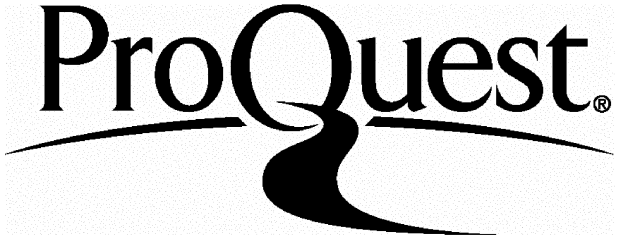
On Some Aspects of the Prequential and Algorithmic Approaches to Probability and Statistical Theory

Thesis submitted to the University of London for
the Degree of Doctor of Philosophy in the Faculty of Science

Marco Minozzo

Department of Statistical Science
University College London

February 1996



Abstract

Following an axiomatic introduction to the *prequential* (predictive sequential) principle to statistical inference proposed by A. P. Dawid, in which we consider some of the questions it raises, we examine a conjecture on the supposed prequential asymptotic behaviour of significance levels based on a particular class of test statistics.

Then, after a presentation of some martingale probability frameworks recently proposed by V. G. Vovk, algorithmic constraints are introduced to give a definition of *random sequences* on the lines of Martin-Löf's classical approach. This definition, instead of being given, as in the classical algorithmic approach, with respect to a Kolmogorovian probability distribution P, is given only with respect to a sequence of measurable functions by using the *principle of the excluded gambling strategy.* The idea underlying this approach is that if we are to play an infinite sequence of fair games against an infinitely rich bookmaker, then, whatever computable strategy we choose, we shall never become richer and richer as the game goes on.

These random sequences, apart from some basic properties, have been shown to satisfy: an analogue of Kolmogorov's strong law of large numbers; an analogue of the upper half of Kolmogorov's law of the iterated logarithm for binary martingales; and an analogue of Schatte's strong central limit theorem for the coin-tossing process. Besides, for these random sequences, we also investigated the distribution of the values of the corresponding infinite single realizations, in the case of two basic processes. These last results, together with the strong central limit theorem, would provide an instance in which 'empirical' distribution functions are derived without the assumption of any Kolmogorovian probability distribution.

Acknowledgements

I am most grateful to Professor A. Philip Dawid for his invaluable guidance during the preparation of this thesis.

I am also grateful to Doctor V. G. Vovk for his precious suggestions and for all his unpublished work I was allowed to use.

Contents

List of Figures

Chapter 1

Introduction

In a sentence, we could say that this thesis is all about one single idea, namely the idea of a sequential interpretation of the concept of probability. Our investigation will start by taking into consideration first, in the classical probability axiomatics of Kolmogorov, Dawid's prequential principle, then Vovk's martingale probability frameworks, and finally a new purely martingale definition of random sequences. Let us consider the problem of assessing the goodness of a sequence of probability forecasts $\mathbf{P}^n = (P_1, P_2, \ldots, P_n)$, for a sequence of random quantities $\mathbf{X}^n = (X_1, X_2, \ldots, X_n)$, in the light of a sequence of realized outcomes $\mathbf{x}^n = (x_1, x_2, \ldots, x_n)$. By the nature of the problem, it would seem sensible to ask for this assessment that it does not depend on the particular way the actual sequence of forecasts has been obtained. That is, it would seem sensible that two different forecasters, who happened to issue two identical sequences of probability forecasts, should receive the same assessment irrespectively of the way they generated those forecasts. On these grounds, Dawid (1984) put forward a principle to statistical inference, the *prequential* (predictive sequential) *principle*, suggesting that the evaluation of a probabilistic model, in the light of an actual sequence of outcomes, should be based only on the sequence of probability forecasts the model made, or would have made, for this sequence of outcomes, and not, for instance, on forecasts the model would have made for outcomes that never materialized. The prequential principle, alongside with some of the questions it addresses, is considered in

Chapter 2 where we also investigate, mainly by means of simulations, a conjecture, related to this principle, about the asymptotic behaviour of significance levels based on test statistics having the form

$$Y_n = \frac{\sum_{i=1}^{n}(Z_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}},$$

where Z_i is a function of $\mathbf{X}^i$, and $\mu_i = \mathrm{E}(Z_i|\mathbf{X}^{i-1})$, $\sigma_i^2 = \mathrm{Var}(Z_i|\mathbf{X}^{i-1})$, are the expectation and variance of Z_i under the conditional distribution $P_i = P(X_i|\mathbf{X}^{i-1})$. It turned out that this simulation work did not corroborate the conjecture in its generality. This, together with some other considerations, cast doubts on the possibility of finding any sequential interpretation of the concept of probability along the lines of the ideas of Seillier-Moiseiwitsch (1986), Dawid (1992), and Seillier-Moiseiwitsch and Dawid (1993).

In Chapter 3 some new frameworks for probability theory, alternative to Kolmogorov's axiomatics, and based on the primitive notion of a martingale, are considered. In the prequential probability framework, put forward by Vovk (1993a) on an idea of Dawid (1985), instead of specifying a full Kolmogorovian probability distribution P on a sample space (a σ-additive set function, normed to one, defined over a σ-algebra), we specify a probability forecasting system π on a tree-like structure giving one-step-ahead probabilities, given the past. Under some restrictions, this probability framework has been shown to be essentially equivalent to another axiomatics for probability theory, Shafer's event tree framework, which has been used by Shafer (1995) to provide a more appropriate framework for the study of the causal foundation of independence graphs. In the other probability framework considered, the purely martingale probability framework, which is due to Vovk (1993c), instead of specifying a probability distribution P, or a probability forecasting system π, we only specify a sequence of measurable functions which we call the basic martingale. Then interpretation, definitions and results are given by using only the *principle of the excluded gambling strategy.*

Then in Chapter 4, after having introduced some algorithmic constraints, we give a definition of *random sequences*, which we called M-typical sequences, in a

purely martingale framework on the lines of the now classical approach proposed by Martin-Löf (1966). This definition, instead of being given, as in the classical approach, with respect to a probability distribution P, is given only with respect to a sequence of measurable functions by using the principle of the excluded gambling strategy. The idea underlying this approach, and giving an interpretation to it, is that if we are to play an infinite sequence of fair games against an infinitely rich bookmaker, then, whatever computable strategy we chose, we will never become richer and richer as the game goes on.

In Chapter 5, these M-typical sequences are shown to satisfy an analogue of Kolmogorov's strong law of large numbers, and an analogue of the upper half of Kolmogorov's law of the iterated logarithm for binary martingales, whereas in Chapter 6 they are shown to satisfy an analogue of Schatte's strong central limit theorem for the coin-tossing process. In Chapter 6, we also investigate the distribution of the values, corresponding to a single M-typical sequence, in the case of two basic processes. These distributions of the values, together with the strong central limit theorem, represent an instance in which distributional properties are obtained without assuming any probability distribution or forecasting system whatsoever.

Chapter 7 and Appendix A contain respectively some general conclusions, and some basic algorithmic notions necessary to the definition of M-typical sequences.

Chapter 2

Dawid's Prequential Principle

2.1 Introduction

Statistical inference is the area of statistics which is concerned with the study of ways of inferring, inductively, on unknown quantities on the basis of observed data. In the tradition of the subjectivist and predictivist approaches, which have been originated by, among others, de Finetti (1937), a new approach to statistical inference, the *prequential* (predictive sequential) *approach*, has been proposed by Dawid (1984). In the light of a sequence of observations, this approach requires the assessment of any probabilistic model to be based only on the probability forecasts the model performed, or would have performed, for this particular sequence of outcomes. This new approach has also found inspirations in the somewhat more applied area of probability weather forecasting, where a great amount of work had already been carried out, independently from the main streams of research in statistics (see, for a review, Dawid, 1986), on the more practical aspects of model assessment.

So far, the preferred inferential techniques of the prequential approach, for the assessment of a probabilistic model, have been probability assessment techniques such as scoring rules and calibration plots (see, e. g., Seillier-Moiseiwitsch, 1986). Prequential inferential techniques have been applied in the areas of probabilistic weather forecasting, educational scaling, patient progress modelling (see Seillier-Moiseiwitsch, 1986), and in the evaluation of the performance of probabilistic expert

systems (see Spiegelhalter, Dawid, Lauritzen and Cowell, 1993).

In Section 2.2 we consider the prequential approach from an axiomatic point of view, namely through the prequential principle. Then in Section 2.3 another principle, the production principle, is considered which in some way tries to balance the requirements of the prequential principle.

By restricting the class of allowable probabilistic models, standard techniques such as the conditional probability integral transform or the standard martingale central limit theorem do lead to assessments which depend only on the realized sequence of outcomes and forecasts. Some of these standard techniques are considered from the point of view of the prequential principle in Section 2.4, whereas a conjecture on a class of test statistics is studied, mainly by means of simulations, in Section 2.5. It should be pointed out that, these techniques, or better their inferential conclusions, even if fulfilling the prequential principle from an 'axiomatic' point of view, do not have any 'within-sequence' interpretation. Indeed, they all have a repeated-sampling interpretation which refers to all possible replications of the experiment that did not materialize.

We conclude the chapter by considering in Section 2.6 a calibration criterion due to Dawid (1982, 1985).

2.2 The Prequential Principle

At the basis of the prequential approach to statistical inference lies the philosophy that it is reasonable, at least in many contexts, to regard Nature as producing, sequentially, an infinite data-string $\mathbf{x} = (x_1, x_2, \ldots)$, and that we should then regard $\mathbf{x}$ as containing all the relevant empirical evidence (past and future). If a probabilistic theory has been suggested for explaining the given data $\mathbf{x}$ as a realization of the sequence of random quantities $\mathbf{X} = (X_1, X_2, \ldots)$, then alternative data-strings which 'might' have been produced should be regarded as strictly theory-dependent and, in this view, they should not have any empirical content.

Let us consider the task of assessing the adequacy of a probabilistic theory for

the sequence of random quantities $\mathbf{X} = (X_1, X_2, \ldots)$, in the light of the sequence of realized outcomes $\mathbf{x}$. Following the prequential approach, we imagine the values arising sequentially. Just before we observe X_{i+1}, uncertainty about its value is measured by the predictive distribution P_{i+1} supplied by the theory.

Any probabilistic theory specifying a predictive distribution P_{i+1}, for all i and $\mathbf{x}^i = (x_1, x_2, \ldots, x_i)$, constitutes a *probability forecasting system* (PFS) for $\mathbf{X}$, and such a PFS essentially determines a joint probability distribution P for $\mathbf{X}$ (generally involving dependence), with the property of having the conditional distribution of X_{i+1} given $\mathbf{X}^i = \mathbf{x}^i$, that is, $P(X_{i+1}|X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i)$, equal to the predictive distribution P_{i+1}. On the other hand, any joint probability distribution P over $\mathbf{X}$, considered as a rule for generating a distribution P_{i+1} for X_{i+1}, for any values of i and $\mathbf{x}^i$, is a PFS (Dawid, 1984). So, the task of obtaining a global assessment of our theory at explaining $\mathbf{x}$ can equivalently be stated in terms of the assessment of the joint probability distribution P.

For this task, a predictive sequential methodology could involve the comparison of each observation x_{i+1} with the uncertainty assessed for it when it was about to be observed, that is with P_{i+1}, using probability assessment techniques such as scoring rules, calibration plots, etc. . More formally, the prequential approach demands that any method of assessing the success of a distribution P at describing the specific data $\mathbf{x}$ should depend only on the two sequences, of realized data-values and of realized forecasts,

$$\begin{array}{lccccccc} \mathbf{x}\text{:} & x_1 & x_2 & x_3 & \ldots & x_n & \ldots, \\ \mathbf{P}\text{:} & P_1 & P_2 & P_3 & \ldots & P_n & \ldots. \end{array}$$

This seemingly natural requirement has been proposed for the first time as an inferential principle by Dawid (1984) who called it the *prequential principle.* In some respect this principle is similar to the classical likelihood principle which, in the parametric case, requires the inference about a parameter to depend only on the observed likelihood. Note that both principles do not give any role to hypothetical outcomes that did not materialize.

Whereas the substantial motivations for the application of the prequential prin-

ciple rest on the above arguments, which have been put forward by Dawid, we will try here to clarify some of the questions it raises by taking a rather formal point of view. In general, statistical inference could be performed by any rule or pattern if no additional constraint would be imposed. The introduction and analysis of inferential principles and assumptions would then permit to characterize which inferential procedures are acceptable to us and which are not. This way of proceeding, even if not extremely common in statistical inference, is similar to other investigations to an axiomatic approach to a logic of statistical inference which have been carried out, among the others, by Birnbaum (1962) and Basu (1975) (see also Dawid, 1983).

Let us consider a class $\mathcal{P}$ of joint probability distributions P for the sequence of random quantities $\mathbf{X} = (X_1, X_2, \ldots)$, defined on the sample space $\mathcal{X} = (\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots)$.

Definition 2.2.1 *An inferential pattern I is a rule producing an inferential statement $I(\mathbf{x}^n, P)$ from data $\mathbf{x}^n$ and model P, for all $\mathbf{x}^n \in \mathcal{X}^n$, $n \in \mathbf{N}$ and for all $P \in \mathcal{P}$.*

(Note that here we do not impose on I any restriction, for instance, we do not impose that I has to be $\mathbf{R}$-valued.) From this more axiomatic point of view, the prequential principle can equivalently be translated as saying that we have to achieve the same inferential conclusions for every probabilistic model P which happened to make the same sequence of probabilistic forecasts $\mathbf{P}^n$ for the realized sequence of outcomes $\mathbf{x}^n$. This is asserted in the following definition.

Definition 2.2.2 *An inferential pattern I is prequential if it is a function only of $(\mathbf{x}^n, \mathbf{P}^n)$.*

So, to say that an inferential pattern I is prequential is the same as to say that I does respect the prequential principle. Note that a prequential inferential pattern I satisfies the first part of metacriterion M1 and metacriterion M2 of Dawid (1985).

As a result of considering only the two sequences $\mathbf{x}$ and $\mathbf{P}$, for any given outcome sequence $\mathbf{x}$ and joint probability distribution P, it makes sense also to consider

the set of all different distributions making the same sequence of forecasts $\mathbf{P}$ for $\mathbf{x}$ as does P. One particular choice is that distribution Q under which the (X_i) are independent, X_i having marginal distribution P_i. If we accept the prequential principle, then any test of the adequacy of P in the light of $\mathbf{x}^n$ (and in general any inferential statement $I(\mathbf{x}^n, P)$) should be identical with the test that we should conduct for Q in the light of $\mathbf{x}^n$ (should be identical with $I(\mathbf{x}^n, Q)$, respectively). Since this particular distribution Q will play a role in the following sections, we put it in the next definition.

Definition 2.2.3 *Given* $\mathbf{x}$ *and* P*, we call prequential independence model (p.i.m.) that model on* $\mathbf{X}$*,* Q *say, under which the* (X_i) *are independent and* X_i *has marginal distribution* $Q_i = P_i$*, for all* i*, where the* (P_i) *are the predictive distributions of* P *for* $\mathbf{x}$.

As in the case of the likelihood principle (which we distinguish from the 'likelihood approach'), the prequential principle does not say actually how a statistical inference has to be performed. It just says that any inferential pattern has to be a function of $(\mathbf{x}^n, \mathbf{P}^n)$. For the prequential principle, any inferential rule making a constant statement for all possible pairs $(\mathbf{x}^n, \mathbf{P}^n)$ would be acceptable as well as any arbitrary one-to-one function of $(\mathbf{x}^n, \mathbf{P}^n)$. In the case of a more common example, such as when we consider the practical task of inducing, given $\mathbf{x}^n$, an ordering on the set $\mathcal{P}$ of distributions for $\mathbf{X}$, it would be acceptable to use any function of $(\mathbf{x}^n, \mathbf{P}^n)$ with values in $\mathbf{R}$. A result of this is that, if we do not want to follow inferential patterns which are completely unacceptable from many reasonable points of view other then the prequential principle, we need to restrict the class of allowable inferential patterns by means of other principles or particular assumptions.

2.3 The Production Principle

In a discussion of the prequential principle, Dawid (1991) singled out an essential feature which should characterize at least those inferential patterns which take the

form of a probabilistic assertion. He started by considering an argument of Fisher. About Fisher (1956a), Dawid wrote:

> 'Discussing the Welch solution to *Behrens's problem* of testing the equality of two normal means when the variances are not supposed equal ... he showed that the nominal 10% test has rejection probability uniformly greater than 10.8%, conditional on the event that the two sample variances are equal. ... Calling such an event a "relevant subset" for the test, he put forward, as a requirement for the inferential validity of a test, that it should not admit any relevant subset.'

