the importance of the principle that an event of very small probability will not happen. He called it Cournot's lemma [11, p. 167] and declared it a basic principle of the logic of the probable [53, pp. 95–96]. The idea was also used by British statisticians, and it was given the name "significance testing" by R. A. Fisher in the 1920s [21].

Saying that an event of very small or vanishingly small probability will not happen is one thing. Saying that probability theory gains empirical meaning only by ruling out the happening of such events is another. Cournot may have been the first to make this second assertion:

> . . . *The physically impossible event is therefore the one that has infinitely small probability*, and only this remark gives substance—objective and phenomenal value—to the theory of mathematical probability [13, p. 78].

As I have already suggested, Cournot's notion of infinitely small was flexible; the small probabilities involved in Bernoulli's theorem qualified [36].

From the time he first taught probability at the *École polytechnique* in Paris in 1919, the celebrated French mathematician Paul Lévy was an articulate advocate of Cournot's viewpoint. In his 1925 *Calcul des probabilités* [34], the most important book on mathematical probability published in the 1920s, Lévy insisted that the only bridge between mathematical probability and the world of experience is provided by predictions that events of vanishingly small probability will not happen. He was thinking both about statistical mechanics, where the probabilities for substantial deviations from equilibrium really are vanishingly small, and about Bernoulli's theorem, which says that there is only a very small probability that the frequency of an event's happening on repeated trials will differ much from the event's probability.

Lévy's views were widely shared in France and elsewhere in Europe. We find them in Italy in Castelnuovo's 1919 textbook [9, p. 180], whose influence is acknowledged by Lévy's French colleagues Maurice Fréchet and Maurice Halbwachs in their 1924 textbook [23]. In the 1940s, the preeminent French probabilist Émile Borel called Cournot's principle "the only law of chance" (la loi unique du hasard) [6, 7]. Neither Lévy nor Borel used the name "Cournot's principle," which was coined by Maurice Fréchet in 1949. Fréchet was responding to Oskar Anderson, who had talked about the Cournotsche Lemma (Cournot's lemma) and the Cournotsche Brücke (Cournot's bridge) [1, 2]. Anderson was following his teacher Chuprov in the use of "lemma," but Fréchet felt that "lemma," like "theorem," should be reserved for purely mathematical results and so suggested "principe de Cournot." Fréchet's coinage was used in the 1950s in French, German, and English [18, 44, 45, 60].

Fréchet and Castelnuovo are sometimes classified as frequentists, because of their emphasis on the objective meaning of probability. They did insist that probabilities are measured by frequencies, but they saw this not as a matter of definition but as a result of Bernoulli's theorem (frequency will approximate probability with high probability) combined with Cournot's principle (an event with high probability will happen). Andrei Kolmogorov is another so-called frequentist who considered Cournot's principle basic to the empirical meaning of probability; he called it "Principle B" in his famous *Grundbegriffe der Wahrscheinlichkeit*, published in 1933.

2.2.2 Fall

After around 1960 the idea that Cournot's principle is basic to the empirical meaning of probability largely disappears [52]. We occasionally find the idea asserted by Kolmogorov's students. Per Martin-Löf asserted it in 1969 [37, p. 616] (though he has told me that he learned it from Borel rather than from Kolmogorov), and Yu. V. Prokhorov and B. A. Sevast'yanov asserted it in their article on probability in the Soviet Mathematical Encyclopedia in the 1970s [43]. I know few other examples.

Why did Cournot's principle disappear? The most obvious factors were geopolitical. The language of mathematics after World War II was English, and to the extent that mathematicians discussed the philosophy of probability, they tended to follow British traditions. Kolmogorov and his Russian colleagues were seen as experts on the mathematics of probability, but they were never loquacious about philosophy, because philosophy was dangerous in the Soviet Union. Perhaps particular influence should be attributed to the United States mathematician J. L. Doob, who put stochastic processes into the measure-theoretic framework championed by Kolmogorov [20]. Doob's philosophy of probability [19] differed little from the pragmatic frequentism of his Harvard teacher Julian Lowell Coolidge [12].

In any case, the disappearance of Cournot's principle is unfortunate, because the principle is precisely the remedy needed for the ailments of frequentism. Although the shift from independent observations to stochastic processes and probability laws conditional on many variables makes it impossible to characterize the empirical content of a model's probabilities fully in terms of stable frequencies, this shift does not alter the fundamental fact that the only empirically meaningful predictions are those to which the model attaches high probability.

2.3 Ville's theorem

The game-theoretic framework that I describe in the next section depends on the fact that we can formulate Cournot's principle game-theoretically. Instead of saying that an event of small probability will not happen, we say that a strategy for placing bets will not multiply the capital risked by a large factor. In the case where the bettor can buy or sell any random variable for its expected value, this game-theoretic formulation is equivalent to the classical formulation; Jean Ville demonstrated the equivalence in his dissertation, published in 1939 [57].

Ville's formulation was actually infinitary: he showed that multiplying the stake one risks by an infinite factor is equivalent to an event of zero probability happening. Ville considered only the infinitary case, where the game continues for an infinite number of rounds and Skeptic tries to become infinitely rich, because he conceived his work as a critique of Richard von Mises's notion of a collective [61–63]; the title of his book was *Étude critique de la notion de collectif*. For von Mises, an event has a probability p only in the context of an infinite sequence of events whose limiting frequency of occurrence is p, and which obeys the further condition that any reasonable rule for selecting a subsequence will produce one with the same limiting frequency. This latter condition, von Mises thought, would keep a gambler from getting rich by betting on some of the events and not others. Ville showed by example that von Mises's condition is inadequate, inasmuch as it does not rule out the gambler's getting rich by varying his bet. Ville then showed that ruling out an arbitrary strategy's making a gambler infinitely rich, even if that strategy does vary how much and on what side to bet, is just the same as ruling out the happening of an arbitrary event of probability zero.

Unfortunately, Ville's work was largely overlooked, in general because of the eclipse of the French school of probability after World War II, and in particular because of the success of Doob's reformulation of Ville's notion of a martingale in measure-theoretic terms. Discussions of von Mises's frequentism still usually overlook Ville's emendation of it, and Ville's theorem is still not well known.

In order to state Ville's theorem, consider a sequence $Y_1, Y_2, \ldots$ of binary random variables with a joint probability distribution P. Suppose, for simplicity, that P assigns every finite sequence $y_1, \ldots, y_n$ of 0s and 1s positive probability, so that its conditional probabilities for Y_n given values of the preceding variables are always unambiguously defined. Following Ville, consider a gambler who begins with \$1 and is allowed to bet as he pleases on each round, provided that he never risks more than he has. (The condition that he never risks more than he has is needed to guarantee that the initial capital of \$1 represents the total capital he risks. If he were allowed to borrow in order to make larger bets, he would be risking more.) We can formalize this with the following protocol, where betting on Y_n is represented as buying some number s_n (possibly zero or negative) of tickets that cost $\$\mathrm{P}(Y_n = 1|Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})$ and pay $\$Y_n$.

Protocol 2. Ville's binary probability protocol

$\mathcal{K}_0 := 1$.

FOR $n = 1, 2, \ldots$:

Skeptic announces $s_n \in \mathbb{R}$.

Reality announces $y_n \in \{0, 1\}$.

$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - \mathrm{P}(Y_n = 1|Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}))$.

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 Reality moves.

Along with all the other protocols I consider in this article, this is a perfect-information sequential protocol; moves are made in the order listed, and each player sees the other player's moves as they are made. The sequence $\mathcal{K}_0, \mathcal{K}_1, \ldots$ is Skeptic's capital process.

2.3.1 Infinite horizon, zero probability

Ville's theorem says that Skeptic's getting infinitely rich in Protocol 2 is equivalent to an event of zero probability happening, in the following sense:

1. When Skeptic follows a strategy that gives s_n as a function of $y_1, \ldots, y_{n-1}$,

$$\mathrm{P}\left(\mathcal{K}_0, \mathcal{K}_1, \ldots \text{ is unbounded}\right) = 0. \tag{1}$$

2. If A is a measurable subset of $\{0, 1\}^\infty$ with $\mathrm{P}(A) = 0$, then Skeptic has a strategy that guarantees

$$\lim_{n\to\infty} \mathcal{K}_n = \infty$$

whenever $(y_1, y_2, \ldots) \in A$.

We can summarize these two statements by saying that Skeptic's being able to multiply his capital by an infinite factor is equivalent to the happening of an event with probability zero.

2.3.2 Infinite horizon, small probability

Practical applications require that we consider small probabilities rather than zero probabilities. In this case, we can use the following finitary version of Ville's theorem, which is proven in Chapter 8 of [51].

1. When Skeptic follows a strategy that gives s_n as a function of $y_1, \ldots, y_{n-1}$,

$$\mathrm{P}\left(\sup_{n=0,1,\ldots} \mathcal{K}_n \geq \frac{1}{\delta}\right) \leq \delta \tag{2}$$

   for every $\delta > 0$. (Although Ville was the first to establish Equation (2), it is now sometimes called *Doob's inequality.*)

2. If A is a measurable subset of $\{0,1\}^\infty$ with $\mathrm{P}(A) \leq \delta$, then Skeptic has a strategy that guarantees

$$\liminf_{n\to\infty} \mathcal{K}_n \geq \frac{1}{\delta}$$

   whenever $(y_1, y_2, \ldots) \in A$.