Later, Buehler (1959) applied this same logic to confidence intervals. After having named *production model* the model P that describes the (known or assumed) probability processes directly leading to the observation of data, Dawid continued:

> 'In other words, it is not good enough for a procedure to have correct overall probabilistic properties in the production model: if these are to have inferential relevance, it must not be possible to challenge them on the basis of specific data—which could be done by demonstrating their incorrectness conditional on some observable event.'

The production model, however, cannot be completely discharged, and Dawid (1991) argued as follows:

> 'Suppose that a proposed method of inference is couched as a probabilistic assertion, purported to have inferential validity conditional on the data. The form of this might be, for example, that the probability that the parameter θ is less than some statistic T is 95%. This could be considered as a confidence statement, a fiducial assertion or a Bayesian posterior probability—the interpretation is immaterial.
>
> In such a case we are surely entitled to require, as a *minimal* validity requirement on any inferential procedure, that its overall probabilistic properties, in the production experiment, are compatible with its purported inferential content. This would not be so in the above example if, for instance, the production model sampling probability that T exceeds θ is less than 93% for all θ, which would mean that the whole space

constitutes a relevant subset! We may term this (admittedly somewhat vague) validity requirement the *production principle*.'

Here, for the case in which an inferential pattern $I(\mathbf{x}^n, P)$ takes the form of a probabilistic assertion, we distinguish between a class of *weak* production principles and a unique *strong* production principle. Whereas a weak production principle would require, for an inferential procedure, that its probabilistic assertions are compatible, on an overall basis to be appropriately defined, with its probabilistic properties under the production model P, the strong production principle would require, for an inferential procedure, that its probabilistic assertions are exactly valid under the model P, in the sense that they can be considered as usual probabilities calculated under P. This strong production principle is the inspiration for the following definition.

Definition 2.3.1 *A probabilistic inferential pattern I is $\mathcal{P}$-valid if, for every $\mathbf{x}^n$, $n \in \mathbf{N}$, the probabilistic assertion $I(\mathbf{x}^n, P)$ holds under P for every $P \in \mathcal{P}$. If $\mathcal{P}$ is the set of all possible distributions on $\mathbf{X}$, we say that I is probabilistically valid.*

To require for I to be probabilistically valid means that its probabilistic assertions should have an unquestionable probabilistic meaning under the production model P. Without this requirement, a probabilistic inferential pattern I might generate probabilistic assertions with no validity at all under P.

2.4 A Prequential Pivotal Transformation

Dawid (1984) noted that if the (X_i) are continuous real variables we might consider test statistics based on the use of the well known 'conditional probability integral transform', that is, based on the quantities $U_i = F_i(X_i)$, F_i being the cumulative distribution function corresponding to the probability distribution P_i. In fact, Rosenblatt (1952) showed that, under any P (for continuous real variables), the (U_i) are independent and that each U_i is uniformly distributed in $[0, 1]$. Consequently any diagnostic test, of the adequacy of P in the light of $\mathbf{x}^n$, based on the

independent identical uniform distributions of the (U_i) will be prequential, since it is a function of the realized (u_i), which depend only on $(\mathbf{x}^n, \mathbf{P}^n)$, and will also be $\mathcal{P}$-valid, in the sense of the strong production principle, in the class $\mathcal{P}$ of all continuous distributions, since this distribution for the (U_i) holds under every $P \in \mathcal{P}$.

Here, we generalize this idea by considering a generic class $\mathcal{P}$ of joint probability distributions P for $\mathbf{X}$.

Lemma 2.4.1 *Let $\mathbf{W}^n = \Psi(\mathbf{X}^n)$, with Ψ depending just on $\mathbf{P}^n$, be a multivariate transformation such that $\mathbf{W}^n$ has a fixed joint probability distribution for all $P \in \mathcal{P}$. If I is an inferential pattern whose statements $I(\mathbf{x}^n, P)$ are expressed as a determinate probabilistic assertion in which $\mathbf{w}^n$ is considered as a realization from $\mathbf{W}^n$ under P, then I is prequential and $\mathcal{P}$-valid.*

Proof. Trivially, I is $\mathcal{P}$-valid by definition. Also, given $\mathbf{x}^n$, $\mathbf{w}^n$ is a function of $\mathbf{P}^n$, and having $\mathbf{W}^n$ a fixed distribution in $\mathcal{P}$, it follows that I is prequential. **Q.E.D.**

Observe that, if a model assuming independence is included in $\mathcal{P}$ and $W_i = \Psi_i(X_i)$, for all i, where Ψ_i depends only on P_i, then, under the fixed joint distribution of the lemma, the (W_i) have to be independent. An example of this particular case is readily supplied by taking the transformation Ψ to be the previous 'probability integral transform' of the predictive distributions P_i, and the class $\mathcal{P}$ to be the class of all continuous distributions on $\mathbf{X}$. Note that Ψ operates a proper reduction of the information contained in $(\mathbf{x}^n, \mathbf{P}^n)$. The use of $U_i = F_i(X_i)$ can yield, for a given $\mathbf{x}^n$, the same sequence $\mathbf{u}^n$ for two different sequences of predictions P_i, and so, an equal assessment for two models which had a different 'historical' behaviour. Another example of the previous lemma is provided by the standardized variables $W_i = (X_i - \nu_i)/\tau_i$ when $\mathcal{P}$ is, for instance, the class of distributions under which the (X_i) have conditional densities P_i of the form $\tau_i^{-1} g\{(x_i - \nu_i)/\tau_i\}$.

The transformation discussed here, unlike the sum test statistics considered in the next section, supplies test statistics which provide prequential and $\mathcal{P}$-valid significance levels for finite n. However, while these sum test statistics are applicable, in great generality, even for discrete random variables, we do not have here an

instance of the transformation Ψ for the class of all discrete distributions on $\mathbf{X}$.

2.5 A Prequential Inferential Model

Dawid (1991), together with the introduction of the production principle, made also another proposal. He started again from a Fisherian idea, namely the idea of an inferential frame of reference. To understand this idea, the following key abstraction is necessary: in any given inferential situation we can distinguish two models for the data. The first model, the *production model*, which has already been described in Section 2.3, is just the standard 'statistical model' available before experimentation. The second model, which Dawid (1991) called the *inferential model*, is supplied by the relevant frame of reference and it is to be used for analysis of the data at hand. Dawid (1991) writes:

> '... Fisher, even when discussing sampling-theoretic concepts such as significance levels, did not accept that this production model was necessarily the appropriate one in which to perform inferential probability calculations: for that purpose we need to discover the relevant frame of reference, which supplies what we may call the *inferential model* to be used for analysis of the data at hand. For example (Cox, 1958), if a coin has been tossed to decide which of two possible experiments to perform, then the randomness in the toss of the coin forms part of the production model; but once the coin has landed and the chosen experiment has been performed, it should be excluded from the inferential model, together with any consideration of hypothetical results that might have been obtained had the other experiment been chosen. Thus an appropriate frame of reference in this case might consist of all the possible results of the experiment that was actually chosen, and the inferential model would then be obtained from the production model by conditioning on the observed result of the coin-toss.'

He comments:

> 'Very generally a frame of reference may be regarded as specifying an inferential model for the data, or for some appropriate reduction of

the data such as the maximum likelihood estimator. An important feature of a frame of reference is that it should be tailored to the data that have been observed, and thus the inferential model will generally depend on those data. ...as conceived by Fisher, the relevant frame of reference need not necessarily be directly constructed from, or related to, the experiment as actually performed ... Consequently, even when the form of a Fisherian inference consists of "sampling probability" statements within the inferential model, it might not have any valid sampling-theoretic interpretation within the context of the production model.'

And then he stresses, in dealing with the likelihood estimator $\widehat{\theta}$ of a one-dimensional parameter θ, that:

'...Fisher (1925), as elsewhere, ...appears to be putting the view that (in large samples) it is appropriate, after observing the data, to use an inferential model which treats $\widehat{\theta}$ as normally distributed with mean θ and (data-dependent) inverse variance $\widehat{J}$ $[= -L''(\widehat{\theta})]$ ...We shall call this asymptotic inferential model for $\widehat{\theta}$ the *Fisher model.* ...Although a sampling model, it is entirely determined by the realized likelihood function based on the given data. Any inferential model with this property may be termed a "likelihood model".'

The Fisher model is thus entirely respecting the likelihood principle. Dawid (1991), analysing to what extent this model is in agreement with the production principle, verified that very generally it does satisfy this sampling criterion asymptotically. And he also found that a suitable extension of the Fisher model to non-regular problems, in which the asymptotic likelihood need not be of approximately normal form, does satisfy the production principle as well.

Turning to the prequential case, Dawid then argued that if we accept the prequential principle, then any test of the adequacy of P in the light of $\mathbf{x}$ should be identical with the test that we should conduct for Q (the prequential independence model) in the light of $\mathbf{x}$. And, to ensure that this will be so, he suggested to take Q, which he called the *prequential model*, as the appropriate inferential null distribution for constructing the test. Analogously to the case of the Fisher model,

the prequential model is entirely determined by the two sequences $(\mathbf{x}, \mathbf{P})$, and its inferential use does permit respect of the prequential principle. Of course, this prequential null distribution is not required at all to lead to sampling inferential statements with any sort of validity in the production model. It would be our duty to find to which extent inferential probabilities, calculated under Q, do satisfy the production principle, that is, are in some sort of correspondence with sampling probabilities in the production model.

An example in which we can certainly use the prequential model is given by test statistics based on the quantities $U_i = F_i(X_i)$, when the (X_i) are continuous random variables. In fact, under the inferential null distribution Q, the (X_i) are independent, and each Q_i is identical with the realized P_i. Hence, under Q, the (U_i) are independent, and being each U_i a 'probability integral transform' they will also be uniformly distributed. But, as we have already noted in Section 2.4, exactly this same joint distribution for the (U_i) will also hold under the more general production model P. Consequently any diagnostic test based on the inferential null distribution of the (U_i), which wrongly assumes that the (X_i) are independent with distribution Q, will nevertheless be exactly valid under the production null model.

In the remainder of the section we will consider a simulation study about a conjecture on the validity, under the production model, of the inferential null distribution Q for a class of test statistics.

2.5.1 A Conjecture

Let us consider the class of test statistics having the form

$$Y_n = \frac{\sum_{i=1}^{n}(Z_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}}, \tag{2.1}$$

where Z_i is a function of $\mathbf{X}^i$, and $\mu_i = \mathrm{E}(Z_i|\mathbf{X}^{i-1})$, $\sigma_i^2 = \mathrm{Var}(Z_i|\mathbf{X}^{i-1})$, are the expectation and variance of Z_i under the conditional distribution P_i. Seillier-Moiseiwitsch and Dawid (1993) noted that, since, under Q, the (Z_i) are independent and μ_i and σ_i^2 are fixed, it requires only weak additional conditions to ensure that the null inferential distribution of Y_n will be asymptotically $\mathcal{N}(0,1)$. And for such a test

statistic they also showed that (still under very weak conditions) this identical null asymptotic distribution for Y_n will continue to be valid under the production null hypothesis P.

Dawid (1991), by analogy with the results for the likelihood model of Fisher in non-regular problems, thought that even when the conditions under which Y_n has an asymptotic normal distribution under Q fail, use of the inferential distribution Q could still have some validity in the production model. By letting $\alpha(\mathbf{x}^n)$ be the observed significance level for Y_n calculated from data $\mathbf{x}^n$ and the corresponding forecasts $P_1, P_2, \ldots, P_n$, assuming the independence null distribution Q, he conjectured that, in great generality, as $n \to \infty$, the distribution of $\alpha(\mathbf{X}^n)$ should be asymptotically uniform over $[0, 1]$ under the production null hypothesis P. And he also said that, if this result were to fail in a given case, it would cast doubt on the validity of basing a test on Y_n, in either the production or the prequential model.

We will see that this conjecture is not supported by the following simulations, but let us first make a couple of remarks about it. First note that, for a given model P, since Q depends on the realized data-sequence $\mathbf{x}$, the asymptotic distribution of Y_n, under Q, has to be conceived, at least at first glance, varying with $\mathbf{x}$. Secondly, observe that the conjecture itself could be corroborated either strongly or weakly. It would be corroborated strongly if we would observe the same asymptotic distribution for Y_n, not necessarily normal, under both P and Q, or if we would observe the same asymptotic observed significance level, under both P and Q. Or, it would be corroborated weakly, on an 'overall' basis, even if the asymptotic observed significance level under Q does not equal that under P, if the distribution of $\alpha(\mathbf{X}^n)$ would be asymptotically uniform over $[0, 1]$ under the production null hypothesis P. In this case, however, the asymptotic distribution of Y_n under Q has necessarily to vary with $\mathbf{x}$, since the conjecture cannot be true for a fixed inferential distribution not valid under P.

2.5.2 Simulation Results

In the following examples, simulations have been carried out for investigating the validity, under the null hypothesis P, of using the inferential null distribution Q in the calculation of the asymptotic observed significance level of Y_n. Chosen a particular probabilistic model P for the sequence of random quantities $\mathbf{X}$, an histogram of the distribution of $\alpha(\mathbf{X}^n)$, the observed significance level of Y_n calculated under Q, has been obtained by generating a large number s of samples of fixed length n under P.

Example 2.1: The probabilistic model P has been taken to be a Gaussian autoregression $X_{i+1}|\mathbf{X}^i \sim \mathcal{N}(0, k^2 X_i^2)$, $X_1 \sim \mathcal{N}(0, k^2)$, where k is a constant, and $Z_i \equiv X_i$ has been considered in the test statistic Y_n. The observed significance level $\alpha(\mathbf{x}^n)$, under the independence null distribution Q, is given by

$$\alpha(\mathbf{x}^n) = Q(Y_n > y_n) = Q\left(\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} x_i\right) = 1 - \Phi(y_n),$$

where y_n is the value of Y_n for $\mathbf{x}^n$, and Φ is the standard normal distribution function. Various combinations of different values of k and n have been considered with $s = 10,000$. Figure 2.1 and 2.2 present the histogram of $\alpha(\mathbf{X}^n)$ respectively for the cases $k = 1$ with $n = 2,500$ and $k = 1.2$ with $n = 100$. In both cases, the simulated distributions cannot be considered uniform over $[0, 1]$ and so, at least in its greatest generality, the asymptotic conjecture cannot be corroborated. For very small and very high values of k it turned out that uniform histograms did actually appear. Figure 2.3 shows, for instance, the histogram of $\alpha(\mathbf{X}^n)$ for the case $k = 100$ with $n = 50$, and, apart from random variation, the asymptotic simulated distribution of $\alpha(\mathbf{X}^n)$ can be considered uniform over $[0, 1]$. This fact, however, relies on reasons that cannot be ascribed to the Dawid's conjecture. Indeed, it is just the consequence of the fact that the first and the last observations determine, respectively in the case with k very small and the case with k very high, the values of the statistics involved, so that under P we obtain, irrespectively of the value of n, an approximate standard normal distribution for Y_n. Note that, for the choice $Z_i \equiv X_i/(kX_{i-1})$, where $X_0 = 1$, Y_n is exactly standard normal for finite n, under

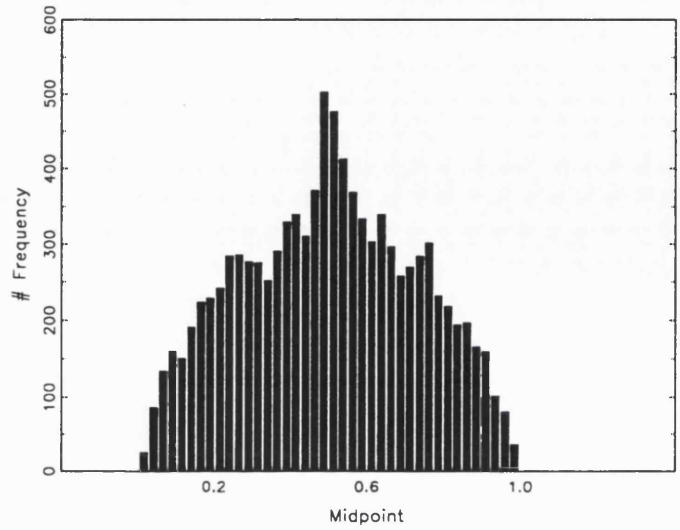


Figure 2.1: Histogram of $\alpha(\mathbf{X}^n)$ under the production model $X_{i+1}|\mathbf{X}^i \sim \mathcal{N}(0, k^2 X_i^2)$, $X_1 \sim \mathcal{N}(0, k^2)$, $k = 1$, with $s = 10,000$, $n = 2,500$, $Z_i \equiv X_i$.