We can summarize these results by saying that Skeptic's being able to multiply his capital by a factor of $1/\delta$ or more is equivalent to the happening of an event with probability δ or less.

2.3.3 Finite horizon, small probability

If we assume that the game is played only for a fixed number of rounds N instead of being continued indefinitely, then the preceding statements simplify to the following:

1. When Skeptic follows a strategy that gives s_n as a function of $y_1, \ldots, y_{n-1}$,

$$\mathrm{P}\left(\max_{0\leq n\leq N} \mathcal{K}_n \geq \frac{1}{\delta}\right) \leq \delta \tag{3}$$

   for every $\delta > 0$.

2. If A is a subset of $\{0,1\}^N$ with $\mathrm{P}(A) < \delta$, then Skeptic has a strategy that guarantees

$$\mathcal{K}_N > \frac{1}{\delta}$$

   whenever $(y_1, \ldots, y_N) \in A$.

3 Game-theoretic probability

According to Cournot's principle, a probabilistic theory predicts an event by assigning it a very high probability. Ville's theorem allows us to say this in a different way: a probabilistic theory predicts an event by singling out a betting strategy that will multiply the capital it risks by a large or infinite factor if the event fails. In the game-theoretic approach to probability, we derive the prediction by constructing the strategy.

In Ville's work, the game-theoretic approach was merely an alternative way to study classical probability. But we can also use it to generalize classical probability to situations where no joint probability distribution is given for all the events and quantities that we may observe. Skeptic can construct strategies as soon as he is offered only a few bets, and it turns out that relatively few bets are needed in order to construct strategies that demonstrate interesting generalizations of classical theorems such as the law of large numbers, the law of the iterated logarithm, and the central limit theorem. The bets offered to Skeptic in the probability forecasting protocol of §1.1, for example, are enough.

I begin this section, in §3.1 and §3.2, by reviewing Ville's treatment of classical probability in terms of betting strategies. This work by Ville marks the beginning of the modern theory of martingales in probability. In §3.1, I review how Ville thought about martingales in the binary probability protocol. In §3.2, I discuss a martingale that Ville used to prove the strong law of large numbers in the special case of his binary protocol in which successive events are independent and all have the same probability p.

I then turn to the generalization to the case where no joint probability distribution for what we may observe is given. In §3.3, I return to the probability forecasting protocol we considered in §1.1. I discuss the weak and strong law of large numbers for this protocol, and as an illustration, I sketch a very simple proof of the weak law. In §3.4, I discuss a protocol in which predictions m_n of quantities y_n are interpreted as prices at which Skeptic can buy or sell the y_n, and I explain how we can obtain laws of large numbers even in this case. In §3.5, I discuss how far the game-theoretic approach applies beyond these simple examples, and in §3.6, I give an example of its power in practice.

3.1 Martingales in Ville's picture

Consider again Ville's binary probability protocol (Protocol 2 on p. 9), but now suppose that Skeptic begins with capital α not necessarily equal to 1, and drop the requirement that he keep his capital nonnegative.

Protocol 3. Ville's binary probability protocol again

Parameters: real number α, positive probability distribution P on $\{0,1\}^\infty$

$\mathcal{K}_0 := \alpha$.

FOR $n = 1, 2, \ldots$:

Skeptic announces $s_n \in \mathbb{R}$.

Reality announces $y_n \in \{0,1\}$.

$$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - \mathrm{P}(Y_n = 1 | Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})). \tag{4}$$

Recall the meaning of the assumption that the probability distribution P is positive: for every sequence $y_1, \ldots, y_n$ of 0s and 1, $\mathrm{P}(Y_1 = y_1, \ldots, Y_n = y_n) > 0$.

A strategy for Skeptic in Protocol 3 is a rule that specifies, for each n and each possible sequence of prior moves $y_1, \ldots, y_{n-1}$ by Reality, a move s_n for Skeptic. When such a strategy is fixed and α is given, Skeptic's capital $\mathcal{K}_n$ becomes a function of Reality's moves $y_1, \ldots, y_n$. Ville was the first to call the sequence of functions $\mathcal{K}_0, \mathcal{K}_1, \ldots$ a *martingale*.

In Protocol 3, knowing the capital process $\mathcal{K}_0, \mathcal{K}_1, \ldots$ produced by a strategy for Skeptic is equivalent to knowing the strategy and the initial capital α. You can find $\mathcal{K}_0, \mathcal{K}_1, \ldots$ from α and the strategy, and you can find α and the strategy from $\mathcal{K}_0, \mathcal{K}_1, \ldots$. This equivalence was Ville's justification for calling a capital process a martingale; before his work, "martingale" was a name for a gambler's strategy, not a name for his capital process. The name stuck; now we call capital processes martingales even in protocols where the strategy cannot be recovered from them.

Let us call a martingale $\mathcal{K}_0, \mathcal{K}_1, \ldots$ a *scoring martingale* if $\mathcal{K}_0 = 1$ and $\mathcal{K}_n$ is always nonnegative—i.e., $\mathcal{K}_n(y_1, \ldots, y_n) \geq 0$ for all n and all $y_1, \ldots, y_n$. Let us call a strategy for Skeptic that produces a scoring martingale when it starts with unit capital a *scoring strategy*. When Skeptic plays a scoring strategy starting with $\mathcal{K}_0 = 1$, he can be sure that $\mathcal{K}_n \geq 0$ for all n no matter how Reality plays. He may be risking the entire initial unit of capital, but he is not putting any other capital—his own or anyone else's—at risk.

The game-theoretic version of Cournot's principle predicts that a scoring martingale will be bounded. This prediction can be used in two ways:

Prediction. The prediction that a particular scoring martingale $\mathcal{K}_n$ is bounded can imply other predictions that are interesting in their own right. It is easy to give examples. In §3.2, we will look at a scoring martingale whose being bounded, as predicted with probability one by (1), implies the strong law of large numbers. In §3.3, we will look at a scoring martingale whose being bounded by $1/\delta$, as predicted with probability δ by (3), implies the weak law of large numbers.

Testing. The actual values of a scoring martingale test the validity of the probability distribution P. The larger these "scores" are, the more strongly P is refuted.

Ville used scoring martingales in both ways.

Equation (4), the rule for updating the capital in Protocol 3, implies that $\mathcal{K}_{n-1}$ is the expected value at time $n-1$ of the future value of $\mathcal{K}_n$:

$$\mathrm{E}(\mathcal{K}_n | y_1, \ldots, y_{n-1}) = \mathcal{K}_{n-1}(y_1, \ldots, y_{n-1}). \quad (5)$$

Because its left-hand side is equal to

$$\mathcal{K}_n(y_1, \ldots, y_{n-1}, 1) \frac{\mathrm{P}(y_1, \ldots, y_{n-1}, 1)}{\mathrm{P}(y_1, \ldots, y_{n-1})} + \mathcal{K}_n(y_1, \ldots, y_{n-1}, 0) \frac{\mathrm{P}(y_1, \ldots, y_{n-1}, 0)}{\mathrm{P}(y_1, \ldots, y_{n-1})},$$

Equation (5) is equivalent to

$$\mathrm{Q}(y_1, \ldots, y_{n-1}, 1) + \mathrm{Q}(y_1, \ldots, y_{n-1}, 0) = \mathrm{Q}(y_1, \ldots, y_{n-1}), \quad (6)$$

where Q is the function on finite strings of 0s and 1s defined by

$$\mathrm{Q}(y_1, \ldots, y_n) = \mathcal{K}_n(y_1, \ldots, y_n)\mathrm{P}(y_1, \ldots, y_n).$$

Equation (6) is necessary and sufficient for a nonnegative function Q on finite strings of 0s and 1s to define a probability measure on $\{0,1\}^\infty$. So a nonnegative martingale with respect to P is the same thing as the ratio of a probability measure Q to P—i.e., a likelihood ratio with P as denominator.

Doob imported Ville's notion of a martingale into measure-theoretic probability, where the gambling picture is not made explicit, by taking Equation (5) as the definition of a martingale. Game-theoretic probability goes in a different direction. It uses Ville's definition of a martingale in more general forecasting games, where the forecasts are given by a player in the game, not by a probability distribution P for Reality's moves. In these games, martingales are not likelihood ratios. They are simply capital processes for Skeptic.

One very important property of martingales that Ville used, which does carry over to the more general games studied in game-theoretic probability, is this: an average of scoring strategies produces the same average of the scoring martingales. In order to avoid any confusion, let me spell out what this means in the case where we are taking a simple average of two strategies and also in the general case where we are averaging a class of strategies.

Simplest case. Suppose the scoring strategies $\mathcal{S}^1$ and $\mathcal{S}^2$ recommend moves $s_n^1(y_1, \ldots, y_{n-1})$ and $s_n^2(y_1, \ldots, y_{n-1})$, respectively, on the nth round of the game. The average of $\mathcal{S}^1$ and $\mathcal{S}^2$—call it $\mathcal{S}$—recommends the move

$$s_n(y_1, \ldots, y_{n-1}) = \frac{1}{2}\left(s_n^1(y_1, \ldots, y_{n-1}) + s_n^2(y_1, \ldots, y_{n-1})\right)$$

on the nth round of the game. If we write $\mathcal{K}_n^1$, $\mathcal{K}_n^2$, and $\mathcal{K}_n$ for the corresponding scoring martingales, then

$$\mathcal{K}_n(y_1, \ldots, y_n) = \frac{1}{2}\left(\mathcal{K}_n^1(y_1, \ldots, y_n) + \mathcal{K}_n^2(y_1, \ldots, y_n)\right)$$

for all n and all $y_1, \ldots, y_n$.

General case. In general, consider strategies $\mathcal{S}^\xi$, where the index ξ ranges over a set Ξ. We may average these strategies with respect to a probability distribution μ on Ξ, obtaining a strategy $\mathcal{S}$ with moves

$$s_n(y_1, \ldots, y_{n-1}) = \int_\Xi s_n^\xi(y_1, \ldots, y_{n-1})\mu(d\xi),$$

provided only that this integral always converges. The corresponding scoring martingales then average in the same way:

$$\mathcal{K}_n(y_1, \ldots, y_n) = \int_\Xi \mathcal{K}_n^\xi(y_1, \ldots, y_n)\mu(d\xi).$$

These assertions are true because the increment $\mathcal{K}_n - \mathcal{K}_{n-1}$ is always linear in Skeptic's move s_n. This linearity is a feature of all the protocols considered in [51].

3.2 Ville's constant probability game

Ville studied some strategies for Skeptic in the special case of the preceding protocol in which $\mathrm{P}(Y_n = 1|Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})$ is always equal to p. In order to give the reader a glimpse of Ville's methods, I will now review the scoring martingale he used to prove the strong law of large numbers for this special case: Skeptic can become infinitely rich unless Reality makes the relative frequency of 1s converge to p.

Under the assumption that $\mathrm{P}(Y_n = 1|Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})$ is always equal to p, our protocol looks like this.

Protocol 4. Forecasting with a constant probability p
Parameter: real number p satisfying $0 < p < 1$
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
Skeptic announces $s_n \in \mathbb{R}$.
Reality announces $y_n \in \{0, 1\}$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p)$.

This protocol represents what classical probability calls independent Bernoulli trials: successive events that are independent and all have the same probability.

The scoring martingale Ville used to prove the strong law of large numbers for this protocol is

$$\mathcal{K}_n(y_1, \ldots, y_n) := \frac{r_n!(n - r_n)!}{(n + 1)!} p^{-r_n} q^{-(n-r_n)}, \tag{7}$$

where q is equal to $1 - p$ and r_n is the number of 1s among $y_1, \ldots, y_n$:

$$r_n := \sum_{i=1}^n y_i$$

[57, p. 52], [58], [46]. To confirm that (7) is a scoring martingale, it suffices to notice that it is a likelihood ratio with P as the denominator. In fact,

$$p^{r_n} q^{n-r_n} = \mathrm{P}(y_1, \ldots, y_n)$$

and

$$\frac{r_n!(n-r_n)!}{(n+1)!} = \int_0^1 \xi^{r_n}(1-\xi)^{n-r_n} d\xi = \mathrm{Q}(y_1, \ldots, y_n), \tag{8}$$

where Q is the probability distribution obtained by averaging probability distributions under which the events $y_i = 1$ are independent and all have a probability ξ possibly different from p.

Proposition 1 ***A classical strong law of large numbers.*** *With probability one,*

$$\lim_{n\to\infty} \frac{r_n}{n} = p. \tag{9}$$

Sketch of Proof Because (7) is a scoring martingale, Ville's theorem (Equation (1)) says that, with probability one, there is a constant C such that

$$\frac{r_n!(n-r_n)!}{(n+1)!} p^{-r_n} q^{-(n-r_n)} < C \tag{10}$$

for $n = 0, 1, \ldots$. One can deduce (9) from (10) using Stirling's formula. ∎

The reader may verify that (7) is Skeptic's capital process when he follows the strategy that prescribes the moves

$$s_n(y_1, \ldots, y_{n-1}) := \frac{1}{pq}\left(\frac{r_{n-1}+1}{n+1} - p\right)\mathcal{K}_{n-1}(y_1, \ldots, y_{n-1}). \tag{11}$$

The ratio $(r_{n-1} + 1)/(n + 1)$ is close to the relative frequency of 1s so far, $r_{n-1}/(n-1)$. Roughly speaking, the strategy (11) bets that y_n will be 1 when the relative frequency is more than p, 0 when the relative frequency is less than p. So whenever the relative frequency diverges from p, Reality must move it back towards p to keep Skeptic from increasing his capital.

The rate of convergence of $r_n/n - p$ to zero implied by (10) is $\sqrt{(\log n)/n}$ [58]. Using other scoring martingales (obtained by averaging $\xi^{r_n}(1-\xi)^{n-r_n}$ as in (8), but with respect to distributions concentrated around p rather than the uniform distribution), Ville also derived the faster convergence asserted by the law of the iterated logarithm [51, chapter 5].

We can also use martingale methods to prove other classical theorems, including the weak law of large numbers [51, chapter 6] and the central limit theorem [51, chapter 7]. Rather than pursue this path here, I will now step outside classical probability theory, into more general protocols like the probability forecasting protocol of §1.1. In these protocols, martingale methods allow us to prove theorems analogous to the classical theorems, but we must express them directly in game-theoretic terms. Instead of saying that something happens with probability one, we say that if it does not happen, Skeptic will multiply the capital he risks by an infinite factor. Instead of saying that something happens with the high probability $1-\delta$, we say that if it does not happen, Skeptic will multiply the capital he risks by the large factor $1/\delta$.

3.3 The probability forecasting game

As a first step outside classical probability, let us return to the protocol we considered in §1.1, in which probability forecasts are made by a player, Forecaster, rather than by a probability distribution P. For the sake of variety, let us consider the weak law of large numbers rather than the strong law. Accordingly, I now assume that the game has a fixed finite horizon: it ends after a large but finite number of rounds, N.

PROTOCOL 5. FINITE-HORIZON PROBABILITY FORECASTING
Parameter: natural number N
$\mathcal{K}_0 := 1$.
FOR $n = 1, \ldots, N$:
Forecaster announces $p_n \in [0,1]$.
Skeptic announces $s_n \in \mathbb{R}$.
Reality announces $y_n \in \{0,1\}$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p_n)$.

Our study of this protocol will focus on

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

The weak law of large numbers says that if N is sufficiently large, the relative frequency of 1s, $(1/N)\sum_{n=1}^{N} y_n$, should be close to the average probability forecast, $(1/N)\sum_{n=1}^{N} p_n$. This means S_N/N should be close to zero.

The weak law can be formulated and proven for Protocol 5 as follows:

Proposition 2 ***A game-theoretic weak law of large numbers.*** *There exists a scoring martingale that will exceed $1/\delta$ unless*

$$\left|\frac{S_N}{N}\right| \le \frac{1}{\sqrt{N\delta}}. \tag{12}$$

Proof Consider the strategy for Skeptic that prescribes the move

$$s_n = \frac{2S_{n-1}}{N}.$$

This produces the capital process

$$\mathcal{K}_n = 1 + \frac{1}{N}\left(S_n^2 - \sum_{i=1}^{n}(y_i - p_i)^2\right),$$

which is a scoring martingale: $\mathcal{K}_n \geq 0$ for $n = 1, \ldots, N$ because $(y_i - p_i)^2 \leq 1$ for each i. For the same reason, the hypothesis $\mathcal{K}_N \leq 1/\delta$ implies (12). ∎

When ϵ and δ are small positive numbers, and N is extremely large, so that

$$N \geq \frac{1}{\epsilon^2 \delta},$$

Proposition 2 implies that we can be as confident that $|S_N/N| \leq \epsilon$ as we are that Skeptic will not multiply his capital by the large factor $1/\delta$. The martingale in the proof appears in measure-theoretic form in Kolmogorov's 1929 proof of the weak law [27].

A positive probability distribution P on $\{0,1\}^N$ can serve as a strategy for Forecaster in Protocol 5, prescribing

$$p_n := \mathrm{P}(Y_n = 1 | Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}).$$

When such a strategy is imposed, Protocol 5 reduces to a finite-horizon version of Ville's binary protocol, Proposition 2 says that there is a scoring martingale in this finite-horizon protocol that will exceed $1/\delta$ unless

$$\left| \frac{1}{N} \sum_{n=1}^{N} (y_n - \mathrm{P}(Y_n = 1 | Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})) \right| \leq \frac{1}{\sqrt{N\delta}}, \tag{13}$$

and by Ville's theorem, this implies that

$$\mathrm{P}\left(\left| \frac{1}{N} \sum_{n=1}^{N} (y_n - \mathrm{P}(Y_n = 1 | Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})) \right| \leq \frac{1}{\sqrt{N\delta}} \right) \geq 1 - \frac{1}{\delta}.$$

When $\mathrm{P}(Y_n = 1 | Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1})$ is always equal to p, we are in a finite-horizon version of Protocol 3, and we get

$$\mathrm{P}\left(\left| \frac{1}{N} \sum_{n=1}^{N} y_n - p \right| \leq \frac{1}{\sqrt{N\delta}} \right) \geq 1 - \frac{1}{\delta},$$

the elementary version of the classical weak law usually derived using Chebyshev's inequality.

Protocol 5 does not require Forecaster to follow a strategy defined by a probability distribution for $y_1, \ldots, y_N$. He can use information from outside the game to decide how to move, or he can simply follow his whims. So instead of having a classical probability distribution, we are in what A. P. Dawid called the *prequential framework* [15]. In this framework, we cannot say that (13) has probability at least $1 - 1/\delta$, because we have no probability distribution, but Proposition 2 has the same intuitive meaning as such a probability statement. We consider Skeptic's multiplying his capital by $1/\delta$ or more against a good Forecaster unlikely in the same way as we consider the happening of an event unlikely when it is given probability δ or less by a valid probability distribution.