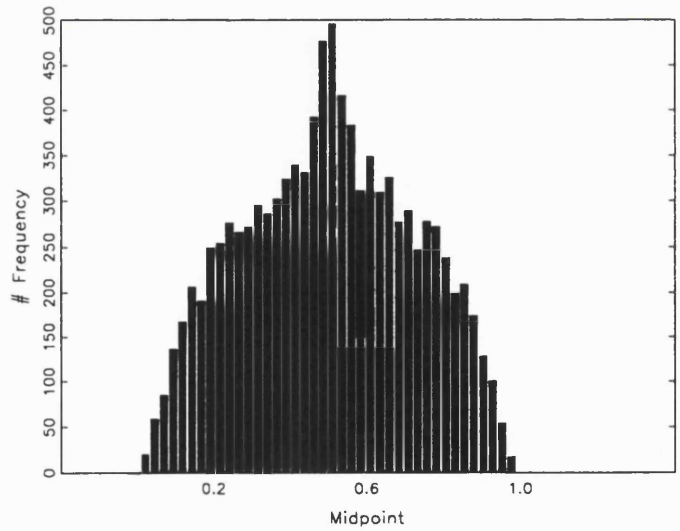


Figure 2.2: Histogram of $\alpha(\mathbf{X}^n)$ under the production model $X_{i+1}|\mathbf{X}^i \sim \mathcal{N}(0, k^2 X_i^2)$, $X_1 \sim \mathcal{N}(0, k^2)$, $k = 1.2$, with $s = 10,000$, $n = 100$, $Z_i \equiv X_i$.

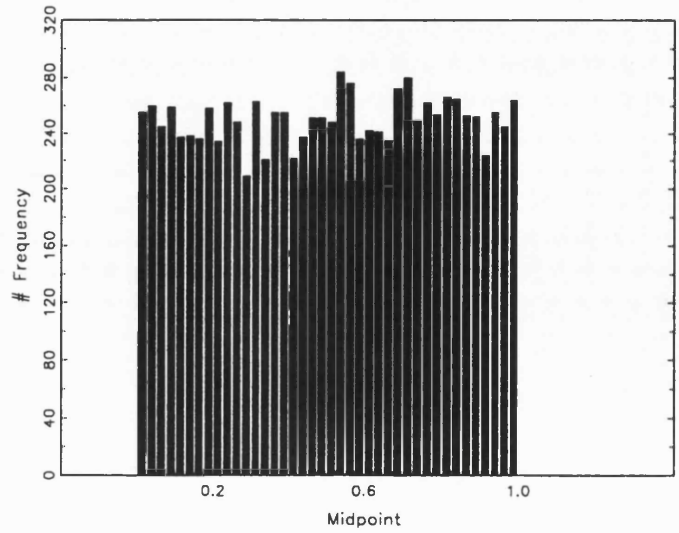


Figure 2.3: Histogram of $\alpha(\mathbf{X}^n)$ under the production model $X_{i+1}|\mathbf{X}^i \sim \mathcal{N}(0, k^2 X_i^2)$, $X_1 \sim \mathcal{N}(0, k^2)$, $k = 100$, with $s = 10,000$, $n = 50$, $Z_i \equiv X_i$.

both P and Q, and $\alpha(\mathbf{X}^n)$ is exactly uniformly distributed over $[0, 1]$.

In the next examples we report the simulated asymptotic distribution of Y_n, under both P and Q. As we have already noted, different asymptotic distributions of Y_n, under P and under Q, do not confute the conjecture, and this could still be corroborated strongly by equal asymptotic observed significance levels, or weakly on an overall basis. But, if the asymptotic distribution of Y_n under Q is independent of the data, then the above discrepancy is enough to confute the conjecture, that is, it implies that the asymptotic distribution of $\alpha(\mathbf{X}^n)$ is not uniform over $[0, 1]$.

Example 2.2: Take for the production model P the exponential autoregression $X_{i+1}|\mathbf{X}^i \sim \mathcal{E}(1/X_i)$, $X_1 \sim \mathcal{E}(1)$, with $\mathrm{E}(X_{i+1}|\mathbf{X}^i) = X_i$ and $\mathrm{Var}(X_{i+1}|\mathbf{X}^i) = X_i^2$, $(X_0 = 1)$, and set $Z_i \equiv X_i$ in the test statistic Y_n given by (2.1). The observed significance level $\alpha(\mathbf{x}^n)$, under the independence null distribution Q, can be calculated directly by convolution and it is given by

$$\alpha(\mathbf{x}^n) \;=\; Q\left(\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} x_i\right)$$

$$= (-1)^{n+1} \sum_{i=1}^{n} e^{-\frac{\sum_{i=1}^{n} x_i}{x_{i-1}}} \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x_{i-1}}{x_{j-1} - x_{i-1}}.$$

The simulated distribution of $\alpha(\mathbf{X}^n)$, for $n = 50$, was concentrated in the interval $[0.4, 1]$, whereas the simulated distributions of Y_n, for $s = 10,000$ and $n = 50$, under Q, were independent of the data (obtained under P) and highly different from the simulated distribution of Y_n, under the null model P. As in the previous example, we note that if $Z_i \equiv (X_i - X_{i-1})/X_{i-1}$, then Y_n has the same identical distribution, under both P and Q, for every n, and $\alpha(\mathbf{X}^n)$ has an exact uniform distribution over $[0, 1]$.

Example 2.3: Assume, for the model P, the uniform autoregressive model $X_{i+1}|\mathbf{X}^i \sim \mathcal{U}(0, bX_i)$, $X_1 \sim \mathcal{U}(0, b)$, $(X_0 = 1)$, where b is a positive parameter, and take $Z_i \equiv X_i$ in the test statistic Y_n given by (2.1). The calculation of the observed significance level $\alpha(\mathbf{x}^n)$, under the independence distribution Q, can be performed using the Laplace transform, yielding

$$\begin{aligned} \alpha(\mathbf{x}^n) &= Q\left(\sum_{i=1}^{n} X_i > \sum_{i=1}^{n} x_i\right) \\ &= 1 - \frac{1}{n! \prod_{i=1}^{n} b_i} \left(w_+^n - \sum_{i=1}^{n} (w - b_i)_+^n + \sum_{i<j}^{n} (w - (b_i + b_j))_+^n \right. \\ &\qquad \left. - \sum_{i<j<n}^{n} (w - (b_i + b_j + b_k))_+^n + \cdots + (-1)^n \left(w - \sum_{i=1}^{n} b_i\right)_+^n \right), \end{aligned}$$

where $b_i = bx_{i-1}$, $w = \sum_{i=1}^{n} x_i$ and a_+^n stands for $(a_+)^n$ with $a_+ = (a + |a|)/2$ for any real number a. The simulations of the distribution of $\alpha(\mathbf{X}^n)$, carried out only for very small values of n, did not result in a uniform distribution over $[0, 1]$. Moreover, the simulated distribution of Y_n, for $s = 10,000$ and $n = 50$, under Q, was independent of the data and completely different from that obtained under P. For the choice $Z_i \equiv (X_i - \nu_i)/\tau_i$, where $\nu_i = bX_{i-1}/2$ and $\tau_i^2 = b^2 X_{i-1}^2/12$, it is easy to check that Y_n has the same distribution, under both P and Q, for finite n, and that $\alpha(\mathbf{X}^n)$ has, still for finite n, a uniform distribution over $[0, 1]$.

Example 2.4: In this example the model P is supplied by a non-central chi-

squared law in which the rule of dependence is given by $X_{i+1}|\mathbf{X}^i \sim \chi^2_1(X_i)$, $X_1 \sim \chi^2_1$. For the choice $Z_i \equiv X_i$ in (2.1), we get $\mu_i = 1 + X_{i-1}$ and $\sigma_i^2 = 2 + 4X_{i-1}$, where $X_0 = 0$. The simulated distribution of the observed significance level $\alpha(\mathbf{X}^n)$ was not uniform over $[0, 1]$. Figure 2.4 gives the histogram of the simulated distribution of Y_n, under P, with $s = 10,000$ and $n = 100$. For the same values of s and n,

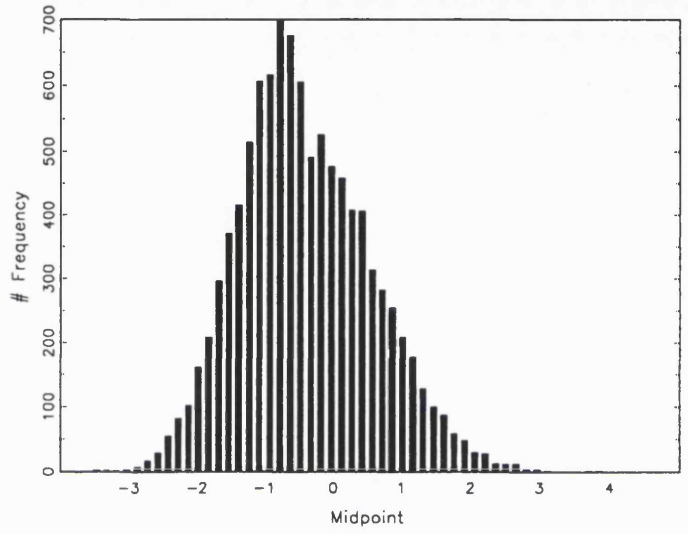


Figure 2.4: Histogram of Y_n under the production model $X_{i+1}|\mathbf{X}^i \sim \chi^2_1(X_i)$, $X_1 \sim \chi^2_1$, with $s = 10,000$, $n = 100$, $Z_i \equiv X_i$.

Figure 2.5 gives instead an histogram of the simulated distribution of Y_n, under Q, which is an asymptotic standard normal, for every data-sequence from P. As before, the two histograms do not provide any support to the conjecture. For the choice $Z_i \equiv (X_i - 1 - X_{i-1})/(2 + 4X_{i-1})^{1/2}$, we get that Y_n is just a sum of standardized random variables, under both P and Q, and so that Y_n has an asymptotic standard normal distribution, again under both P and Q.

The next and last example provides an instance in which, unlike the previous ones, the asymptotic distribution of the observed significance level $\alpha(\mathbf{X}^n)$ is uniform over $[0, 1]$ for the choice $Z_i \equiv X_i$.

Example 2.5: Consider the binary Markov Chain model with $\theta_1 = P_i(X_i = 1|X_{i-1} = 0)$, $\theta_2 = P_i(X_i = 1|X_{i-1} = 1)$, where $P_1(X_1 = 1) = \theta_1$. Taking $Z_i \equiv X_i$ in

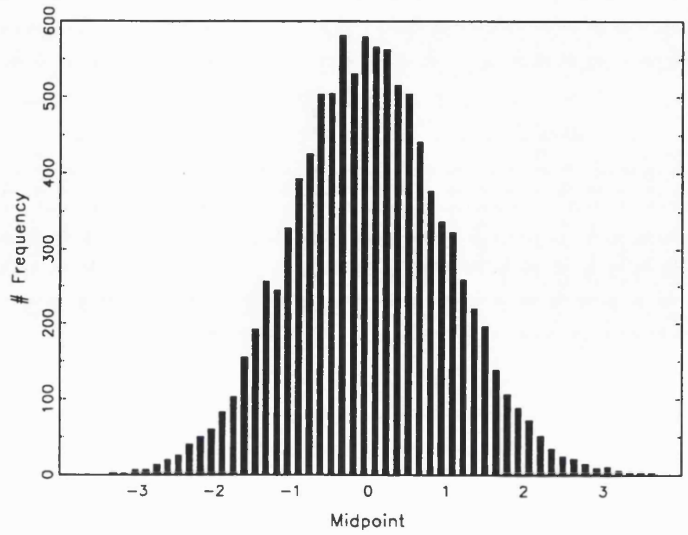


Figure 2.5: Histogram of Y_n under the prequential model Q for the production model $X_{i+1}|\mathbf{X}^i \sim \chi^2_1(X_i)$, $X_1 \sim \chi^2_1$, with $s = 10,000$, $n = 100$, $Z_i \equiv X_i$.

(2.1), we have $\mu_i = \theta_1(1 - X_{i-1}) + \theta_2 X_{i-1}$ and $\sigma_i^2 = \theta_1(1-\theta_1)(1 - X_{i-1}) + \theta_2(1 - \theta_2)X_{i-1}$, $(X_0 = 0)$. With some algebra, we can see that the observed significance level $\alpha(\mathbf{x}^n)$, under the independence distribution Q, is given by

$$\begin{aligned}\alpha(\mathbf{x}^n) &= Q\left(\sum_{i=1}^n X_i > \sum_{i=1}^n x_i\right) \\ &= \sum_{k>n_1}\left[\sum_{\substack{i+j=k \\ i\le n_0+x_n \\ j\le n_1-x_n}} \binom{n_0+x_n}{i}\theta_1^i(1-\theta_1)^{n_0+x_n-i}\binom{n_1-x_n}{j}\theta_2^j(1-\theta_2)^{n_1-x_n-j}\right],\end{aligned}$$

where $n_1 = \sum_{i=1}^n x_i$ and $n_0 = n - n_1$. Figure 2.6 shows the simulated cumulative distributions of $\alpha(\mathbf{X}^n)$, when $\theta_1 = 0.4$ and $\theta_2 = 0.7$, for $n = 10$ and $n = 50$ (the closest to the diagonal) respectively. In this case, Y_n has an asymptotic standard normal distribution under P, as well as under Q, which implies that the asymptotic distribution of the observed significance level $\alpha(\mathbf{X}^n)$ is uniform over $[0, 1]$.

The first conclusion we can draw from these simulations is that, in its generality, the original conjecture has not been corroborated. It would seem, quite strongly, that the asymptotic use of the test statistic Y_n does not guarantee to respect both the prequential principle and the production principle, even on an overall 'α-level'

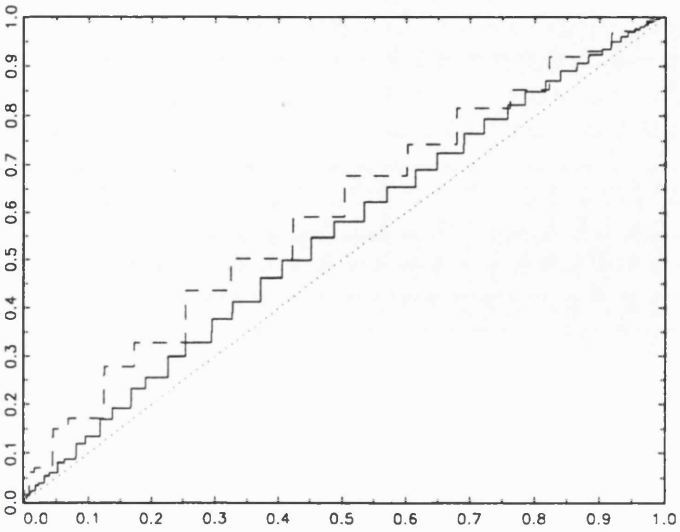


Figure 2.6: Distributions of $\alpha(\mathbf{X}^n)$ under the binary Markov Chain production model with $\theta_1 = 0.4$, $\theta_2 = 0.7$, and $s = 5,000$, $n = 10$ ($--$), $n = 50$ (——), $Z_i \equiv X_i$.

basis. From this point of view, these simulations would better sustain the form of Y_n in which the Z_i are standardized quantities.

Besides, the rejection of the conjecture seems to cast doubts also on the general inferential relevance of the prequential independence model Q itself. In this case further investigations about the prequential model Q, for instance in the direction of possible analogues to the concepts of sufficiency and ancillarity (in the classical sense of, e. g., Fisher (1956b)) when the production model is restricted to an element of a parametric family, would seem of little interest.

2.6 Dawid's Calibration Criterion

In the previous sections we considered the problem of assessing the 'goodness' of a probabilistic model in the light of a sequence of outcomes, so as to respect the prequential principle. That is, in such a way that the assessment depends only on the realized sequence of probability forecasts the model would have performed for a given sequence of outcomes. A general rule satisfying this requirement has been proposed by Dawid (1982, 1985) with his calibration criterion.

Following Dawid (1986), an intuitive introduction to these ideas can be given by considering the practical task of assessing the goodness of a sequence of probability forecasts. Dawid writes:

> 'At its simplest, probability forecasting refers to the process of attaching a numerical probability to an uncertain event. ... A single non-categorical probability forecast (i. e., not 0 or 1) can never be said to have been "right" or "wrong". But when a forecaster has issued a long string of such forecasts, it becomes possible to apply checks of external validity.
>
> Suppose that of n sequential forecasts, the ith is p_i and the realized outcome of the associated event A_i ("rain on day i") is a_i (= 0 or 1). Then we can compare the overall average forecast probability $\overline{p}_n = n^{-1}\sum_{i=1}^{n} p_i$ with the overall relative frequency of occurrence $\overline{a}_n = n^{-1}\sum_{i=1}^{n} a_i$. If $\overline{p}_n \simeq \overline{a}_n$, the set of forecasts may be regarded as approximately valid on an overall basis. ...
>
> A more incisive test looks at that subset of occasions i for which the forecast probability p_i was at, or suitably close to, some preassigned value p^*, and compares the observed relative frequency in this subset, $\overline{a}(p^*)$ say, with p^*. If $\overline{a}(p^*) \simeq p^*$ for all p^*, the forecasts have been variously termed "unbiased in the small", "reliable", "valid", or "well calibrated".'