Although I have focused on the weak law here, it is also straightforward to prove the strong law for the probability forecasting protocol (Protocol 5, but with the game played for an infinite number of rounds instead of only N rounds). Kumon and Takemura have given a proof using a modified form of Ville's martingale (7) [30]. In [51, chapter 3], Vovk and I gave an alternative proof using a martingale extracted from Kolmogorov's 1930 proof of the strong law [28].

3.4 The bounded forecasting game

Stepping yet farther away from classical probability, I now ask the reader to consider a protocol where Forecaster's task on each round is to give not a probability for an event but a prediction of a bounded quantity. To fix ideas, I assume that the quantity and the prediction are always in the interval from 0 to 100.

PROTOCOL 6. FINITE-HORIZON BOUNDED FORECASTING
Parameter: natural number N
$\mathcal{K}_0 := 1$.
FOR $n = 1, \ldots, N$:
Forecaster announces $m_n \in [0, 100]$.
Skeptic announces $s_n \in \mathbb{R}$.
Reality announces $y_n \in [0, 100]$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - m_n)$.

As it turns out, the game-theoretic treatment of Protocol 6 is scarcely different from that of Protocol 5. Setting

$$S_n := \sum_{i=1}^{n}(y_i - m_i),$$

we reason almost as before.

Proposition 3 ***A weak law for bounded forecasting.*** *There exists a scoring martingale that will exceed $1/\delta$ unless*

$$\left|\frac{S_N}{N}\right| \leq \frac{100}{\sqrt{N\delta}}. \tag{14}$$

Proof Consider the strategy for Skeptic that prescribes the move

$$s_n = \frac{2S_{n-1}}{10000N}.$$

This produces

$$\mathcal{K}_n = 1 + \frac{1}{10000N}\left(S_n^2 - \sum_{i=1}^{n}(y_i - m_i)^2\right),$$

which is nonnegative because $(y_i - m_i)^2 \leq 10000$ for each i. For the same reason, the hypothesis $\mathcal{K}_N \leq 1/\delta$ implies (14). ∎

The proofs of the strong law for the infinite-horizon version of Protocol 5 that we just cited, in [30] and [51, chapter 3], also apply to the infinite-horizon version of Protocol 6.

3.5 Scope of the framework

According to the conventional wisdom, probability theory is based on a handful of axioms and definitions that were formulated concisely by Andrei Kolmogorov in 1933 [29, 52]. The reader might ask for a similarly concise statement of the game-theoretic framework. I have given examples, but where is the general definition of what counts as game-theoretic probability? Is there a single protocol of which all probabilistic protocols are special cases?

I think not. Game theory is remarkably flexible, and the simple perfect-information protocols I have presented can be varied in many interesting and perhaps useful ways. It seems futile to try to draw sharp boundaries between games that are part of game-theoretic probability and those that are not. We should think of game-theoretic probability as a Wittgensteinian object, with canonical instances but fuzzy boundaries.

The canonical instances are perfect-information games with three players, Forecaster, Reality, and Skeptic. The payoff to Skeptic is linear in Skeptic's move, so that we can average scoring strategies and thereby average scoring martingales, as explained in §3.2. The players' moves may be one-dimensional, as in the examples we have considered so far, or they may be multidimensional. In addition to determining the outcome and therefore the increment in Skeptic's capital, Reality may also provide auxiliary information to Forecaster and Skeptic before they make their moves. This suggests the following protocol.

PROTOCOL 7. LINEAR FORECASTING
Parameters: set $\mathbf{X}$, subset $\mathbf{Y}$ of $\mathbb{R}^k$
$\mathcal{K}_0 := 1$.
FOR $n = 1, 2, \ldots$:
Reality announces $x_n \in \mathbf{X}$.
Forecaster announces $f_n \in \mathbb{R}^k$.
Skeptic announces $s_n \in \mathbb{R}^k$.
Reality announces $y_n \in \mathbf{Y}$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n \cdot (y_n - f_n)$.

Here $s_n \cdot (y_n - f_n)$ is the dot product of the k-dimensional vectors s_n and $y_n - f_n$.

Protocol 7 covers many prediction problems considered in statistics (where x and y are often called *independent* and *dependent* variables, respectively) and machine learning (where x is called the *object* and y the *label*) [25, 56, 73]. As we will see in a moment, it also covers some market games. Yet many other interesting protocols relevant to probability, including many in [51], are not special cases of Protocol 7. These include protocols in which Skeptic can bet on only one side (i.e., cannot necessarily replace a permitted move s_n by $-s_n$) and others involving additional players. Some of the results of game-theoretic probability also generalize from Euclidean spaces to more general spaces [51, §4.6].

3.6 Example: the $\sqrt{dt}$ effect

Organized exchanges, in which a buyer or seller can always find a ready price for a particular commodity or security, are forecasting games. It is often said that in an efficient market, an investor cannot make a lot of money without taking undue risk. Cournot's principle makes this precise by saying that he will not multiply a fixed stake without additional risk. To illustrate the power of the game-theoretic framework, I want now to point out that Cournot's principle alone can account for the stylized fact that changes in market prices over an interval of time of length dt scale as $\sqrt{dt}$.

In a securities market where shares are traded 252 days a year, for example, the typical change in price of a share from one year to the next is $\sqrt{252}$, or about 16, times as large as the typical change from one day to the next. There is a standard way of explaining this. We begin by assuming, in the spirit of Neyman, that price changes are governed by some "chance mechanism"; they have an objective but unknown probability distribution. Then we argue that successive changes must be uncorrelated; otherwise someone who knew the correlation (or learned it by observation) could devise a trading strategy with positive expected value. Uncorrelatedness of 252 successive daily price changes implies that their sum, the annual price change, has variance 252 times as large and hence standard deviation, or typical value, $\sqrt{252}$ times as large. This is a simple argument, but it uses the mysterious chance mechanism twice, first when we use the probabilistic concept of correlation, and then when we interpret market efficiency as the absence of a trading strategy with positive expected value. As I now explain, we can replace the appeal to a chance mechanism with a purely game-theoretic argument, in which Cournot's principle expresses the assumption of market efficiency.

For simplicity, consider the following protocol, which describes a market in shares of a corporation. Investor plays the role of Skeptic; he tries to make money, and Cournot's principle says he will not do so without risking more than his initial stake, which we take to be \$1. Market plays the roles of Forecaster (by giving opening prices) and Reality (by giving closing prices). We suppose that today's opening price is yesterday's closing price, so that Market gives only one price each day, at the end of the day. When Investor holds s_n shares during day n, he makes $s_n(y_n - y_{n-1})$, where y_n is the price at the end of day n.

PROTOCOL 8. DAILY TRADING IN THE MARKET FOR A SECURITY

$\mathcal{K}_0 := 1$.
Market announces $y_0 \in \mathbb{R}$.
FOR $n = 1, 2, \ldots, N$:
Investor announces $s_n \in \mathbb{R}$.
Market announces $y_n \in \mathbb{R}$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - y_{n-1})$.

For simplicity, I ignore the fact that the price y_n of a share cannot be negative. This does not invalidate our results; they are worst case results, telling us what Investor can accomplish regardless of Market's behavior, and so they are not invalidated by a further restriction on Market's moves. Stronger results may be obtained, however, when the restriction $y_n \geq 0$ is imposed, see [75].

Since no chance mechanism is assumed to be operating, we cannot appeal to the idea of the variance of a probability distribution for price changes to explain what $\sqrt{dt}$ scaling means. But we can use

$$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - y_{n-1})^2} \tag{15}$$

as the typical daily change, and we can compare it to the magnitude of the change we see over the whole game, say

$$\max_{0<n\leq N} |y_n - y_0|. \tag{16}$$

The quantity (16) should have the same order of magnitude as $\sqrt{N}$ times the quantity (15). Equivalently, we should have

$$\sum_{n=1}^{N}(y_n - y_{n-1})^2 \sim \max_{0<n\leq N}(y_n - y_0)^2, \tag{17}$$

where $\sim$ is understood to mean that the two quantities have the same order of magnitude.

Does Cournot's principle give us any reason to think that (17) should hold? Indeed it does. As it turns out, there is a scoring martingale that becomes large (makes a lot of money for Investor) if (17) is violated. Market (whose moves are affected by all the participants in the market, but only slightly by Investor) tends to set prices so that Investor will not make a lot of money, and as we will see in §4, he can more or less do so. So we may expect (17) to hold.

The scoring strategy that makes money if (17) is violated is an average of two scoring strategies, a momentum strategy (hold more shares after the price goes up), and a contrarian strategy (hold more shares after the price goes down).

1. The momentum strategy turns \$1 into \$$D/E$ or more if $\sum(y_n - y_{n-1})^2 \leq E$ and $\max(y_n - y_0)^2 \geq D$.

2. The contrarian strategy turns \$1 into $\$E/D$ or more if $\sum(y_n - y_{n-1})^2 \geq E$ and $\max(y_n - y_0)^2 \leq D$.

For details, see [75].

4 Defensive Forecasting

I now turn from strategies for Skeptic, which test Forecaster, to strategies for Forecaster, which seek to withstand such tests. Forecaster engages in *defensive forecasting* when he plays in order to defeat a particular strategy for Skeptic.

As it turns out, Forecaster can defeat any particular strategy for Skeptic, provided only that each move prescribed by the strategy varies continuously with respect to Forecaster's previous move. Forecaster wants to defeat more than a single strategy for Skeptic. He wants to defeat simultaneously all the scoring strategies Skeptic might use. But as we will see, Forecaster can often amalgamate the scoring strategies he needs to defeat by averaging them, and then he can play against the average. Defeating the average may be good enough, because when any one of the scoring strategies rejects Forecaster's validity, the average will generally reject as well, albeit less strongly.

I begin this section, in §4.1, with an overview of related work on prediction. Then I explain defensive forecasting in the case of binary forecasting, leaving the reader to consult other expositions for other protocols [71, 74]. In §4.2, I review the simple proof that Forecaster can defeat any continuous strategy for Skeptic in the binary case. In §4.3, I explain how to average strategies for Skeptic to test for the calibration of binary probability forecasts, and in §4.4, I explain how resolution can also be achieved. In §4.5, I make some final comments about the significance of these results.

4.1 Context

Although the names "Forecaster," "Skeptic," and "Reality" may not have been used systematically prior to [51], a good deal of prior work on probabilistic prediction can be understood in terms of averaging strategies in a game in which, for successive values of n, Forecaster predicts the value of y_n, Skeptic then bets on the value of y_n, and Reality then determines the value of y_n.

This work can be classified into the four cells of a 2×2 table based to these two distinctions:

Forecast-based vs. test-based. Work is *forecast-based* if it involves averaging or otherwise amalgamating strategies for Forecaster. Work is *test-based* if it amalgamates strategies for Skeptic and then seeks a strategy for Forecaster that is optimal against the resulting strategy for Skeptic.

Universalizing vs. amalgamating. Work is *universalizing* if it seeks to amalgamate all possible strategies for a particular player, thus obtaining a strategy that is universal for that player. Universal strategies are usually not computable. Work is (merely) *amalgamating* if it attempts only to amalgamate a more manageable class of strategies of the player concerned, which may be sufficient in practice.

Defensive forecasting is test-based and merely amalgamating.

A substantial amount of work has been done in each of the four cells.

Forecast-based and universalizing. The only work in this cell of which I am aware is that of Ray Solomonoff, who has argued for decades that it is possible to average all computable strategies for Forecaster, thus obtaining a universal strategy that will be successful regardless of Reality's behavior [54]. Solomonoff's universal strategy is only semi-computable, and it is not clear that useful approximations can be implemented.

Forecast-based and amalgamating. An immense amount of work falls in this cell, including all the work on Bayesian prediction [3] and prediction with expert advice [10, 64]. All this work seeks to improve strategies for Forecaster (probability distributions for Reality's moves $y_1, y_2, \dots$) by averaging or otherwise amalgamating them. Averaging strategies for Forecaster can be effective asymptotically because the average will eventually behave like the strategy in the average that best matches the behavior of Reality. Whether this happens in practice depends on how many rounds the prediction game runs and how close Reality's behavior is to any of the strategies being averaged.

Test-based and universalizing. The best known work in this cell is that of Leonid Levin and Peter Gács. Their work has roots in a celebrated 1966 article by Per Martin-Löf [37], which established the existence of universal tests in the case of Bernoulli trials. In the early 1970s, Levin generalized Martin-Löf's result beyond the binary case and showed that the amalgamation of individual tests can be carried out by averaging [31, 32]. Levin established the existence of a "neutral" probability distribution, one that defeats the universal test [32, 33], and this result has been developed further by Gács [24]. The result is interesting to those who study randomness, because it asserts the existence of a probability distribution with respect to which any sequence $y_1, y_2, \dots$ is random.

Test-based and amalgamating. This includes work on well-calibrated randomized forecasting by Foster and Vohra [22, 47], and more recent work on well-calibrated deterministic calibration by Kakade and Foster [26].

Defensive forecasting goes beyond the other test-based amalgamating work primarily because it uses the game-theoretic framework and can therefore defend against any strategy by Skeptic, not just strategies that test calibration and resolution. This is analogous to the difference between von Mises, who was concerned only with frequencies, and Ville, who considered all properties of probabilities. The added power of the game-theoretic framework may be of limited importance in the binary case, where calibration and resolution are most important. But it is key for the generalization to other forecasting games.

4.2 Defeating a continuous strategy for Skeptic

In this section, I repeat a simple proof, first given in [77], showing that Forecaster can defeat any particular fixed strategy for Skeptic, provided that each move prescribed by the strategy varies continuously with respect to Forecaster's previous move

Consider a strategy $\mathcal{S}$ for Skeptic in Protocol 1, the binary forecasting protocol we studied in §1.1. Write

$$\mathcal{S}(p_1, y_1, \ldots, p_{n-1}, y_{n-1}, p_n) \tag{18}$$

for the move that $\mathcal{S}$ prescribes on the nth round of the game. At the beginning of the nth round, just before Forecaster makes his move p_n, the earlier moves $p_1, y_1, p_1, \ldots, p_{n-1}, y_{n-1}$ are known and therefore fixed, and so (18) becomes a function of p_n only, say

$$\mathcal{S}_n(p_n).$$

This defines a function $\mathcal{S}_n$ on the interval $[0, 1]$. If $\mathcal{S}_n$ is continuous for all n and all $p_1, y_1, \ldots, p_{n-1}, y_{n-1}$, let us say that $\mathcal{S}$ is *forecast-continuous.*

When we fix the strategy $\mathcal{S}$, Skeptic no longer has a role to play in the game, and we can omit him from the protocol.

PROTOCOL 9. BINARY PROBABILITY FORECASTING AGAINST A FIXED TEST **Parameter:** Strategy $\mathcal{S}$ for Skeptic $\mathcal{K}_0 := \alpha$. FOR $n = 1, 2, \ldots$: Forecaster announces $p_n \in [0, 1]$. Reality announces $y_n \in \{0, 1\}$. $\mathcal{K}_n := \mathcal{K}_{n-1} + \mathcal{S}_n(p_n)(y_n - p_n)$.

Proposition 4 *If the strategy $\mathcal{S}$ is forecast-continuous, Forecaster has a strategy that ensures $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \geq \cdots$.*

Proof By the intermediate-value theorem, the continuous function $\mathcal{S}_n$ is always positive, always negative, or else satisfies $\mathcal{S}_n(p) = 0$ for some $p \in [0, 1]$. So Forecaster can use the following strategy:

- if $\mathcal{S}_n$ is always positive, take $p_n := 1$;
- if $\mathcal{S}_n$ is always negative, take $p_n := 0$;
- otherwise, choose p_n so that $\mathcal{S}_n(p_n) = 0$.

This guarantees that $\mathcal{S}_n(p_n)(y_n - p_n) \leq 0$, so that $\mathcal{K}_n \leq \mathcal{K}_{n-1}$. ∎

If the reader finds it confusing that the notation $\mathcal{S}_n(p_n)$ leaves the dependence on the earlier moves $p_1, y_1, p_1, \ldots, p_{n-1}, y_{n-1}$ implicit, he or she may wish to think about the following alternative protocol, which leaves Skeptic in the game and has him announce the function $\mathcal{S}_n$ just before Forecaster makes his move p_n:

PROTOCOL 10. WHEN SKEPTIC CHOOSES A STRATEGY ON EACH ROUND
$\mathcal{K}_0 := \alpha$.
FOR $n = 1, 2, \ldots$:
Skeptic announces continuous $\mathcal{S}_n : [0, 1] \to \mathbb{R}$.
Forecaster announces $p_n \in [0, 1]$.
Reality announces $y_n \in \{0, 1\}$.
$\mathcal{K}_n := \mathcal{K}_{n-1} + \mathcal{S}_n(p_n)(y_n - p_n)$.

Protocol 10 gives Skeptic a little more flexibility than Protocol 9 does; it allows Skeptic to take into account information coming from outside the game as well as the previous moves $p_1, y_1, \ldots, p_{n-1}, y_{n-1}$ when he decides on $\mathcal{S}_n$. But it still requires that $\mathcal{S}_n(p_n)$ depend on p_n continuously, and it still makes sure that Forecaster knows $\mathcal{S}_n$ before he makes his move p_n. (Like all protocols in this article, Protocol 10 is a perfect-information protocol; the players move in sequence and see each other's moves as they are made.) So the proof and conclusion of Proposition 4 still hold: Forecaster has a strategy that ensures $\mathcal{K}_0 \geq \mathcal{K}_1 \geq \mathcal{K}_2 \geq \cdots$.

As I argued in (1.1), the requirement that $\mathcal{S}_n$ be continuous does not restrict the practical significance of the result, because a continuous function can change with arbitrarily great abruptness. This is consistent with the views of L. E. J. Brouwer, who argued that the idealized concept of computability for real-valued functions (idealized because real numbers are already idealized objects) should include the requirement of continuity [8, 38]. Notice also that the strategies for Skeptic used in §3 to establish classical results in probability theory (as well as all those used in [51]) are continuous. Finally, as I noted in §1.1, we can get a version of Proposition 4 even if we permit Skeptic to be discontinuous, provided we allow Forecaster to randomize his forecasts a little [66, 76].

4.3 Calibration

I now exhibit a forecast-continuous strategy $\mathcal{S}$ for Skeptic in Protocol 9 that multiplies Skeptic's capital by a large factor whenever Forecaster's probabilities fail to be well calibrated. According to Proposition 4, Forecaster can give probabilities that are well calibrated by playing against $\mathcal{S}$.

This subsection follows closely the more general exposition in Vladimir Vovk's "Non-asymptotic calibration and resolution" [72].