Elaborating this intuitive idea, Dawid (1982) introduced an extended calibration criterion. He considered an arbitrary subsequence of $(1, 2, \ldots, n)$ subject to the constraint that the decision on whether or not i is to be included in the subsequence should be determined only by the previous outcomes $(a_1, a_2, \ldots, a_{i-1})$, and not by a_i or any later outcomes. Then, indicating with $\overline{p}'_n$ the average forecast probability, and with $\overline{a}'_n$ the empirical relative frequency, for the events in this subsequence, he required as a validity criterion that $\overline{p}'_n \simeq \overline{a}'_n$.

Dawid (1982) justified this by showing that if the (p_i) are constructed sequentially as appropriate conditional probabilities from a joint distribution P for the events $(A_1, A_2, \ldots)$, so that $p_i = P(A_i | a_1, a_2, \ldots, a_{i-1})$, then $\overline{a}'_n - \overline{p}'_n \to 0$, as $n \to \infty$, with probability one, under P, provided that the cardinality of the subsequence goes

to infinity. He observed that if $\overline{a}'_n$ and $\overline{p}'_n$ are not close enough together, there is a suggestion that, in fact, $\overline{a}'_n - \overline{p}'_n$ does not tend to zero, and this would serve to discredit the probability assignments made by P.

Elaborating further these ideas, Dawid (1985), in an algorithmic probability setting, calling *completely calibrated* a sequence of forecasts which satisfies the calibration criterion for every 'admissible' subsequence, showed that if $(p_i^{(1)})$ and $(p_i^{(2)})$ are each completely calibrated computable forecast sequences, for a given infinite sequence of outcomes, then $p_i^{(1)} - p_i^{(2)} \to 0$, as $i \to \infty$. And, since this criterion of complete calibration is strong enough to exclude all but one limiting assignment of probabilities, he argued in favour of the existence of what he called calibration-based empirical probabilities, in an attempt also to provide a probabilistic foundation to the prequential principle.

2.7 Discussion

In this chapter we have presented some of the problems which arose from the seemingly natural requirement of respecting the prequential principle. Related problems had already been highlighted. With regard to the use of test statistics Y_n of the form (2.1), Seillier-Moiseiwitsch, Sweeting and Dawid (1992), in the case of tests of composite null hypotheses, showed, in some specific cases, that the asymptotic distribution of Y_n is standard normal, thus providing, in those circumstances, a straightforward test for the validity of the statistical model. However, they were not able to prove the above result under more general conditions and put forward, instead, only some heuristic arguments. Besides, in facing the problem of assessing the goodness of a probabilistic model for a given data-sequence, respecting the prequential principle, Seillier-Moiseiwitsch and Dawid (1993) said that the asymptotic distribution of the test statistics considered had been studied imposing conditions on the null distribution P. They argued that it would had been more satisfactory, in the derivation of the asymptotic standard normal distribution of Y_n, to impose conditions only on the actual realized sequence of probability forecasts, thus tak-

ing the prequential principle more seriously. However, at the present moment, a solution to this problem is still to come, and there are reasons to believe that the problem itself might have to be reconsidered.

From a standard algorithmic point of view, the complete calibration criterion of Dawid (1985) could be seen, as noted by Dawid himself, as an attempt to give a definition of randomness for non-Bernoulli distributions by means of the frequency approach, as initially conceived by von Mises (1951), with his idea of a 'collective'. However, such an attempt has in itself some fundamental differences which make it difficult to consider it simply as a generalization of von Mises's approach. Indeed, Uspenskii, Semenov and Shen' (1990) remark, still from a standard algorithmic point of view, that if we want to give a definition of randomness on the lines of von Mises frequency approach, it is essential that the distribution considered is a Bernoulli one.

On the issue of the relationship with other approaches to statistical inference, Dawid (1992) studied the similarities between the prequential approach and the rather new approach to statistical inference called stochastic complexity, which is based on the connections between probability distributions and coding systems. The underlying philosophy of this approach (see, e. g., Rissanen, 1989) considers a transmission problem in which a sender, who observes a (finite) data-string, wishes to transmit this, by means of a coded message, to a receiver; and the success of a coding system is measured by the shortness of its code for the observed data-string. Dawid (1992) showed that the empirical assessment of a model based on the minimal length of a coded message, needed to transmit the data, is essentially equivalent to the prequential assessment based on the 'logarithmic scoring rule'.

Chapter 3

Martingale Probability Frameworks

3.1 Introduction

Vovk (1993a), inspired by the ideas of Dawid (1984) and Dawid (1985, Section 13.2), but extending them in a different direction, put forward a probability framework, alternative to that of Kolmogorov, based on the idea of a probability forecasting system. In this framework, which makes use of a 'martingale calculus', Vovk (1990a, 1990b, 1991) showed versions of the weak and strong law of large number, the central limit theorem, and the law of the iterated logarithm. We present this framework in Section 3.2. Let us note here that the difference between Vovk's calculus, which, in other words, was motivated by the idea of building a probability framework starting from the sequences of conditional probabilities given the past (not just the sequence of the actually observed conditional probabilities for the actual realized sequence of outcomes, as in the prequential principle), and Kolmogorov's probability framework is mainly foundational, and the same applications can equivalently be treated in both frameworks.

Before going further, we have here to make a note about the different ideas behind the world 'prequential'. With respect to Dawid's prequential principle, this word has a meaning mostly in terms of the two sequences of actual probability

forecasts and of actual realized outcomes. On the other hand, when related to Vovk's prequential probability framework, this term acquires a somewhat wider meaning referring to all possible sequences of probability forecasts and outcomes, and not just to the realized ones. Moreover, apart from this, the word 'prequential' may also be related to results or statements, in whichever probability framework, which are based on local properties, that is, on properties which take account only of conditional probabilities given the past.

In Section 3.3, we present an axiomatics of probability theory, due to Shafer (1985, 1993), which, even if it originated from quite different considerations, is strongly related to Vovk's prequential probability framework. This axiomatics, which will be formally connected to Vovk's prequential framework in Section 3.4, has been used by Shafer to provide a framework for the study of the causal foundation of independence graphs. Whereas in the previous chapter the prequential principle had been considered in Kolmogorov's probability axiomatics, in Section 3.5 it will briefly be considered in Vovk's prequential probability framework. Very recently, another foundation for probability theory, based entirely on the primitive notion of a martingale, has been proposed by Vovk (1993b, 1993c, 1995a). In this foundation, neither a probability distribution nor a probability forecasting system are introduced. An instance of this purely martingale probability framework is considered in Section 3.6.

3.2 Vovk's Prequential Probability Framework

In this section we present the prequential probability framework of Vovk (1993a). Let $\mathbf{N}$ denote the set of positive integers $1, 2, \ldots$, and $\mathbf{R}$ the set of real numbers. We introduce the following notation. Let us fix an *observation space* Ω, with generic element ω, endowed with a σ-algebra. Ω^* is the set of all finite sequences $x = \omega_1\omega_2\ldots\omega_n$ of elements of Ω; Ω^* includes the empty sequence $\Box$. If $x \in \Omega^*$, then $|x|$ denotes the length of the sequence x, and, for $\omega \in \Omega$, $x * \omega$ denotes the sequence obtained from x by adding ω on the right-hand side. The set of infinite

sequences $\omega_1\omega_2\ldots$ of elements of Ω is denoted by Ω^∞. A *forecasting system* (over the observation space Ω) is defined as a function $\pi\colon D \to \Pi$, where D is an arbitrary measurable set in Ω^* and Π is the set of probability distributions in Ω. Sequences x in D are called π-prior. This forecasting system π represents for us a probabilistic theory about the world that, for some given sequences of past data x, is able to make probability forecasts for the results of future observations ω. Note that, π is not required to make predictions for every past sequence $x \in \Omega^*$, but just for those sequences $x \in D$. When $D = \Omega^*$, the forecasting system π is *total*, that is, it makes predictions for all $x \in \Omega^*$. We denote by $\pi(x, A)$, where $x \in D$ and $A \subseteq \Omega$, the probability $\pi(x)(A)$ of the set A for the observed sequence x.

For a fixed forecasting system π, let us give the basic elements of this framework. These are the measurable functions, called *non-negative π-supermartingales*, $S\colon \Omega^* \to [0, \infty]$ such that

$$S(x) \geq \int_\Omega S(x * \omega)\pi(x, \mathrm{d}\omega),$$

for any π-prior x, and $S(x) = S(x * \omega)$, for all ω, for any x which is not π-prior. A non-negative π-supermartingale S is interpreted as a gambling strategy against an infinitely rich holder of the theory π.

With these ingredients, Vovk (1993a) gave his martingale calculus of probability. He gave a definition of 'global probability' of an *event* E, which is a measurable set in Ω^∞, in terms of the 'local probabilities' $\pi(x)$. We follow here the second of his two equivalent formulations. Being $\xi = \omega_1\omega_2\ldots$ an infinite sequence of elements of Ω and ξ^n, $n \in \mathbf{N}$, the initial segment $\omega_1\omega_2\ldots\omega_n$ of ξ of length n, we say that a non-negative π-supermartingale S is *successful* on an event E if, for all $\xi \in E$, $\sup_{n\in\mathbf{N}}\{S(\xi^n)\} > 1$. The *global π-probability* of E is then defined as

$$\mathrm{pr}_\pi(E) = \inf\{S(\Box)\},$$

where S ranges over the non-negative π-supermartingales which are successful on E.

If π is non-total, global π-probability is not additive, and even if π is total, it is not necessarily σ-additive. However, global π-probability is subadditive. That is,

for any forecasting system π and any sequence, finite or infinite, of events $E_1, E_2, \ldots$,

$$\mathrm{pr}_\pi\Big(\bigcup_k E_k\Big) \leq \sum_k \mathrm{pr}_\pi(E_k).$$

Global π-probability is in agreement with Kolmogorov's calculus of probability. Let us consider a probability distribution P in Ω^∞. With P we can associate total forecasting systems π such that $\pi(\omega_1\omega_2\ldots\omega_n)(A)$ is a variant of the conditional probability, relative to P, that the $(n+1)$th observation will lie in A given that the first n observations have been $\omega_1\omega_2\ldots\omega_n$. All such π are in some sense equivalent, and any one of them is said to be generated by P. Vovk (1993a, Theorem 4) proved that, if Ω is a Borel space and P is a probability distribution in Ω^∞ generating a total forecasting system π, then $\mathrm{pr}_\pi = P$.

3.3 Shafer's Protocols and Event Trees

We consider here the basics of a mathematical structure for probability theory which has been proposed by Shafer. In a discussion on the use of conditional probability, Shafer (1985) convincingly argued that its use is fully justified only in the presence of a *protocol*, that is, of a set of rules that tell, at each step, what can happen next. He gave the following definition.

Definition 3.3.1 *A protocol is a non-empty collection $\mathcal{S}$ of subsets of a non-empty set Υ such that*

(i) any two elements of $\mathcal{S}$ are either disjoint or nested, and

(ii) if $v \in \Upsilon$, $S \in \mathcal{S}$, and $v \notin S$, then there is an element of $\mathcal{S}$ that contains v and is disjoint from S.

Note that, a protocol is essentially a tree-like structure added to a sample space Υ. The elements of $\mathcal{S}$ are called *situations*. If E is an event in the sample space Υ and S is a situation, then every situation S corresponds to an event E, but most events do not correspond to a situation. Trees whose root is Υ and whose nodes are elements of $\mathcal{S}$ are called by Shafer *event trees*. The main point of Shafer (1985) was

to show that the classical probabilistic rule of conditioning on an event is justified only if the conditioning event is a situation. With a protocol, he said, the rule of conditioning can be treated as a theorem; otherwise its use is questionable.

On this tree-like structure, Shafer (1985) considered a set of axioms for probability, and proved, to show the richness of the framework, a version of Bernoulli law of large numbers. Here, we will present his axiomatics as presented in Shafer (1993). When an event E contains a situation S, we say that E is *certain* in S. When E is disjoint from S, we say that E is *impossible* in S. When E is impossible or certain in S, E is *determinate* in S. Otherwise, it is *indeterminate* in S. We say that an event E *precedes* an event F if E is determinate in any situation in which F is determinate. And we say that E and F are *incompatible* in S if the intersection of the three events is empty.

Then, by letting E and F be events, and S, T and U be situations, Shafer encapsulated in the following axioms the properties of the probabilities $P_S(E)$ of event E in situation S.

S1. $0 \leq P_S(E) \leq 1$ *for every situation* S *and every event* E.

S2. $P_S(E) = 0$ *if and only if* E *is impossible in* S.

S3. $P_S(E) = 1$ *if and only if* E *is certain in* S.

S4. *If* E *and* F *are incompatible in* S, *then* $P_S(E \cup F) = P_S(E) + P_S(F)$.

S5. *If* S *precedes* T *and* T *precedes* U, *then* $P_S(U) = P_S(T)P_T(U)$.

Axioms S1–S5 are satisfied whenever probabilities in situations are taken to be conditional probabilities with respect to some overall probability measure on the sample space. Conversely, if the numbers $P_S(E)$ satisfy axioms S1–S5, then for each situation S, the mapping that assigns the number $P_S(E)$ to each subset E of Υ is a probability measure on Υ.

Without S5, the remaining axioms would specify just a set of unlinked probability measures on the sample space, one for each situation. S5 supplies the connection

among these probability measures by means of which any set of probabilities in situations satisfying S1–S5 can be regarded as a set of conditional probabilities with respect to an essentially unique overall probability measure on the sample space.

Though axioms S1–S5 are in agreement with Kolmogorov's axiomatics, Shafer maintained the view that it would be better to abandon it for the event tree framework. Event trees, he says (see Shafer, 1990, 1991, 1993), provide the best framework for the philosophical study of probability, and in Shafer (1995) they are used to provide a probabilistic foundation to the study of probabilistic causation.

3.4 A Basic Prequential Framework

The two probability frameworks we have just presented in Section 3.2 and 3.3 share two basic fundamental features. They both are based on a tree-like structure, and conceived in terms of local probabilities by means of which it is then possible to specify a global probability on the sample space. To the end of this section we will see how under some restrictions they happen to integrate.

For a fixed set Ω, let us consider the infinite product space $\Omega^\infty = \Omega \times \Omega \times \cdots$. This is the set of all infinite data-sequences $\xi = \omega_1\omega_2\ldots$. The set Ω is called the observation space, Ω^∞ is called the sample space, and, to emphasize the tree-like structure of this space, $\Omega \times \Omega \times \cdots$ is called the event tree. This structure, in which finite data-sequences and situations coincide, will be the basis of what we will call the *basic prequential framework.*

Building on the idea of a 'sub-forecasting system' (Dawid, 1993), we introduce the following notation. For $x \in \Omega^*$ and $\omega \in \Omega$, $x * \omega$ denotes the sequence obtained from x by adding ω on the right-hand side. In the same way, for $A \subseteq \Omega$, $y \in \Omega^*$ and $E \subseteq \Omega^\infty$, the quantities $x * A$, $x * y$, $x * E$, etc., are similarly defined. For instance, $x * A$ is the set of sequences of form $x * \omega$ with $\omega \in A$. For every sequence $x \in \Omega^*$, the *cylinder* set $\Gamma_x \subseteq \Omega^\infty$ is defined by

$$\Gamma_x = \{\xi : \xi^{|x|} = x\}.$$

When $x, y \in \Omega^*$, $x \subseteq y$ means that x is a *prefix* of y. Of course, $x \subseteq y$ implies

$\Gamma_y \subseteq \Gamma_x$. If π is a forecasting system with domain $D \subseteq \Omega^*$, then for $x \in \Omega^*$, we denote by π_x the forecasting system, with domain $D_x = \{y \in \Omega^* : x*y \in D\}$, which for $y \in D_x$ and $A \subseteq \Omega$ satisfies $\pi_x(y, A) = \pi(x * y, A)$. The global probabilities associated with π and π_x are denoted with pr and pr_x respectively. And for every event $E \subseteq \Omega^\infty$, we denote by E_x the event in Ω^∞ given by

$$E_x = \{\xi : x * \xi \in E, \xi \in \Omega^\infty\},$$

for which $x * E_x = E \cap \Gamma_x$. With these definitions, we are now in a position to work with the probabilities $\mathrm{pr}_x(E_x)$ which can be thought of as the basic elements of our framework.

If π is a total forecasting system, generated by a probability distribution P in Ω^∞, then $\mathrm{pr}(E) = P(E)$ and $\mathrm{pr}_x(E_x) = P(x * E_x | \Gamma_x) = P(E|\Gamma_x)$. A basic property enjoyed by the probabilities $\mathrm{pr}_x(E_x)$, for a generic π and for $x, y, z \in \Omega^*$, is

$$\mathrm{pr}_x(\Gamma_{y*z}) = \mathrm{pr}_x(\Gamma_y)\mathrm{pr}_{x*y}(\Gamma_z).$$

Similarly, for $x, y \in \Omega^*$ and $E \subseteq \Omega^\infty$, we also have the formula

$$\mathrm{pr}_x(E_x \cap \Gamma_y) = \mathrm{pr}_x(\Gamma_y)\mathrm{pr}_{x*y}(E_{x*y}).$$

Another interesting property relating the probabilities $\mathrm{pr}_x(E_x)$ is given by the recursive formulae

$$\mathrm{pr}_x(E_x) = \begin{cases} \int \mathrm{pr}_{x*\omega}(E_{x*\omega})\pi(x, \mathrm{d}\omega), & x \in D, \\ \sup_\omega \{\mathrm{pr}_{x*\omega}(E_{x*\omega})\}, & x \notin D, \end{cases} \tag{3.1}$$

which can also be used to calculate $\mathrm{pr}(E) \equiv \mathrm{pr}_\Box(E_\Box)$. (Compare these formulae with the similar but different formulae of Dawid (1993).)

We can now start to formally connect, in our basic prequential framework, Vovk's prequential probability framework to Shafer's axiomatics. To this end, to facilitate the exposition, we first recall the definition of global probability based on non-negative π-martingales (Vovk, 1993a), and secondly translate into our notation Shafer's definition of probability in situations.

Definition 3.4.1 *For a given forecasting system π, a measurable function $S: \Omega^* \to [0, \infty]$, is a non-negative π-martingale if $S(x) = \int_\Omega S(x * \omega)\pi(x, d\omega)$, for π-prior x, and $S(x) = S(x * \omega)$, for all ω, for x which are not π-prior. Then, for $E \subseteq \Omega^\infty$, Vovk's global probability is defined as*

$$\mathrm{pr}(E) = \inf\{S(\Box)\},$$

where S ranges over the non-negative π-martingales which are successful on E, that is, such that for all $\xi \in E$, $\sup_{n \in \mathbf{N}}\{S(\xi^n)\} > 1$.

Definition 3.4.2 *Let $x, y, z \in \Omega^*$ and $E, F \subseteq \Omega^\infty$, then $\mathrm{pr}_x(E_x)$ is a probability in situations if it satisfies the following axioms.*

S1. $0 \leq \mathrm{pr}_x(E_x) \leq 1$ *for every sequence x and every event E.*

S2. $\mathrm{pr}_x(E_x) = 0$ *if and only if E is impossible in x.*

S3. $\mathrm{pr}_x(E_x) = 1$ *if and only if E is certain in x.*

S4. *If E and F are incompatible in x, then $\mathrm{pr}_x(E_x \cup F_x) = \mathrm{pr}_x(E_x) + \mathrm{pr}_x(F_x)$.*

S5. *If $\Gamma_y \subseteq \Gamma_x$ and $\Gamma_z \subseteq \Gamma_y$, then $\mathrm{pr}_x((\Gamma_z)_x) = \mathrm{pr}_x((\Gamma_y)_x)\mathrm{pr}_y((\Gamma_z)_y)$.*

We can now consider the following propositions.

Proposition 3.4.1 *Vovk's global probability satisfies Shafer's axioms when π is total and the event tree is $\Omega \times \Omega \times \cdots$.*

Proof. For every $x \in \Omega^*$, let pr_x be defined in accordance with Vovk's definition of global probability. By theorem 4 of Vovk (1993a), if P is a probability distribution in Ω^∞ generating a total forecasting system π, then $\mathrm{pr} = P$. Also, $\mathrm{pr}_x(E_x) = P(E|\Gamma_x)$ is a probability distribution in Ω^∞. Thus, for all $x \in \Omega^*$ and $E \subseteq \Omega^\infty$, $\mathrm{pr}_x(E_x)$ satisfies S1, S2, S3 and S4. Moreover, for $x, y, z \in \Omega^*$, $\Gamma_z \subseteq \Gamma_y$ and $\Gamma_y \subseteq \Gamma_x$, the standard product rule applied to P yields

$$P(\Gamma_z|\Gamma_x) = P(\Gamma_y|\Gamma_z)P(\Gamma_z|\Gamma_y),$$

and then $\mathrm{pr}_x(E_x)$ satisfies S5 too. **Q.E.D.**

In the next proposition we assume Ω to be countable.

Proposition 3.4.2 *For a countable set Ω, Shafer's probability in situations satisfies Vovk's definition of global probability when π is total and the event tree is $\Omega \times \Omega \times \cdots$.*

Proof. Suppose the probability $\mathrm{pr}_x(E_x)$, defined for all $x \in \Omega^*$ and for all $E \subseteq \Omega^\infty$, satisfies Shafer's axioms. Then, for an $x \in \Omega^*$, taking the event E to be of the form $E = \{\xi : \xi^{|x|+1} = x*\omega, \omega \in A\}$, where $A \subseteq \Omega$, and writing $\pi(x, A) = \mathrm{pr}_x(\bigcup_{\omega \in A} \Gamma_\omega) = \mathrm{pr}_x(E_x)$, we see that, for every $x \in \Omega^*$, $\pi(x, A)$ satisfies the usual axioms for a probability distribution in Ω. To show that $\mathrm{pr}(E)$ satisfies Vovk's definition of global probability, we note that, for every $x \in \Omega^*$ and a generic $E \subseteq \Omega^\infty$,

$$\mathrm{pr}_x(E_x) = \sum_{\omega \in \Omega} \mathrm{pr}_x(\Gamma_\omega)\mathrm{pr}_{x*\omega}(E_{x*\omega}).$$

By writing $\mathrm{pr}_x(\Gamma_\omega)$ as $\pi(x, \omega)$, we can see that, for every fixed $E \in \Omega^\infty$, $\mathrm{pr}_x(E_x)$ is a non-negative π-martingale which we call $S'(x)$. Thus,

$$\mathrm{pr}(E) \equiv S'(\Box) = \inf\{S(\Box)\},$$

where S ranges over the non-negative π-martingales which are successful on E. In fact, $S'(x)$ is successful on E because $\mathrm{pr}_x(E_x) = 1$ when E is certain in x, and $S'(\Box)$ is the infimum because $\mathrm{pr}_x(E_x) = 0$ when E is impossible in x. **Q.E.D.**

Note that the correspondence we have just proved, between Shafer's axiomatics and Vovk's prequential probability framework, is bound to a particular special case. In the more general case of a non-total forecasting system π, we just remember that Vovk's global probability satisfies, on the event tree $\Omega \times \Omega \times \cdots$, just axioms S1, S2 and S5.

3.5 On the Prequential Principle

In this section we briefly discuss Dawid's prequential principle, which has been considered in Chapter 2 in the classical Kolmogorov probability axiomatics, in Vovk's prequential probability framework. For a discrete observation space $\Omega =$

$\{a_1, a_2, \ldots, a_m, \ldots\}$, consider a total forecasting system π on the event tree $\Omega \times \Omega \times \cdots$ (and a corresponding distribution P on Ω^∞). Suppose we perform an experiment. An experiment allows us to observe one, and only one, single path down the tree. And once the experiment has taken place, all our empirical information will be encapsulated in the finite data-sequence $\mathbf{x}^n = (\omega_1, \omega_2, \ldots, \omega_n)$, where $\omega_i \in \Omega$, $i = 1, 2, \ldots, n$. In the light of $\mathbf{x}^n$, the prequential principle says that we should evaluate our forecasting system π only on the basis of the two sequences of realized data-values and of realized probability forecasts,

$$\begin{array}{lccccc} \mathbf{x}^n\text{:} & \omega_1 & \omega_2 & \omega_3 & \ldots & \omega_n, \\ \boldsymbol{\pi}^n\text{:} & \pi(\Box, \cdot) & \pi(\mathbf{x}^1, \cdot) & \pi(\mathbf{x}^2, \cdot) & \ldots & \pi(\mathbf{x}^{n-1}, \cdot), \end{array}$$

where the $\pi(\cdot, \cdot)$'s are the predictive distributions supplied by π. The pair of realized sequences $(\mathbf{x}^n, \boldsymbol{\pi}^n)$ will be called the *prequential path.*

Given these two sequences, let us consider the sets $D_{ij} = \{\xi : \xi^i = \omega_1 \omega_2 \ldots \omega_{i-1} a_j\}$, for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m, \ldots$. Each D_{ij} is a set in Ω^∞ which depends on the realized data-sequence $\mathbf{x}^n = (\omega_1, \omega_2, \ldots, \omega_n)$. We call the *prequential path partition* the partition on Ω^∞ defined by

$$\Pi_x = \{D_{ij} : i = 1, 2, \ldots, n \text{ and } a_j \neq \omega_i \text{ if } i < n\}.$$

This partition depends just on the sequence $\mathbf{x}^n$ of realized outcomes, and it includes all those events, and only those, for which the forecasting system π was asked to give its predictions, while our knowledge about the world unfolded down the tree $\Omega \times \Omega \times \cdots$ along the path $\mathbf{x}^n$. Note that the probability (under P, we can say) of the sets $D_{ij} \in \Pi_x$ is completely defined (unlike that of a generic event E) by the prequential path $(\mathbf{x}^n, \boldsymbol{\pi}^n)$, through the formula

$$P(D_{ij}) = \pi(\Box, \omega_1)\pi(\mathbf{x}^1, \omega_2) \cdots \pi(\mathbf{x}^{i-2}, \omega_{i-1})\pi(\mathbf{x}^{i-1}, a_j).$$

For a generic event $E \subseteq \Omega^\infty$, let us now define its upper probability by

$$\overline{P}_x(E) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ a_j \neq \omega_i, i<n}}^{m} \{P(D_{ij}) : D_{ij} \cap E \neq \emptyset\}.$$

Note immediately that, for the sets $D_{ij} \in \Pi_x$, we have $\overline{P}_x(D_{ij}) = P(D_{ij})$. The probability $\overline{P}_x(\cdot)$ can be thought of as a way of transporting the probabilistic information at hand from the prequential path $(\mathbf{x}^n, \boldsymbol{\pi}^n)$ to the sample space Ω^∞. Indeed, $\overline{P}_x(\cdot)$ is in one-to-one correspondence with the couple $(\mathbf{x}^n, \boldsymbol{\pi}^n)$, and every inference based on the probabilities $\overline{P}_x(\cdot)$ will thus respect the prequential principle.

This upper probability has an interesting interpretation in terms of Vovk's global probability. Consider the 'minimal' forecasting system $\overline{\pi}$, which is non-total, whose $\overline{\pi}$-prior are the initial fragments of the realized data-sequence $\mathbf{x}^n$ and whose predictions, for these π-prior, are identical to the predictions made by π. Then

$$\overline{P}_x(E) = \mathrm{pr}_{\overline{\pi}}(E),$$

where $\mathrm{pr}_{\overline{\pi}}(E)$ is Vovk's global probability for the forecasting system $\overline{\pi}$. This can be seen using the recursive formulae (3.1) given in Section 3.4. In this simple case, put a one on the nodes of the event tree at which E happens and put a zero on the nodes at which E fails. (Using the language of Section 3.3, we say that E *happens at* S if E is certain in S, but not in the situation immediately above it. And we say that E *fails at* S if E is impossible in S, but not in the situation immediately above it.) These ones and zeros are just the probabilities $\mathrm{pr}_x(E_x)$ when x is a node at which E happens or fails. Then the global probability $\mathrm{pr}_{\overline{\pi}}(E)$ can be obtained by applying the recursive formulae towards the root of the tree.

Another interpretation of the upper probability $\overline{P}_x(E)$ comes from Dempster–Shafer theory of belief functions (see Shafer, 1976). In accordance to this theory, $\overline{P}_x(E)$ is a *plausibility function*, whereas

$$\underline{P}_x(E) = 1 - \overline{P}_x(\Omega^\infty \setminus E),$$

is a *belief function.*

Consider now what happens if the observation space Ω is uncountable. For a total forecasting system π on the event tree $\Omega \times \Omega \times \cdots$, consider, as before, an observed sequence of outcomes $\mathbf{x}^n = (\omega_1, \omega_2, \ldots, \omega_n)$, and an observed sequence of probability forecasts $\boldsymbol{\pi}^n = (\pi(\Box, \cdot), \pi(\mathbf{x}^1, \cdot), \ldots, \pi(\mathbf{x}^{n-1}, \cdot))$. For $i = 1, 2, \ldots, n$, and

$a \in \Omega$, define the sets $D_{ia} \subseteq \Omega^\infty$ by

$$D_{ia} = \{\xi : \xi^i = \omega_1\omega_2 \ldots \omega_{i-1}a\},$$

where $\xi \in \Omega^\infty$. Then the analogue of the prequential path partition of the discrete case is given by the partition on Ω^∞ defined by

$$\Pi_x = \{D_{ia} : i = 1, 2, \ldots, n,\ a \in \Omega \text{ and } a \neq \omega_i \text{ if } i < n\},$$

and, for every $E \subseteq \Omega^\infty$, the upper probability is now given by the formula

$$\overline{P}_x(E) = \sum_{i=1}^{n} \int_{\xi \in \Omega^\infty} I_{\bigcup\{D_{ia} : D_{ia} \cap E \neq \emptyset, a \in \Omega, a \neq \omega_i \text{if } i<n\}}(\xi) P(\mathrm{d}\xi).$$

These quantities now lose much of their original appeal. If $\pi(\Box, \mathrm{d}\omega)$ is an absolutely continuous distribution with respect to Lebesgue measure, only the first term in the sum can be positive. Indeed, all other terms would be identically zero, being the integral, for $i = 2, 3, \ldots, n$, over a set of Lebesgue measure zero. Similar considerations would arise for a non-total forecasting system π.

3.6 A Purely Martingale Probability Framework

In Section 3.2, we considered a probability framework in which no Kolmogorovian probability distribution P over Ω^∞ was being introduced. In that framework, a definition of probability over Ω^∞, called global π-probability, was given by means of martingales, namely π-martingales, which were defined with respect to a probability forecasting system π. Vovk (1993b, 1993c, 1995a), instead of a probability distribution P, or of a probability forecasting system π, proposed, as a foundation for probability theory, to use only sequences of measurable functions, which are called M-martingales, applying directly to them the principle of the excluded gambling strategy (see also Shafer, 1995). In this section, we will present a variant of this purely martingale framework, which will also be used in the following chapters. Our exposition will follow to a considerable degree Vovk (1993c).

Let $(\Omega^\infty, (\mathcal{F}_0, \mathcal{F}_1, \ldots), \mathcal{F})$ be a fixed filtered space with $\mathcal{F}_0 = \{\emptyset, \Omega^\infty\}$. To this we add an $\mathbf{R}^k$-valued, $k \in \mathbf{N}$, stochastic sequence M, that is, a sequence of $\mathbf{R}^k$-valued

random elements $M_0, M_1, \ldots$ such that $M_0 = 0$, and each M_n is $\mathcal{F}_n$-measurable. The stochastic sequence M is called the *basic martingale*, and the stochastic sequence $M_1 - M_0, M_2 - M_1, \ldots$ is called the *basic martingale difference sequence.* The filtered space $(\Omega^\infty, (\mathcal{F}_0, \mathcal{F}_1, \ldots), \mathcal{F})$ complemented with the basic martingale M forms a *finite-dimensional martingale model.*

If $V = (V_1, V_2, \ldots)$ is an $\mathbf{R}^k$-valued predictable sequence (that is, a sequence of $\mathbf{R}^k$-valued random elements such that each V_n is $\mathcal{F}_{n-1}$-measurable), then the stochastic sequence defined by

$$(V \cdot M)_n = \sum_{i=1}^{n} V_i \cdot (M_i - M_{i-1}),$$

where $\cdot$ in the right-hand side stands for the inner product of vectors in $\mathbf{R}^k$, is called the *martingale transform* $V \cdot M$. Stochastic sequences of the form $c + V \cdot M$, with $c \in \mathbf{R}$, are called *M-martingales.*

To these definitions are attached the following gambling interpretations. Each element of the, in general multivariate, basic martingale M can be considered the evolution of our capital in an infinite sequence of fair games against an infinitely rich bookmaker in which at each trial we bet a unit of our capital. Each element of the multivariate value M_n is interpreted as our capital after n games, in the corresponding infinite sequence of fair games. The predictable sequence V and the martingale transform $V \cdot M$ represent, respectively, a combined strategy of varying bets over all the k infinite sequences of fair games represented by M, and the evolution of our capital which corresponds to this strategy.

With these elements, Vovk (1993c) gave the following definition of a null set. Note that a null set is not necessarily a measurable subset of Ω^∞.