4.3.1 Goal

Leaving aside Skeptic for a moment, consider binary forecasting with Forecaster and Reality alone:

Protocol 11. Forecaster & Reality

FOR $n = 1, \dots, N$:

Forecaster announces $p_n \in [0, 1]$.

Reality announces $y_n \in \{0, 1\}$.

If a forecaster is doing a good job, we expect about 30% of the events to which he gives probabilities near 30% to happen. If this is true not only for 30% but also for all other probability values, then we say the forecaster is *well calibrated*. Here are two ways to apply this somewhat fuzzy concept to the probabilities $p_1, \dots, p_N$ and outcomes $y_1, \dots, y_N$.

1. For a given $p \in [0, 1]$, divide the p_n into those we consider near p and those we consider not near. Write $\sum_{n:p_n \approx p} 1$ for the number in the group we consider near p. Divide this group further into those for which y_n is equal to 0 and those for which y_n is equal to 1. Write $\sum_{n:p_n \approx p} y_n$ for the number in the latter group. Thus the fraction of 1s on those rounds where p_n is near p is
$$\frac{\sum_{n:p_n \approx p} y_n}{\sum_{n:p_n \approx p} 1}. \tag{19}$$
We say the forecasts are well calibrated at p if (19) is approximately equal to p. We say the forecasts are well calibrated overall if they are well calibrated at p for every p for which the denominator of (19), the number of p_n near p, is large.

2. Alternatively, instead of dividing the p_n sharply into those we consider near p and those we consider not near, we can introduce a continuous measure $K(p_n, p)$ of nearness, a function such as the *Gaussian kernel*
$$K(p_n, p) := e^{-\gamma(p_n - p)^2} \tag{20}$$
(with $\gamma > 0$), which takes the value 1 when p_n is exactly equal to p, a value close to 1 when p_n is near p, and a value close to 0 when p_n is far from p. Then we say the forecasts are well calibrated if
$$\frac{\sum_{n=1}^{N} K(p_n, p) y_n}{\sum_{n=1}^{N} K(p_n, p)} \approx p \tag{21}$$
for every p for which the denominator of the left-hand side is large.

The second approach, making the closeness of p_n and p continuous, fits here, because Skeptic can enforce (21) using a forecast-continuous strategy.

Recall that a function $K : Z^2 \to \mathbb{R}$ is a called a *kernel on Z* if it satisfies two conditions:

- It is symmetric: $K(p, p') = K(p', p)$ for all $p, p' \in Z$.
- It is positive definite: $\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j K(z_i, z_j) \geq 0$ for all real numbers $\lambda_1, \dots, \lambda_m$ and all elements $z_1, \dots, z_m$ of Z.

The Gaussian kernel (20) is a kernel on $[0,1]$ in this sense. It also satisfies $0 \le K(p,p') \le 1$ for all $p,p' \in Z$. And as I have already noted, $K(p,p')$ is 1 when $p = p'$, close to 1 when p' is near p, and close to 0 when p' is far from p. Our reasoning in this section works when K is any kernel on $[0,1]$ satisfying these conditions.

The goal that (21) hold for all p for which $\sum_{n=1}^N K(p_n,p)$ is large is fuzzy on two counts: "$\approx p$" is fuzzy, and "large" is fuzzy. But reasonable interpretations of these two fuzzy predicates lead to the following precise goal.

Goal

$$\left|\sum_{n=1}^N K(p_n,p)(y_n - p_n)\right| \le \sqrt{N} \tag{22}$$

Heuristic Proposition 1 *Under reasonable interpretations of "$\approx p$" and "large," (22) implies that (21) holds for all p for which $\sum_{n=1}^N K(p_n,p)$ is large.*

Explanation Condition (21) is equivalent to

$$\frac{\sum_{n=1}^N K(p_n,p)(y_n - p)}{\sum_{n=1}^N K(p_n,p)} \approx 0. \tag{23}$$

We are assuming that $K(p_n,p)$ is small when p is far from p_n. This implies that

$$\frac{\sum_{n=1}^N K(p_n,p)(p_n - p)}{\sum_{n=1}^N K(p_n,p)} \approx 0$$

when $\sum_{n=1}^N K(p_n,p)$ is large. So (23) holding when $\sum_{n=1}^N K(p_n,p)$ is large is equivalent to

$$\frac{\sum_{n=1}^N K(p_n,p)(y_n - p_n)}{\sum_{n=1}^N K(p_n,p)} \approx 0 \tag{24}$$

holding when $\sum_{n=1}^N K(p_n,p)$ is large. If we take $\sum_{n=1}^N K(p_n,p)$ being large to mean that

$$\sum_{n=1}^N K(p_n,p) \gg \sqrt{N},$$

then this condition together with (22) implies

$$\frac{\left|\sum_{n=1}^N K(p_n,p)(y_n - p_n)\right|}{\sum_{n=1}^N K(p_n,p)} \ll 1,$$

which is a reasonable interpretation of (24). ∎

4.3.2 Betting strategy and forecasting algorithm

In order to derive a strategy for Forecaster that guarantees the goal (22) in Protocol 11, let us imagine that Skeptic is allowed to enter the game and bet, as in Protocol 10.

Our strategy for Forecaster will be to play against the following forecast-continuous strategy for Skeptic.

K29 Betting Strategy

$$\mathcal{S}_n(p) = \sum_{i=1}^{n-1} K(p, p_i)(y_i - p_i)$$

We call this strategy for Skeptic the K29 betting strategy because it is modeled on the martingale that Kolmogorov used in [27] to prove the weak law of large numbers (see §3.3).

According to the proof of Proposition 4, playing against the K29 betting strategy means using the following algorithm to choose p_n for $n = 1, \ldots, N$.

K29 Forecasting Algorithm

- If the equation
$$\sum_{i=1}^{n-1} K(p, p_i)(y_i - p_i) = 0 \qquad (25)$$
has at least one solution p in the interval $[0, 1]$, set p_n equal to such a solution.
- If $\sum_{i=1}^{n-1} K(p, p_i)(y_i - p_i) > 0$ for all $p \in [0, 1]$, set p_n equal to 1.
- If $\sum_{i=1}^{n-1} K(p, p_i)(y_i - p_i) < 0$ for all $p \in [0, 1]$, set p_n equal to 0.

4.3.3 Why the forecasting algorithm works

Why does the K29 forecasting algorithm work? I will give a formal proof that it works and then discuss its success briefly from an intuitive viewpoint.

Proposition 5 *Suppose Forecaster plays in Protocol 11 using the K29 algorithm with a kernel K satisfying $K(p, p) \leq 1$ for all $p \in [0, 1]$. Then (22) will hold no matter how Reality chooses $y_1, \ldots, y_N$.*

Proof Mercer's theorem says that for any kernel K on $[0, 1]$, there is a mapping Φ (called a *feature mapping*) of $[0, 1]$ into a Hilbert space H such that

$$K(p, p') = \Phi(p) \cdot \Phi(p') \qquad (26)$$

for all p, p' in $[0, 1]$, where $\cdot$ is the dot product in H [39, 48].

Under the K29 betting strategy, Skeptic's moves are

$$s_n = \sum_{i=1}^{n-1} K(p_n, p_i)(y_i - p_i).$$

Proposition 4 says that when Forecaster plays the K29 forecasting algorithm against this strategy, Skeptic's capital does not increase. So

$$\begin{aligned}0 \geq \mathcal{K}_N - \mathcal{K}_0 &= \sum_{n=1}^{N} s_n(y_n - p_n) = \sum_{n=1}^{N} \sum_{i=1}^{n-1} K(p_n, p_i)(y_n - p_n)(y_i - p_i) \\ &= \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{N} K(p_n, p_i)(y_n - p_n)(y_i - p_i) - \frac{1}{2} \sum_{n=1}^{N} K(p_n, p_n)(y_n - p_n)^2 \\ &= \frac{1}{2} \left\| \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n) \right\|^2 - \frac{1}{2} \sum_{n=1}^{N} \left\| (y_n - p_n)\Phi(p_n) \right\|^2 .\end{aligned}$$

Hence

$$\left\| \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n) \right\|^2 \leq \sum_{n=1}^{N} \left\| (y_n - p_n)\Phi(p_n) \right\|^2 . \tag{27}$$

From (27) and the fact that

$$\|\Phi(p)\| = \sqrt{K(p,p)} \leq 1,$$

we obtain

$$\left\| \sum_{n=1}^{N} (y_n - p_n)\Phi(p_n) \right\| \leq \sqrt{N}. \tag{28}$$

Using (26), the Cauchy-Schwartz inequality, and then (28), we obtain

$$\begin{aligned}\left| \sum_{n=1}^{N} K(p_n, p)(y_n - p_n) \right| &= \left| \left(\sum_{n=1}^{N} \Phi(p_n)(y_n - p_n) \right) \cdot \Phi(p) \right| \\ &\leq \left\| \sum_{n=1}^{N} \Phi(p_n)(y_n - p_n) \right\| \|\Phi(p)\| \leq \sqrt{N}.\end{aligned}$$

∎

Additional insight about the success of the K29 forecasting algorithm can be gleaned from Equation (25):

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

In practice, this equation usually has a unique solution in $[0,1]$, and so this solution is Forecaster's choice for p_n. Because $K(p, p_i)$ is small when p and p_i

are far apart, the main contribution to the sum comes from terms for which p_i is close to p. So the equation is telling us to look for a value of p such that $y_i - p_i$ average to zero for i such that p_i is close to p. This is precisely what calibration requires. So we can say that on each round, the algorithm *chooses as p_n the probability value where calibration is the best so far.*

The pertinence of this formulation becomes clear when we recognize that Forecaster's calibration cannot be rejected because of what Reality does on a single round of the forecasting game. A statistical test of calibration (such as the K29 betting strategy) can reject calibration (multiply its initial capital by a large factor) only as a result of a trend involving many trials. To reject calibration at p, we must see many rounds with p_i near p, and the relative frequency of 1s for these p_i must diverge substantially from p. The K29 formulation avoids any such divergence by always putting the next p_n at values of p where no trend is emerging—where, on the contrary, calibration is excellent so far.