Definition 3.6.1 *A set $E \subset \Omega^\infty$ is M-null if there is a non-negative M-martingale S such that $S_0 = 1$ and $S_n(\xi) \to \infty$, as $n \to \infty$, for all $\xi \in E$.*

In accordance with this definition, we also say that a set $E \subseteq \Omega^\infty$ is *M-almost sure* if the set $\Omega^\infty \setminus E$ is M-null. This definition, which provides the foundations of the purely martingale framework we present, is interpreted in terms of the infinitary

principle of the excluded gambling strategy (Vovk, 1993a, Section 3). If S is a pre-specified non-negative M-martingale such that $S_0 = 1$ and ξ is the realized outcome, then, as long as we believe in our martingale model M, we can be practically sure that $S_n(\xi)$ does not tend to infinity, as $n \to \infty$. So, if a set E is M-null, we can be practically sure that it will not happen.

As given by Vovk, this definition of an M-null set does not make use of any probability distribution, and so does not make use of Kolmogorov's axioms of probability. Nevertheless, consider extending our filtered space to a probability model by adding a probability distribution P in $\mathcal{F}$. In this case, if a pre-specified event $E \in \mathcal{F}$ satisfies $P(E) = 0$, we can be (almost) sure that E will not happen (provided we believe in the model).

Theorem 3.6.1 (Vovk, 1993c) *If M is a local martingale (see, e. g., Shiryayev, 1984) with respect to a probability distribution P in $(\Omega^\infty, \mathcal{F})$ and an event $E \in \mathcal{F}$ is M-null, then $P(E) = 0$.*

Proof. Let $E \in \mathcal{F}$ be M-null. There is a predictable $\mathbf{R}^k$-valued sequence V such that the M-martingale $S = 1 + V \cdot M$ is non-negative and $S_n(\xi) \to \infty$, as $n \to \infty$, for all $\xi \in E$. With respect to P, S is a non-negative local martingale, and hence, by Fatou's lemma, a non-negative supermartingale for which $\mathrm{E}|S_n| < \infty$, for all $n \geq 0$. Then, for any constant $c > 0$,

$$P(E) \leq P\{\xi : S_n(\xi) \to \infty, \text{ as } n \to \infty\} \leq P\{\xi : \exists\, n \text{ s.t. } S_n(\xi) \geq c\},$$

and, by Doob's inequality (Shiryayev, 1984, p. 464),

$$P\{\xi : \exists\, n \text{ s.t. } S_n(\xi) \geq c\} \leq \frac{1}{c}.$$

Therefore $P(E) = 0$. **Q.E.D.**

This theorem asserts that Vovk's martingale definition of an M-null set, which does not require the introduction of any probability distribution, does not lead to any contradiction when such a probability distribution is present. The converse statement that, for any $E \in \mathcal{F}$, if $P(E) = 0$, then E is M-null, is not true in general.

For instance, if the basic martingale is such that, for all $n \in \mathbf{N}$, $\min(M_n|\mathcal{F}_{n-1}) \leq M_{n-1} \leq \max(M_n|\mathcal{F}_{n-1})$, then no $\mathcal{F}_n$-measurable event, for n finite, can be M-null.

To show the power of these foundations, Vovk (1993c) proved a version of Kolmogorov's strong law of large numbers, and, by enriching the framework, Vovk (1995a) proved also a version of Lindeberg's central limit theorem. Note that Vovk's definition of an M-null set provides a definition of 'probability zero'. Probabilities other that zero could be defined by adapting the martingale construction in Vovk (1993a, Section 4). To this end, however, we would have to consider only basic martingales which are in some sense coherent, that is, basic martingales for which no gambling strategy can ensure a guaranteed profit at some future time.

3.7 Discussion

From an empirical point of view, subjective theory of coherence leaves open the problem of how to cope with coherent probabilistic assertions which seem to disagree with the actual observations. In the tradition of de Finetti and Savage, the classical non-contradictory way to assess subjective probabilistic assertions is to use scoring rules, without attaching to them any probabilistic meaning. Dawid (1982), in an attempt to find a more objective solution to this problem, proposed to discredit any probability assessment which failed to give positive probability to a pre-specified event that actually materialized, and Dawid (1985), using a generalization of the calibration criterion, also argued in favour of the existence of calibration-based empirical probabilities. On the other hand, Vovk (1993a) proposed to tackle the problem of the relation of disagreement between theory and observations in a different way. Attempts to create a general empirical theory of probability, he said, should be abandoned, and he argued that we should content ourselves with what he called the logic of probability, establishing relations between probabilistic theories and observations. To him, the problem of disagreement can only be solved by the introduction of an appropriate principle, and, inspired by the ideas of Dawid (1985, Section 13.2), he proposed to base the measure of disagree-

ment on martingales taking very large values, by using a version of the principle of the excluded gambling strategy.

Chapter 4

M-Typical Sequences

Random sequences are usually defined with respect to a probability distribution assuming Kolmogorov's axioms for probability theory. In this chapter, without using this axiomatics, we will give a definition of random (typical) sequences taking as primitive the notion of a martingale and using the principle of the excluded gambling strategy, on the lines of the purely martingale probability framework of Section 3.6.

To ease our exposition, we recall here some notation from the previous sections which will be used in this and the next two chapters. The set of positive integers $1, 2, \ldots$ is denoted by $\mathbf{N}$, and the sets of rational and real numbers are denoted respectively by $\mathbf{Q}$ and $\mathbf{R}$. For any set Ω, Ω^* is the set of finite sequences $\omega_1\omega_2\ldots\omega_n$ of elements of Ω; Ω^* includes the empty sequence $\Box$. The set Ω is called the observation space. The length of a sequence $x \in \Omega^*$ is denoted by $|x|$, whereas the concatenation of x with an element $\omega \in \Omega$ is denoted by $x * \omega$. For every $x \in \Omega^*$, the cylinder set $\Gamma_x \subseteq \Omega^\infty$ is defined by

$$\Gamma_x = \{\xi : \xi^{|x|} = x\}.$$

The set of infinite sequences $\omega_1\omega_2\ldots$ of elements of Ω is denoted by Ω^∞. If $\xi = \omega_1\omega_2\ldots$ is an infinite sequence of elements of Ω and $n \in \mathbf{N}$, ξ^n is the initial segment $\omega_1\omega_2\ldots\omega_n$ of ξ of length n.

Moreover, in what follows, for any real number z, the largest integer not greater

than z will be denoted by $\lfloor z \rfloor$. We will write $\ln z$ to indicate the natural (base e) logarithm of z. And the indicator function of the interval $(-\infty, z]$, $z \in (-\infty, \infty)$, will be denoted by $I_{(-\infty,z]}(\cdot)$.

4.1 The Notion of Randomness

Consider the following two sequences of twenty zeros and ones

$$0\,1\,0\,0\,1\,1\,0\,1\,1\,0\,0\,1\,1\,1\,0\,0\,1\,0\,1\,1,$$

$$0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0,$$

and suppose that we were asked to decide whether they are or not the result of the random flipping of a fair coin. Almost every one would agree that the first sequence could be the result of such a random experiment, and that the second sequence could not. In favour of this thesis, many would argue that the probability of the second sequence is too small, being equal to 2^{-20}, but they would forget that the first sequence has the same exact probability! The point is that when we are presented with something like the first sequence we do not consider it individually, but consider it as a class of sequences of the same type. And in thinking at the probability of the first sequence we actually think at the probability of a much bigger event. In some sense, this is due to the intrinsic complexity of the succession of zeros and ones in the sequence and not only at a priori probability.

This is not however just an isolated example. According to classical probability theory, every time we are faced with an experiment or a situation in which elements of uncertainty are present we are usually led to consider just a probability distribution governing the probabilities of all a priori possible outcomes. And in a situation like that above we are led to consider only the a priori probabilities of the two sequences without considering any other possible element. From the point of view of classical probability theory, even if our intuition is different, the above two sequences should both be regarded as possible results of the random flipping

of a fair coin. A more satisfactory answer to this problem is provided by the algorithmic approach. By enriching probability theory with some algorithmic notions it becomes possible to distinguish between the above two sequences and to justify the choice of the first sequence in accordance with our intuition.

4.2 Typical Sequences

Consider the set Ω^∞ of all possible infinite sequences of outcomes of an infinite sequence of random experiments described by a probability distribution P (not necessarily assuming independence). Also, consider a set of simple properties, each of them characterizing a small subset of Ω^∞. Following our intuition, an infinite sequence $\xi \in \Omega^\infty$ is felt to be *non-random* if there is one of these properties by means of which it can be characterized, whereas it is felt it could be *random* if there is no such simple property. Statistically, a simple property characterizing a small subset of Ω^∞ is represented by a test of 'absence of regularity' (randomness), which is a partition $(E, \Omega^\infty \setminus E)$ of Ω^∞, such that $P(E) = 1$, which is required to be *algorithmically computable.* Every sequence $\xi \in E$ is regarded as having passed the test and a sequence which passes all computable tests of randomness is said to be random. That we have to examine only tests which are computable is justified by considering that no sequence could ever pass all possible conceivable tests. This is in agreement with what our intuition would suggest, that an extremely complicated and convoluted property (that is a property that cannot be given by a finite amount of information) would be rejected as a characterization of non-random sequence. Since, by this criterion, a random sequence does not belong to any small fraction of all sequences, that is, to any (computable) subset having measure zero, such a sequence can be thought of as a typical representative of the class of all sequences. This property of typicalness, as an appropriate property characterizing the intuitive notion of randomness, was first proposed by Martin-Löf (1966).

In accordance with Kolmogorov and Uspenskii (1987), random sequences characterized in this way are called *typical*, to distinguish them from those random

sequences which arise from the *chaotic* and the *stochastic* approaches (the current use of the terms 'typical', 'chaotic' and 'stochastic' has been set in the above paper). The property of typicalness leads to the same class of random sequences as the property of chaoticness, whereas the property of stochasticness leads to a distinct definition of randomness with many drawbacks (see, for a review, Uspenskii, Semenov and Shen', 1990).

Once the concept of a random (typical or chaotic) sequence has been defined, it is possible to give 'pointwise' algorithmic counterparts to many of the classical almost sure limit results of probability theory such as the strong law of large numbers and the law of the iterated logarithm.

4.3 Computable Martingales

Assuming Kolmogorov's axioms for probability theory, martingale processes have been widely used and extensively studied (see, e. g., Schnorr, 1971, 1977) as test functions for defining typical sequences with respect to a probability distribution P. Our present goal, however, is that to make use of the concept of a martingale in a more direct way. Instead of defining a typical sequence with respect to a probability distribution P, we will define a typical sequence with respect to a sequence of measurable functions, which we declare to be a martingale, by using the principle of the excluded gambling strategy, and without introducing any probability distribution. This use of the principle of the excluded gambling strategy as a foundation for probability theory, parallels, in an algorithmic framework, the purely martingale approach of Section 3.6.

To give our definition of a typical sequence, we will introduce in this purely martingale framework some algorithmic concepts, but first we will have to reconsider the framework itself. In this and the following two chapters, we will restrict ourselves to finite-dimensional martingale models for which each σ-algebra $\mathcal{F}_n$ is generated by a countable partition. We will always consider the observation space Ω to be a subset of $\mathbf{Q}$ (a classical example being $\Omega = \{0,1\}$). And, for a fixed observation

space Ω, we will always consider the family of σ-algebras $(\mathcal{F}_0, \mathcal{F}_1, \ldots)$ on Ω^∞ to be the filtration generated by the cylinder sets Γ_x, $x \in \Omega^*$. Stochastic and predictable sequences are now regarded as functions from Ω^* into $\mathbf{R}^k$ $(k \geq 1)$. A stochastic sequence S is a function $S: \Omega^* \to \mathbf{R}^k$, (for every fixed $n \in \mathbf{N}$, $S(\xi^n)$ is measurable with respect to $\mathcal{F}_n$), and a predictable sequence V is a function $V: \Omega^* \to \mathbf{R}^k$ such that, for every fixed $n \in \mathbf{N}$, $V(\xi^n)$ is measurable with respect to $\mathcal{F}_{n-1}$. In what follows, quantities like $S(\xi^i)$, $V(\xi^i)$, etc., will often be abbreviated to S_i, V_i, etc.. It will be evident from the context if they refer to a stochastic sequence or to the realized sequence of values for a given ξ. The gambling picture of Section 3.6 will still be used to give an interpretation to the framework. The evolution of our capital, in an infinite sequence of 'fair games' against an infinitely rich bookmaker, in which at each trial we bet a unit of our capital, is represented by a given scalar stochastic sequence M_0, M_1, ..., which is called the basic martingale. The value M_n is interpreted as our capital after n games. In the same way, a finite collection of infinite sequences of 'fair games' is represented by a multivariate M. A strategy for varying the sizes of the bets is represented by a predictable sequence V_1, V_2, ..., and then the evolution of our capital corresponding to the application of this strategy is represented by the martingale transform

$$(V \cdot M)_n = \sum_{i=1}^{n} V_i \cdot (M_i - M_{i-1}).$$

Stochastic sequences of the form $c + V \cdot M$, where $c \in \mathbf{R}$ represents a starting capital, are called M-martingales.

Let us now give the following algorithmic definitions. More general definitions and some more elements of the theory of algorithms are given in the Appendix. We say that a stochastic sequence $S: \Omega^* \to \mathbf{R}$ is *computable* if there is an algorithm $\mathcal{U}$ which transforms any input $x \in \Omega^*$ and positive integer n into a rational number r satisfying $|S(x) - r| \leq 2^{-n}$. That is, the stochastic sequence S is computable if its values can be computed arbitrarily accurately by some fixed algorithm. We also say that a stochastic sequence S is *lower semicomputable* if there is an algorithm $\mathcal{U}$ which, when fed with a rational number r and an input $x \in \Omega^*$, eventually stops if $S(x) > r$ and never stops otherwise. Lower semicomputability of S means that if

$S(x) > r$ this fact will sooner or later be learned whereas if $S(x) \leq r$ we may be for ever uncertain. For a computable basic martingale M, M-martingales obtained from a computable predictable sequence V by $c + V \cdot M$, where c is a rational number and $V \cdot M$ is a martingale transform, are computable. Since there is only a countable number of algorithms, the sets of all computable M-martingales, and of all lower semicomputable M-martingales, are countable.

4.4 M-Typical Sequences

Under the usual Kolmogorov axiomatics for probability theory, every definition of typical sequences, with respect to a probability distribution P, corresponds to a definition of an effectively null set (see, e. g., Uspenskii, Semenov and Shen', 1990, Section 2.1). In our purely martingale framework, we define an effectively null set as follows (cf. Definition 3.6.1).

Definition 4.4.1 *A set $E \subset \Omega^\infty$ is effectively M-null if there is a lower semicomputable non-negative M-martingale S such that $S_0 = 1$ and $S(\xi^n) \to \infty$, as $n \to \infty$, for all $\xi \in E$.*

As before, we also say that a set $E \subseteq \Omega^\infty$ is *effectively M-almost sure* if the set $\Omega^\infty \setminus E$ is effectively M-null. Notice that, to define an effectively M-null set it is not necessary to assume that the basic martingale M is computable. All that is algorithmically needed is the set of lower semicomputable non-negative M-martingales with respect to an arbitrary basic martingale M. (For a criticism on the arbitrariness of the choice of this set, see Howard (1993).)

Definition 4.4.2 *An infinite sequence $\xi \in \Omega^\infty$ is M-typical if, for any lower semicomputable non-negative M-martingale S, such that $S_0 = 1$, $S(\xi^n)$ does not tend to infinity.*

An M-typical sequence may be interpreted as follows. Let us suppose that in our gambling picture, starting with a positive amount of money, at each trial we bet a fraction of our capital, with the constraint that we can never incur a

debt. Of course, for any possible sequence of outcomes there would always be a winning strategy among all possible strategies. But if we limit ourselves to only those strategies which are not too complicated, viz. those which can be effectively calculated by means of some algorithm, then we would expect that, whatever betting strategy we might decide to employ, we will never become richer and richer as the game goes on (see Figure 4.1). If it really happens that we become richer and richer, this is because the actual sequence of outcomes is not random at all.

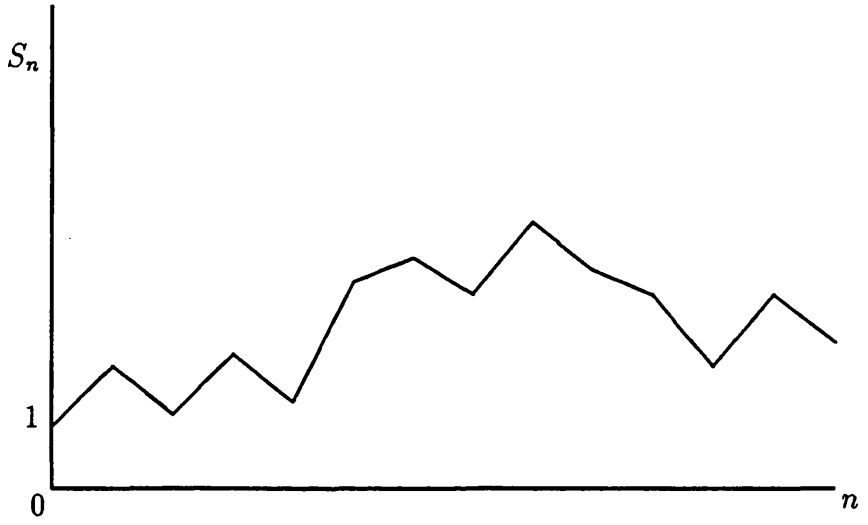


Figure 4.1: Realization of a lower semicomputable non-negative M-martingale S, $S_0 = 1$, for an M-typical sequence ξ.

An effectively M-null set E cannot contain any M-typical sequence. In fact, if E were to contain an M-typical sequence ξ, then for some lower semicomputable non-negative M-martingale S, with $S_0 = 1$, we would have $S(\xi^n) \to \infty$, as $n \to \infty$, in contradiction with ξ being M-typical. We denote the set of all M-typical sequences by T_M. Then this set is the intersection of all effectively M-almost sure sets or, equivalently, the set $\Omega^\infty \setminus T_M$ is the union of all effectively M-null sets.

In the traditional treatment of random sequences, assuming Kolmogorov's axiomatics, the enumerability of the set of all test functions leading to some notion of typicalness guarantees that the set of all typical sequences has measure one (by using, e. g., Proposition 2.4(a) of Williams (1991)). In our context, this property translates into the statement that the set of all M-typical sequences is M-almost

sure (see Definition 3.6.1).

Lemma 4.4.1 *The set T_M of all M-typical sequences is M-almost sure.*

Proof. We have to prove that the set $\Omega^\infty \setminus T_M$ is M-null. That is, we have to find a non-negative M-martingale S such that $S_0 = 1$ and $S(\xi^n) \to \infty$, as $n \to \infty$, for all $\xi \in \Omega^\infty \setminus T_M$. Now, since the set of all lower semicomputable M-martingales is countable, we can consider an enumeration of all lower semicomputable non-negative M-martingales starting at one, $S^{(1)}, S^{(2)}, \ldots$ say, and consider the non-negative M-martingale

$$S_n = \sum_{i=1}^{\infty} \frac{1}{2^i} S_n^{(i)}.$$

The non-negative M-martingale S is such that $S_0 = 1$ and $S(\xi^n) \to \infty$, as $n \to \infty$, whenever there is an i such that $S^{(i)}(\xi^n) \to \infty$, as $n \to \infty$. **Q.E.D.**

This lemma tells us that we can reasonably substitute the set Ω^∞ of all infinite sequences with the smaller set T_M of all M-typical sequences.

A similar result, but of a somewhat different nature, would state that the set T_M is also effectively M-almost sure, that is, that the union of all effectively M-null sets is also an effectively M-null set. The existence of such a 'maximal' effectively null set is guaranteed, for instance, under Kolmogorov's axiomatics, by the classical definition of typical sequences proposed by Martin-Löf (1966, Section III). In our framework, to prove that the set T_M is effectively M-almost sure, we would have to consider M-supermartingales instead of M-martingales in Definitions 4.4.1 and 4.4.2 of an effectively M-null set and of an M-typical sequence respectively (Vovk, 1995b).

4.5 More on M-Typical Sequences

In this section we consider the problem of whether an M-typical sequence can also be typical with respect to some other stochastic sequence different from M. For a given basic martingale M and for an arbitrary stochastic sequence N, we will say that an infinite sequence $\xi \in \Omega^\infty$ is N-typical if Definition 4.4.2 holds for ξ when

N is taken to be the basic martingale. With this terminology, let us now consider the following lemma.

Lemma 4.5.1 *Let $M_n = \sum_{i=1}^n X_i$ be an $\mathbf{R}^k$-valued ($k \geq 1$) computable basic martingale and let $N_n = c + \sum_{i=1}^n V_i \cdot X_i$, where c is a rational number and V is an $\mathbf{R}^k$-valued ($k \geq 1$) computable predictable sequence, be a computable M-martingale. Then every sequence $\xi \in \Omega^\infty$ which is M-typical is also N-typical.*

Proof. Consider an M-typical sequence $\xi \in \Omega^\infty$. We have to show that ξ is N-typical, that is, that for any lower semicomputable non-negative N-martingale S with $S_0 = 1$, $S(\xi^n)$ does not tend to infinity. Let S be a lower semicomputable non-negative N-martingale with $S_0 = 1$, then

$$\begin{aligned} S_n &= 1 + \sum_{i=1}^n W_i(N_i - N_{i-1}) \\ &= 1 + \sum_{i=1}^n (W_i V_i) \cdot X_i, \end{aligned}$$

for some predictable sequence W. So, S is also a lower semicomputable non-negative M-martingale with $S_0 = 1$. Then, since ξ is M-typical, $S(\xi^n)$ does not tend to infinity, and ξ is also N-typical. **Q.E.D.**

In plain words, this lemma says that the set of all lower semicomputable non-negative N-martingales, starting at one, is included in the set of all lower semicomputable non-negative M-martingales, starting at one. A more general result, which also includes the previous lemma, is the next one.

Lemma 4.5.2 *Let $M_n = \sum_{i=1}^n X_i$ be an $\mathbf{R}^k$-valued ($k \geq 1$) computable basic martingale, and consider the computable M-martingales $N_n^{(j)} = c_j + \sum_{i=1}^n V_i^{(j)} \cdot X_i$, $j = 1, 2, \ldots, J$, where c_j are rational numbers and $V^{(j)}$ are $\mathbf{R}^k$-valued ($k \geq 1$) computable predictable sequences. Define the computable stochastic sequence $N_n = [N_n^{(1)}, N_n^{(2)}, \ldots, N_n^{(J)}]'$. Then every sequence $\xi \in \Omega^\infty$ which is M-typical is also N-typical.*

Proof. Since for an $\mathbf{R}^J$-valued computable predictable sequence W we have that

$$1 + \sum_{i=1}^n W_i \cdot (N_i - N_{i-1})$$

$$= 1+\sum_{i=1}^{n}[W_{1i},\ldots,W_{Ji}][V_i^{(1)}\cdot X_i,\ldots,V_i^{(J)}\cdot X_i]'$$
$$= 1+\sum_{i=1}^{n}[(W_{1i}V_{1i}^{(1)}+\cdots+W_{Ji}V_{1i}^{(J)}),\ldots,(W_{1i}V_{ki}^{(1)}+\cdots+W_{Ji}V_{ki}^{(J)})][X_{1i},\ldots,X_{ki}]',$$

every lower semicomputable non-negative N-martingale S, with $S_0 = 1$, is also a lower semicomputable non-negative M-martingale, with $S_0 = 1$. Then, for an M-typical sequence ξ, for any lower semicomputable non-negative N-martingale S, with $S_0 = 1$, $S(\xi^n)$ does not tend to infinity, and ξ is also N-typical. **Q.E.D.**

Note that, in general, we cannot say that a sequence $\xi \in \Omega^\infty$ is $[M^{(1)}, M^{(2)}]'$-typical just because it is simultaneously $M^{(1)}$-typical and $M^{(2)}$-typical, where $M^{(1)}$ and $M^{(2)}$ are computable stochastic sequences.

4.6 Discussion

The above definition of M-typical sequences provides the basis of an approach to randomness in which, unlike the standard algorithmic approach, assuming Kolmogorov's axioms of probability, no probability distribution P over Ω^∞ needs to be introduced. All that we need is an infinite sequence of 'fair games', and an appeal to the long-term impossibility of winning against an infinitely rich bookmaker. Even if M-typical sequences are so unpredictable that no computable gambling strategy can hope to gain anything against them, this does not mean that they cannot satisfy interesting asymptotic statistical properties. Indeed the contrary is true. In the next two chapters we show some algorithmic versions for M-typical sequences of the strong law of large numbers, of the upper half of the law of the iterated logarithm, and of the strong central limit theorem. This algorithmic use of the principle of the excluded gambling strategy, as a foundation for probability theory, parallels its use made in Section 3.6 in the purely martingale probability framework of Vovk (1993b, 1993c, 1995a). On the other hand, the algorithmic counterpart of the prequential probability framework of Vovk (1993a) has been considered by Vovk and V'yugin (1993, 1994).

The algorithmic framework that has been laid in this chapter is powerful enough

to handle any situation in which the observation space is a subset of the rational numbers. For dealing with the more general situation in which the observation space is a subset of the computable real numbers, we would have to deal with computable functions defined over sets of computable real numbers, which requires a non-trivial extension of the algorithmic definitions used here. It should also be stressed that this approach is suitable for dealing with limiting results, that is, with results for infinite sequences of outcomes. For dealing with finite sequences of outcomes we would have to replace our definition of M-typical sequences with some purely martingale equivalent to the standard notion of 'deficiency of randomness' (see Kolmogorov and Uspenskii, 1987). The resulting algorithmic approach would probably become much less appealing, as is the case in the standard Kolmogorov framework, in which, with respect to the non-algorithmic formulation, exact equalities and inequalities have to be replaced by equalities and inequalities to within a constant factor.

Chapter 5

Strong Law of Large Numbers and Law of the Iterated Logarithm

In this chapter we will prove some versions for M-typical sequences of the strong law of large numbers and of the upper half of the law of the iterated logarithm. Let us note that these results will be derived in a framework in which no probability distribution P over Ω^∞ is being introduced, and so, without assuming Kolmogorov's axioms of probability. Due to its importance, the convergence lemma used in the proof of the strong law of large numbers and of the upper half of the law of the iterated logarithm is presented in Section 5.1. In Section 5.2 and 5.3 we give the strong law of large numbers with some of its variants. In Section 5.4 and 5.5 we give a calibration theorem and a classical refinement of the strong law of large numbers respectively. In Section 5.6 and 5.7 we then give the upper half of the law of the iterated logarithm, again with some of its variants, for the case of a binary basic martingale. Finally, in Section 5.8 we consider a strong law of large numbers in the case of sampled martingales.

5.1 The Convergence Lemma

The following lemma, which will be used in the proof of the strong law of large numbers and of the upper half of the law of the iterated logarithm, is an analogue of

Doob's martingale convergence theorem (Doob, 1953, Ch. VII, Section 4), restricted to non-negative martingales. The proof, which is in a way archetypical, depends on a reductio ad absurdum which involves the definition of M-typical sequences and the asymptotic behaviour of an appropriate computable non-negative M-martingale.

Lemma 5.1.1 *If S is a computable non-negative M-martingale with $S_0 > 0$ and ξ is an M-typical sequence, then the limit $\lim_{n\to\infty} S(\xi^n)$ (of the sequence of values of S for the M-typical sequence ξ) exists and is finite.*

Proof. (Reductio ad absurdum.) Suppose the limit $\lim_{n\to\infty} S(\xi^n)$ does not exist. Then there are rational numbers a_1, a_2, b_1, b_2 such that

$$\liminf_{n\to\infty} S(\xi^n) < a_1 < a_2 < b_1 < b_2 < \limsup_{n\to\infty} S(\xi^n).$$

Fix an algorithm computing S, $\mathcal{U}_S$ say, and consider the algorithm, which takes as input (ξ^n, m), $m = 1, 2, \ldots$, and yields as output zero or one for $n = 1, 2, \ldots$ and never stops for $n = 0$, represented by the flowchart of Figure 5.1.

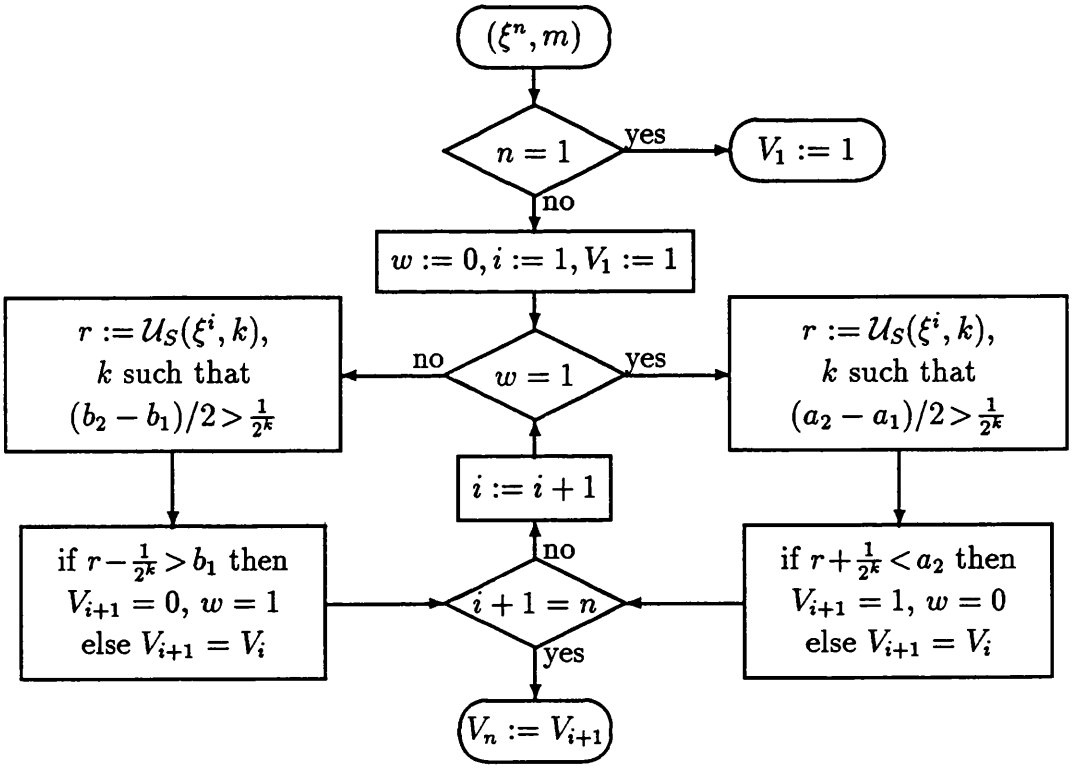


Figure 5.1: Flowchart of the algorithm $\mathcal{U}_V$ computing the binary predictable sequence V in the proof of Lemma 5.1.1.

(Note that, in the flowchart, the value of m is not actually used, and r is the rational provided by the algorithm $\mathcal{U}_S$ when the input is (ξ^i, k), $k \in \mathbf{N}$.) This algorithm, $\mathcal{U}_V$ say, is completely determined by the rational approximations to S given by $\mathcal{U}_S$. It defines a binary predictable sequence, V say, which is computable, by definition, by $\mathcal{U}_V$ itself (here whatever the value of m, V is always computed exactly). The sequence V, starting at one, takes value one until S crosses b_1 or b_2, then takes value zero until S crosses a_1 or a_2, then again it takes value one until S crosses b_1 or b_2, and so forth.

We define

$$S^*(\xi^n) = 1 + \frac{1}{S_0} \sum_{i=1}^{n} V(\xi^i)\big[S(\xi^i) - S(\xi^{i-1})\big].$$

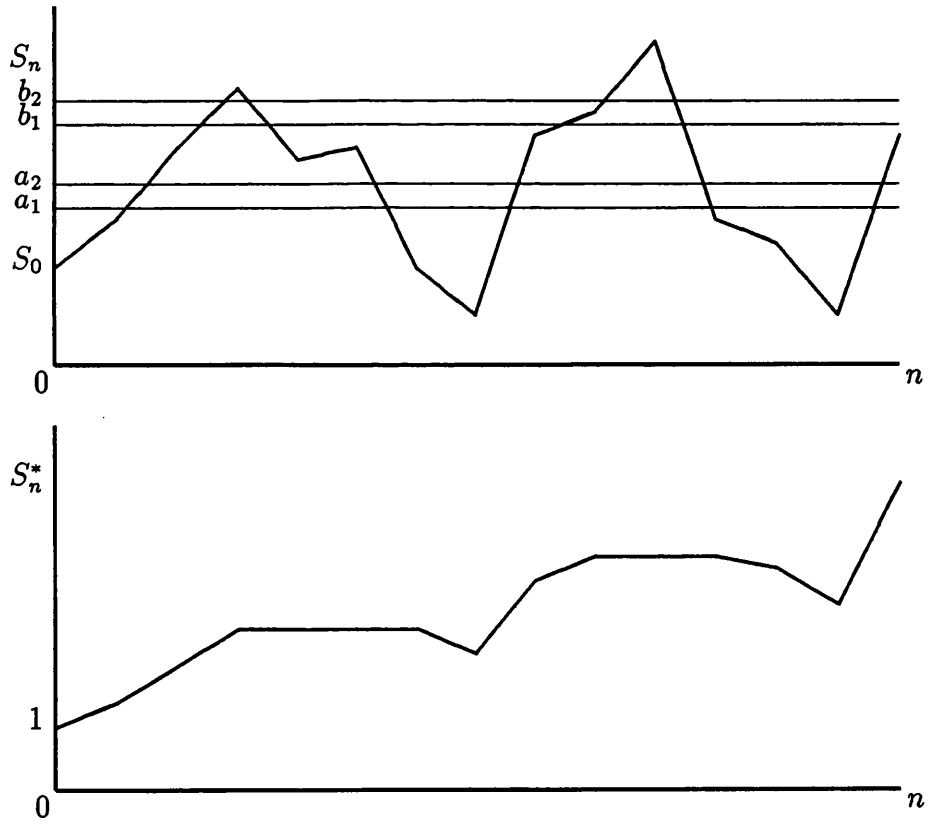


Figure 5.2: Non-converging realization of the M-martingale S and corresponding realization of the non-negative M-martingale S^*.