This is only natural. In general, we avoid choices that have worked out poorly in the past. The interesting point here is that this is sufficient to avoid rejection by a statistical test. Forecaster can make sure he is well calibrated by acting as if the future will be like the past, regardless of what Reality does on each round.

4.3.4 More to say

Calibration is only one probabilistic property of the forecasts $p_1, \ldots, p_N$ in Protocol 10 that we might demand in order to count Forecaster as a good probability forecaster. For example, we might demand that the relative frequency of 1s on rounds for which p_i is close to p converge to p at the rate described by the law of the iterated logarithm. This demand can also be satisfied using Proposition 4. Because violation of the law of the iterated logarithm is an event of small probability in classical probability, Ville's theorem tells us that there is a strategy for Skeptic that multiplies the capital risked by a large factor if the law is violated. Such a strategy is constructed in [51, Chapter 5], and Proposition 4 tells us that Forecaster will satisfy the law of the iterated logarithm by playing against it.

There is much more to say. For example:

- Proposition 4 does more than prevent Skeptic from multiplying the capital he risks by a large factor: it prevents him from making any money at all. This means that defensive forecasts do even better, with respect to tests being defended against, than classical probability theory expects. Rejection by these tests is avoided for certain, not merely with high probability.
- Forecaster's strategy on each round n does not depend on the horizon N. So if the game is played indefinitely, the goal (calibration in the case of the K29 forecasting algorithm) is achieved for every N.

See [72] for a more comprehensive discussion.

4.4 Resolution

Probability forecasting is usually based on more information than the success of previous forecasts for the various probability values. If rainfall is more common in April than May, for example, then a weather forecaster should take this into account. It should rain on 30% of the April days for which he gives rain a probability of 30% and also on 30% of the May days for which he gives rain a probability 30%. This property is stronger than mere calibration, which requires only that it rain on 30% of all the days for which the forecaster says 30%. It is called *resolution*.

To see that defensive forecasting can achieve resolution as well as mere calibration, we can introduce the auxiliary information x_n in the way explained in §3.5:

> PROTOCOL 12. FORECASTING WITH AUXILIARY INFORMATION x_n
> **Parameter:** natural number N, set $\mathbf{X}$
> $\mathcal{K}_0 := \alpha$.
> FOR $n = 1, \dots, N$:
> 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 $y_n \in \{0,1\}$.
> $\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n)$.

In this context, we need a kernel $K : ([0,1] \times \mathbf{X})^2 \to [0,1]$ to measure the nearness of (p,x) to (p',x'). We may choose it so that $K((p,x)(p',x'))$ is 1 when $(p,x) = (p',x')$, close to 1 when (p,x) is near (p',x'), and close to 0 when (p,x) is far from (p',x'). Once we have chosen such a kernel, we may say that the forecasts have good resolution if

$$\frac{\sum_{n=1}^{N} K((p_n,x_n)(p,x))y_n}{\sum_{n=1}^{N} K((p_n,x_n)(p,x))} \approx p$$

for every pair (p,x) for which the denominator of the left-hand side is large.

This is a straightforward generalization of calibration, and the entire theory that I have reviewed for calibration generalizes directly. In the generalization, the K29 betting strategy is

$$\mathcal{S}_n(p) = \sum_{i=1}^{n-1} K((p,x_n),(p_i,x_i))(y_i - p_i),$$

and K29 forecasting algorithm achieves good resolution by playing against it. The K29 forecasting strategy is again very natural. It chooses p_n to satisfy

$$\sum_{i=1}^{n-1} K((p_n,x_n),(p_i,x_i))(y_i - p_i) = 0$$

In other words, it chooses p_n so that we already have good resolution for (p_i, x_i) near (p_n, x_n). This is analogous to the practice of varying one's actions with the situation in accordance with past experience in different situations. If x is the friend I am spending time with and p is our activity, and experience tells me that bowling has been the most enjoyable activity with Tom, bridge with Dick, and tennis with Harry, then this is how I will choose in the future.

See [72] for details.

4.5 Implications

There is much more to say about defensive forecasting, much of it in recent papers by Vladimir Vovk [65–72]. I cannot begin to summarize this work here, but I should mention that probabilities produced by defensive forecasting can often be used in decision problems, where calibration and resolution often suffice to guarantee that decisions derived from them are optimal in the same sense as decisions derived from classical probabilities.

Some readers may find it unsurprising that experience can always be used to make probability forecasts that pass statistical tests. It is well known, after all, that probabilities can be estimated consistently using a random sample (see, e.g., [55]). But this well-known result is based on a strong assumption about reality's behavior in the future: that past and future observations are independently drawn from the same probability distribution. Defensive forecasting, in contrast, gives probabilities that pass statistical tests without using any advance knowledge about how reality will behave. For me, at least, this is surprising and full of implications.

In my own previous work, I tried to understand the weighing of evidence when we do not have classical probabilities [49] and the meaning of probabilistic causality [50]. Learning about the possibility of defensive forecasting has altered my thinking on both these topics.

- I now think that the main feature distinguishing the domain where we can use probabilities from the domain where we need other methods (such as those I studied in [49]) is the presence of a structure for repetition. As soon as we have a game with repeated rounds and we are concerned with long run performance in that structure, not with one particular case, we can use probability theory, at least in its game-theoretic form. But when we are concerned with a particular case, which different people may place in different sequences of similar cases or different games, we must weigh arguments in ways that go beyond probability theory.

- In [50], I argued that probabilistic causal relations can be fully understood in terms of the possibilities for probabilistic prediction; counterfactual worlds are not needed. I continue to believe this, but I now see the possibility of probabilistic prediction as depending only on the availability of the auxiliary information on which it is based (the x_n in Protocol 12), not on prior knowledge of probabilities.

These are the ideas I would most like to clarify in future work.

References

[1] Oskar N. Anderson. *Einführung in die mathematische Statistik.* Springer, Vienna, 1935.

[2] Oskar N. Anderson. Die Begründung des Gesetzes der grossen Zahlen und die Umkehrung des Theorems von Bernoulli. *Dialectica*, 3(9/10):65–77, 1949.

[3] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 2000.

[4] Jacob Bernoulli. *Ars Conjectandi.* Thurnisius, Basel, 1713. In 2006, The Johns Hopkins University Press published Edith Sylla's translation of the book into English, *The Art of Conjecturing together with Letter to a Friend on Sets in Court Tennis.* Oscar Sheynin's translation of Part IV into English, dated 2005, is at www.sheynin.de.

[5] Georg Bohlmann. Lebensversicherungs-Mathematik. In *Encyklopädie der mathematischen Wissenschaften, Bd. I, Teil 2*, pages 852–917. Teubner, Leipzig, 1901.

[6] Émile Borel. *Les probabilités et la vie.* Presses Universitaires de France, Paris, 1943. Later editions in 1946, 1950, 1958, and 1967. The 1958 edition was translated into English by Maurice An English translation of the 1958 edition, *Probabilities and Life*, was published in 1962 by Dover, New York.

[7] Émile Borel. *Probabilité et certitude.* Presses Universitaires de France, Paris, 1950. An English translation, *Probability and Certainty*, was published in 1963 by Walker, New York.

[8] 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.

[9] Guido Castelnuovo. *Calcolo delle probabilitá.* Albrighi e Segati, Milan, Rome, and Naples, 1919. Second edition in two volumes, 1926 and 1928. Third edition 1948.

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

[11] Aleksandr A. Chuprov. Очерки по теории статистики *(Essays on the theory of statistics).* Sabashnikov, Saint Petersburg, second edition, 1910. The first edition appeared in 1909. The second edition was reprinted by the State Publishing House, Moscow, in 1959.

[12] Julian Lowell Coolidge. *An Introduction to Mathematical Probability.* Oxford University Press, London, 1925.

[13] Antoine-Augustin Cournot. *Exposition de la théorie des chances et des probabilités*. Hachette, Paris, 1843. Reprinted in 1984 as Volume I (B. Bru, editor) of [14].

[14] Antoine-Augustin Cournot. *Œuvres complètes*. Vrin, Paris, 1973–1984. 10 volumes, with an eleventh to appear.

[15] A. Philip Dawid. Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society. Series A*, 147:278–292, 1984.

[16] 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 [41].

[17] 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.

[18] Bruno de Finetti. Notes de M. B. de Finetti sur le "Rapport général". In *Les mathématiques et le concret*, pages 232–241. Presses Universitaires de France, Paris, 1955.

[19] Joseph L. Doob. Probability as measure. *Annals of Mathematical Statistics*, 12:206–214, 1941. This article stems from a discussion with Richard von Mises at a meeting of the Institute of Mathematical Statistics in Hanover, New Hampshire, in September 1940.

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

[21] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, first edition, 1925. Thirteenth edition 1958.

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

[23] Maurice Fréchet and Maurice Halbwachs. *Le calcul des probabilités à la portée de tous*. Dunod, Paris, 1924.

[24] Peter Gács. Uniform test of algorithmic randomness over a general space. *Theoretical Computer Science*, 341:91–137, 2005.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning Theory: Data Mining, Inference, and Prediction*. Springer, New York, 2001.

[26] Sham Kakade and Dean 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.

[27] Andrei N. Kolmogorov. Sur la loi des grands nombres. *Atti della Reale Accademia Nazionale dei Lincei, Serie VI, Rendiconti*, 9:470–474, 1929.

[28] Andrei N. Kolmogorov. Sur la loi forte des grands nombres. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, 191:910–912, 1930b.

[29] Andrei N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. An English translation by Nathan Morrison appeared under the title *Foundations of the Theory of Probability* (Chelsea, New York) in 1950, with a second edition in 1956.

[30] Masayuki Kumon and Akimichi Takemura. On a simple strategy weakly forcing the strong law of large numbers in the bounded forecasting game. *Annals of the Institute of Statistical Mathematics*, 2007. Forthcoming, doi:10.1007/s10463-007-0125-5.

[31] Leonid A. Levin. On the notion of a random sequence. *Soviet Mathematics Doklady*, 14:1413–1416, 1973.

[32] Leonid A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17:337–340, 1976. The Russian original in: Доклады АН СССР 227(1), 1976.

[33] Leonid A. Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61:15–37, 1984.

[34] Paul Lévy. *Calcul des probabilités*. Gauthier-Villars, Paris, 1925.

[35] Thierry Martin. *Probabilités et critique philosophique selon Cournot*. Vrin, Paris, 1996.

[36] Thierry Martin. Cournot, philosophe des probabilités. In Thierry Martin, editor, *Actualité de Cournot*, pages 51–68. Vrin, Paris, 2005.

[37] Per Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

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

[39] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909.

[40] Jerzy Neyman. Indeterminism in science and new demands on statisticians. *Journal of the American Statistical Association*, 55:625–639, 1960.

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

[42] Siméon-Denis Poisson. *Recherches sur la probabilité des judgments en matière criminelle et en matière civile, précédés des règles générale du calcul des probabilités.* Bachelier, Paris, 1837.

[43] Yu. V. Prokhorov and B. A. Sevast'yanov. Probability theory. In *Encyclopaedia of mathematics: An updated and annotated translation of the Soviet "Mathematical Encyclopaedia" (managing editor M. Hazewinkel)*, volume 7, pages 307–313. Reidel, Boston, 1987–1994.

[44] Hans Richter. Zur Begründung der Wahrscheinlichkeitsrechnung. *Dialectica*, 8:48–77, 1954.

[45] Hans Richter. *Wahrscheinlichkeitstheorie*. Springer, Berlin, 1956.

[46] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41:1397–1409, 1970.

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

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

[49] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[50] Glenn Shafer. *The Art of Causal Conjecture*. The MIT Press, Cambridge, MA, 1996.

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

[52] Glenn Shafer and Vladimir Vovk. The origins and legacy of Kolmogorov's *Grundbegriffe*, Working Paper 4, www.probabilityandfinance.com, 2005. An abridged version, emphasizing the period before 1933, appeared as "The sources of Kolmogorov's *Grundbegriffe*" in *Statistical Science*, 26:70–98, 2006.

[53] Oscar Sheynin. *Aleksandr A. Chuprov: Life, Work, Correspondence. The making of mathematical statistics.* Vandenhoeck & Ruprecht, Göttingen, 1996.

[54] Ray Solomonoff. The universal distribution and machine learning. *The Computer Journal*, 46, 2003.

[55] Charles J. Stone. Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5:595–645, 1977.

[56] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1996.

[57] Jean-André Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939. This differs from Ville's dissertation, which was defended in March 1939, only in that a 17-page introductory chapter replaces the dissertation's one-page introduction. For a translation into English of the passages where Ville spells out an example of a sequence that satisfies von Mises's conditions but can still be beat by a gambler who varies the amount he bets, see www.probabilityandfinance.com/misc/ville1939.pdf.

[58] Jean-André Ville. Notice sur les travaux scientifiques de M. Jean Ville, 1955. Prepared by Ville when he was a candidate for a position at the University of Paris and archived by the Institut Henri Poincaré in Paris.

[59] Ladislaus von Bortkiewicz. Anwendungen der Wahrscheinlichkeitsrechnung auf Statistik. In *Encyklopädie der mathematischen Wissenschaften, Bd. I, Teil 2*, pages 821–851. Teubner, Leipzig, 1901.

[60] Guy von Hirsch. Sur un aspect paradoxal de la théorie des probabilités. *Dialetica*, 8:125–144, 1954.

[61] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919.

[62] Richard von Mises. *Wahrscheinlichkeitsrechnung, Statistik und Wahrheit.* Springer, Vienna, 1928. Second edition 1936, third 1951. A posthumous fourth edition, edited by his widow Hilda Geiringer, appeared in 1972. English editions, under the title *Probability, Statistics and Truth*, appeared in 1939 and 1957.

[63] Richard von Mises. *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*. Deuticke, Leipzig and Vienna, 1931.

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

[65] Vladimir Vovk. Predictions as statements and decisions, Working Paper 17, www.probabilityandfinance.com, June 2006.

[66] Vladimir Vovk. Continuous and randomized defensive forecasting: unified view, Working Paper 21, www.probabilityandfinance.com, August 2007.

[67] Vladimir Vovk. Defensive forecasting for optimal prediction with expert advice, Working Paper 20, www.probabilityandfinance.com, August 2007.

[68] Vladimir Vovk. Leading strategies in competitive on-line prediction, Working Paper 18, www.probabilityandfinance.com, August 2007.

[69] Vladimir Vovk. Competing with wild prediction rules, Working Paper 16, www.probabilityandfinance.com, December 2005 (revised January 2006).

[70] Vladimir Vovk. Competitive on-line learning with a convex loss function, Working Paper 14, www.probabilityandfinance.com, May 2005 (revised September 2005).

[71] Vladimir Vovk. Non-asymptotic calibration and resolution, Working Paper 13, www.probabilityandfinance.com, November 2004 (revised JUly 2006).

[72] Vladimir Vovk. On-line regression competitive with reproducing kernel Hilbert spaces, Working Paper 11, www.probabilityandfinance.com, November 2005 (revised January 2006.

[73] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[74] Vladimir Vovk, Ilia Nouretdinov, Akimichi Takemura, and Glenn Shafer. Defensive forecasting for linear protocols, Working Paper 10, www.probabilityandfinance.com, February 2005. An abridged version appeared in *Algorithmic Learning Theory: Proceedings of the 16th International Conference, ALT 2005, Singapore, October 8-11, 2005*, edited by Sanjay Jain, Simon Ulrich Hans, and Etsuji Tomita, www-alg.ist.hokudai.ac.jp/ thomas/ALT05/alt05.jhtml on pp. 459–473 of *Lecture Notes in Computer Science*, Volume 3734, Springer-Verlag, 2005.

[75] Vladimir Vovk and Glenn Shafer. A game-theoretic explanation of the $\sqrt{dt}$ effect, Working Paper 5, www.probabilityandfinance.com, 2003.

[76] Vladimir Vovk and Glenn Shafer. Good randomized sequential probability forecasting is always possible, Working Paper 7, www.probabilityandfinance.com, June 2003 (revised January 2005). An abridged version appeared in *Journal of the Royal Statistical Society, Series B*, 67:747–764, 2005.

[77] Vladimir Vovk, Akimichi Takemura, and Glenn Shafer. Defensive forecasting, Working Paper 8, www.probabilityandfinance.com, September 2004 (revised January 2005). An abridged version appeared in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, www.gatsby.ucl.ac.uk/aistats/.

Game-Theoretic Probability and Finance Project
Defensive Forecasting Subseries Working Papers

7. *Good randomized sequential probability forecasting is always possible*, by Vladimir Vovk and Glenn Shafer, June 2003 (revised August 2007).

8. *Defensive forecasting*, by Vladimir Vovk, Akimichi Takemura, and Glenn Shafer, September 2004 (revised January 2005).

9. *Experiments with the K29 algorithm*, by Vladimir Vovk, October 2004.

10. *Defensive forecasting for linear protocols*, by Vladimir Vovk, Ilia Nouretdinov, Akimichi Takemura, and Glenn Shafer, February 2005 (revised September 2005).

11. *On-line regression competitive with reproducing kernel Hilbert spaces*, by Vladimir Vovk, November 2005 (revised January 2006).

13. *Non-asymptotic calibration and resolution*, by Vladimir Vovk, November 2004 (revised July 2006).

14. *Competitive on-line learning with a convex loss function*, by Vladimir Vovk, May 2005 (revised September 2005).

16. *Competing with wild prediction rules*, by Vladimir Vovk, December 2005 (revised January 2006).

17. *Predictions as statements and decisions*, by Vladimir Vovk, June 2006.

18. *Leading strategies in competitive on-line prediction*, by Vladimir Vovk, August 2007.

20. *Defensive forecasting for optimal prediction with expert advice*, by Vladimir Vovk, August 2007.

21. *Continuous and randomized defensive forecasting: unified view*, by Vladimir Vovk, August 2007.

22. *Game-theoretic probability and its uses, especially defensive forecasting*, by Glenn Shafer, August 2007.