The function S_n^* is a lower semicomputable (indeed computable) non-negative M-martingale, with $S_0^* = 1$, such that $S^*(\xi^n) \to \infty$, as $n \to \infty$, for the M-typical

sequence ξ (see Figure 5.2). So, by contradiction, the limit $\lim_{n\to\infty} S(\xi^n)$ exists. To prove that this limit is also finite, just consider the lower semicomputable non-negative M-martingale S/S_0. **Q.E.D.**

In Lemma 5.1.1, and Lemma 5.2.1 below, we use flowcharts to depict our algorithms. It is a remarkable fact that the notion of a flowchart leads to a specific theory of flowchart computable functions which is equivalent to the classical theory of computability based on recursive functions (see, e. g., Odifreddi, 1992).

5.2 The Strong Law of Large Numbers

Let us consider, under a classical probability distribution P, the following generalization of Kolmogorov's strong law of large numbers. If $X_1, X_2, \ldots$ is a martingale difference sequence with respect to a filtration $(\mathcal{F}_0, \mathcal{F}_1, \ldots)$, then

$$\sum_{i=1}^{\infty} \frac{\mathrm{E}(X_i^2|\mathcal{F}_{i-1})}{i^2} < \infty \implies \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = 0,$$

almost surely. Vovk (1993c), as an application of his purely martingale framework, proved a version of this result using only his definition of an M-null set (see Section 3.6). And because of Theorem 3.6.1, his result also implied the above generalization under Kolmogorov's axiomatics. In this section we will consider the 'pointwise' algorithmic version of this result based on the definition of M-typical sequences. In proving this version, we will parallel the proof given by Vovk (1993c) which resembles, in turn, the proof of Kolmogorov's classical result given in Shiryayev (1984, Theorem VII.5.4). This proof takes several steps, the first being provided by the lemma of the previous section.

An *M-submartingale* is defined as a stochastic sequence of the form $T = S + A$, where $S{:}\,\Omega^* \to \mathbf{R}$ is an M-martingale and $A{:}\,\Omega^* \to \mathbf{R}$ is a non-decreasing predictable sequence. If S and A are both computable, then T is computable as well. Any such sequence A is called a *compensator* of T.

Lemma 5.2.1 *If T is a computable non-negative M-submartingale, A is one of its*

computable compensators, and ξ is an M-typical sequence, then,

$$A_\infty(\xi) < \infty \implies T(\xi^n) \text{ converges.}$$

Proof. Suppose that the M-typical sequence ξ is such that $A_\infty(\xi) < \infty$. Then there is a $C \in \mathbf{N}$ such that $A(\xi^n) < C - 1$, for all n.

Fix an algorithm computing A, $\mathcal{U}$ say, and consider the algorithm $\mathcal{U}_A$ that takes as input (ξ^n, m), $m, n = 1, 2, \ldots$, feeds $\mathcal{U}$ with $(\xi^{n-1} * \omega, m)$, where ω is a fixed element of Ω, and yields as output the rational supplied by $\mathcal{U}$. This algorithm computes A yielding as output rational approximations that are predictable. We build out of $\mathcal{U}_A$ the algorithm $\mathcal{U}_V$, which takes as input (ξ^n, m), $m, n = 1, 2, \ldots$, and yields as output zero or one, given by the flowchart of Figure 5.3. (Note that, as in Figure 5.1, the value of m is not used.)

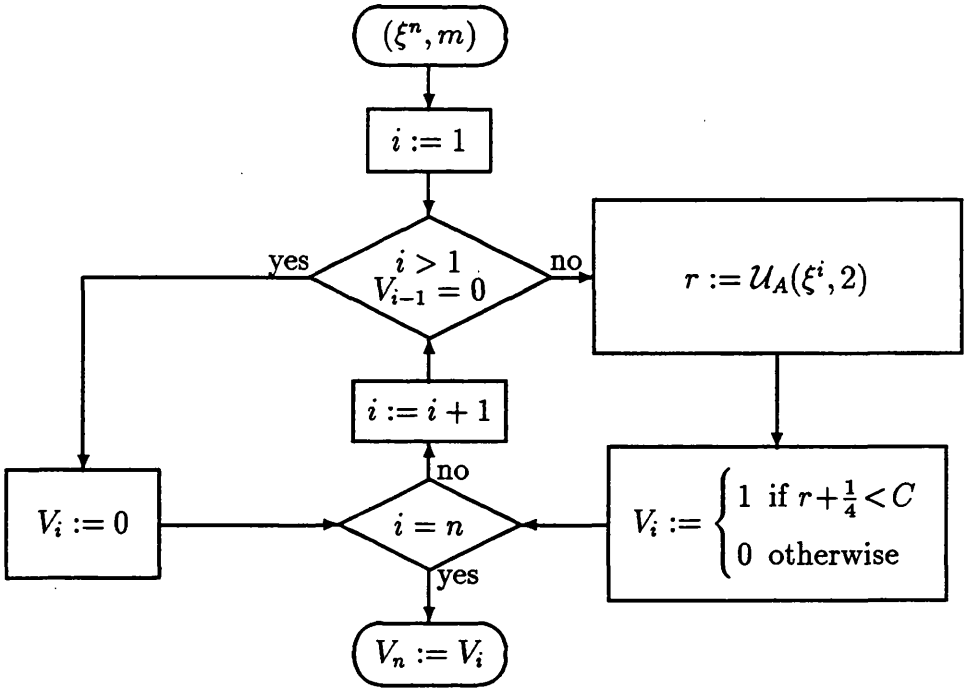


Figure 5.3: Flowchart of the algorithm $\mathcal{U}_V$ computing the binary predictable sequence V in the proof of Lemma 5.2.1.

This algorithm specifies a computable predictable sequence V with values zero and one. The paths of the sequence V are step functions, with only one jump, such

that, for $n = 1, 2, \ldots$,

$$V_n = \begin{cases} 1, & A_n \leq C - 1, \\ 1 \text{ or } 0, & C - 1 < A_n < C, \\ 0, & A_n \geq C. \end{cases}$$

Call S the computable M-martingale $T - A$ and consider the stochastic sequence

$$S_n^{(C)} = 1 + \sum_{i=1}^{n} V_i^{(C)}(S_i - S_{i-1}),$$

where $V^{(C)}$ is the computable predictable sequence defined by

$$V_n^{(C)} = \frac{V_n}{S_0 + C}.$$

It is easy to see that $S^{(C)}$ is a computable non-negative M-martingale with $S_0^{(C)} = 1$. By Lemma 5.1.1, the limit $\lim_{n\to\infty} S^{(C)}(\xi^n)$ exists and is finite and since, for $n = 1, 2, \ldots$,

$$S(\xi^n) = S^{(C)}(\xi^n)[S_0 + C] - C,$$

(for the M-typical sequence ξ), also the limit $\lim_{n\to\infty} S(\xi^n)$ exists and is finite. So, $T(\xi^n) = S(\xi^n) + A(\xi^n)$ converges, $A(\xi^n)$ being a bounded non-decreasing sequence by hypothesis. **Q.E.D.**

Lemma 5.2.2 *Let S be a computable M-martingale, let S^2 be a computable M-submartingale, and let A be one of its computable compensators. Then, for every M-typical sequence ξ,*

$$A_\infty(\xi) < \infty \implies S(\xi^n) \text{ converges.}$$

Proof. If A is a computable compensator of the computable M-submartingale S^2, then

$$(S_n + 1)^2 = (N_n + 2S_n + 1) + A_n,$$

where N is a computable M-martingale, and so A is also a computable compensator of the computable M-submartingale $(S+1)^2$. By Lemma 5.2.1, $S^2(\xi^n)$ and $(S(\xi^n)+1)^2$ converge when $A_\infty(\xi) < \infty$. Then, since

$$S(\xi^n) = \frac{1}{2}\Big[(S(\xi^n) + 1)^2 - S^2(\xi^n) - 1\Big],$$

$S(\xi^n)$ converges as well. **Q.E.D.**

Theorem 5.2.1 *Consider the computable basic martingale*

$$M_n = \begin{bmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n (X_i^2 - d_i) \end{bmatrix},$$

where X is a computable stochastic sequence and d is a computable non-negative predictable sequence. For every M-typical sequence ξ,

$$\sum_{i=1}^{\infty} \frac{d(\xi^i)}{i^2} < \infty \implies \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(\xi^i) = 0.$$

Proof. Let $S_0 = 0$, $S_i - S_{i-1} = X_i/i$; S is a computable M-martingale. Then, since

$$\Big(\sum_{i=1}^{n} \frac{X_i}{i}\Big)^2 = \Big[\sum_{i=1}^{n} \frac{(X_i^2 - d_i)}{i^2} + 2\sum_{i>j}^{n} \frac{X_i}{i}\frac{X_j}{j}\Big] + \sum_{i=1}^{n} \frac{d_i}{i^2},$$

and the term in square brackets is a computable M-martingale, S^2 is a computable M-submartingale and d_i/i^2 is one of its computable compensator difference sequences. By Lemma 5.2.2, applied to the computable M-martingale S,

$$\sum_{i=1}^{\infty} \frac{d(\xi^i)}{i^2} < \infty \implies \sum_{i=1}^{\infty} \frac{X(\xi^i)}{i} \text{ converges.}$$

Finally, by Kronecker's lemma (Stout, 1974, Lemma 3.2.3), which is valid for any sequence of real numbers,

$$\sum_{i=1}^{\infty} \frac{X(\xi^i)}{i} \text{ converges} \implies \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(\xi^i) = 0.$$

Q.E.D.

In our picturesque gambling interpretation, the strong law of large numbers can be described as saying that for a given M-typical sequence of outcomes the average of losses and winnings tends to zero, as $n \to \infty$, if the corresponding realized sequence of fair games satisfies some regularity conditions.

5.3 Variants of the Strong Law of Large Numbers

Here we will consider some variants of the strong law of large numbers which follow more or less directly from the law of Section 5.2. The first of these variants is for a generic martingale transform.

Theorem 5.3.1 *Consider the computable basic martingale*

$$M_n = \begin{bmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n (X_i^2 - d_i) \end{bmatrix},$$

where X is a computable stochastic sequence and d is a computable non-negative predictable sequence, and also consider a computable predictable sequence V. For every M-typical sequence ξ,

$$\sum_{i=1}^{\infty} \frac{V^2(\xi^i)d(\xi^i)}{i^2} < \infty \implies \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} V(\xi^i)X(\xi^i) = 0.$$

Proof. Let ξ be an M-typical sequence and consider the computable stochastic sequence

$$N_n = \begin{bmatrix} \sum_{i=1}^n V_i X_i \\ \sum_{i=1}^n V_i^2 (X_i^2 - d_i) \end{bmatrix}.$$

Since the components of N are computable M-martingales, ξ is also N-typical, and the desired result follows by applying the strong law of large numbers of Theorem 5.2.1 with N as the basic martingale. **Q.E.D.**

As an illustration of this theorem, take the above basic martingale with $X_i \in \{-1, 1\}$ and $d_i = 1$, $i = 1, 2, \ldots$. Then, for $V_i = X_{i-1}$, the theorem says that for every M-typical sequence the empirical autocorrelation at lag one

$$\frac{1}{n} \sum_{i=1}^{n} X_{i-1} X_i,$$

tends to zero, as $n \to \infty$.

The next variant is a strong law of large numbers for sampled basic martingales.

Theorem 5.3.2 *Consider the computable basic martingale*

$$M_n = \begin{bmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n (X_i^2 - d_i) \end{bmatrix},$$

where X is a computable stochastic sequence and d is a computable non-negative predictable sequence, and let $\{n_k\}$ be a computable predictable subsequence. Then for every M-typical sequence ξ,

$$\sum_{j=1}^{\infty} \frac{d_{n_{j-1}+1} + \cdots + d_{n_j}}{j^2} < \infty \implies \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} Y_j = 0,$$

where $Y_j = X_{n_{j-1}+1} + \cdots + X_{n_j}$.

Proof. Let $S_i - S_{i-1} = V_i X_i$, where $V_i = 1/j$, $n_{j-1} < i \leq n_j$. Then

$$\begin{aligned} S_n^2 &= \Big(\sum_{i=1}^{n} V_i X_i\Big)^2 \\ &= \Big[\sum_{i=1}^{n} V_i^2 (X_i^2 - d_i) + 2\sum_{i=1}^{n}\Big(V_i \sum_{r=1}^{i-1} V_r X_r\Big) X_i\Big] + \sum_{i=1}^{n} V_i^2 d_i, \end{aligned}$$

is a computable M-submartingale since the quantity in square brackets is a computable M-martingale and $V_i^2 d_i$ is a computable non-negative predictable sequence. By Lemma 5.2.2, applied to the computable M-martingale S,

$$\sum_{i=1}^{\infty} V_i^2 d_i < \infty \implies \sum_{i=1}^{\infty} V_i X_i \text{ converges},$$

for any M-typical sequence ξ. Also, since

$$\sum_{i=1}^{\infty} V_i^2 d_i = \sum_{j=1}^{\infty} \frac{1}{j^2}(d_{n_{j-1}+1} + \cdots + d_{n_j}),$$

and

$$\sum_{i=1}^{\infty} V_i X_i \text{ converges} \implies \sum_{j=1}^{\infty} \frac{1}{j}(X_{n_{j-1}+1} + \cdots + X_{n_j}) \text{ converges},$$

we can write

$$\sum_{j=1}^{\infty} \frac{1}{j^2}(d_{n_{j-1}+1} + \cdots + d_{n_j}) < \infty \implies \sum_{j=1}^{\infty} \frac{1}{j}(X_{n_{j-1}+1} + \cdots + X_{n_j}) \text{ converges}.$$

So, by Kronecker's lemma,

$$\sum_{j=1}^{\infty} \frac{Y_j}{j} \text{ converges} \implies \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} Y_j = 0.$$

Q.E.D.

A simple example of this theorem is given when $d_i = 1$, $i = 1, 2, \ldots$. In this case, the theorem states that for every M-typical sequence ξ,

$$\sum_{j=1}^{\infty} \frac{n_j - n_{j-1}}{j^2} < \infty \implies \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k}(X_{n_{j-1}+1} + \cdots + X_{n_j}) = 0,$$

and so, the limit is guaranteed for every M-typical sequence, if n_j is not growing too fast.

We now give a variant of the strong law of large numbers involving a sampled martingale transform.

Theorem 5.3.3 *Consider the computable basic martingale*

$$M_n = \begin{bmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n (X_i^2 - d_i) \end{bmatrix},$$

where X is a computable stochastic sequence and d is a computable non-negative predictable sequence. Let V be a computable predictable sequence and $\{n_k\}$ be a computable predictable subsequence. Then for every M-typical sequence ξ,

$$\sum_{j=1}^{\infty} \frac{(V_{n_{j-1}+1}^2 d_{n_{j-1}+1} + \cdots + V_{n_j}^2 d_{n_j})}{j^2} < \infty \implies \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} (V_{n_{j-1}+1} X_{n_{j-1}+1} + \cdots + V_{n_j} X_{n_j}) = 0.$$

Proof. Let ξ be an M-typical sequence and consider the computable stochastic sequence

$$N_n = \begin{bmatrix} \sum_{i=1}^n V_i X_i \\ \sum_{i=1}^n V_i^2 (X_i^2 - d_i) \end{bmatrix}.$$

Since the components of N are computable M-martingales, ξ is also N-typical and so Lemma 5.1.1, Lemma 5.2.1 and Lemma 5.2.2 still hold for ξ when N is taken to be the basic martingale. Then by an application of the strong law of large numbers of Theorem 5.3.2 with N as the basic martingale we have the result. **Q.E.D.**

We end this section by considering a strong law of large numbers a bit more general than Theorem 5.2.1, which involves a slightly more general version of Kronecker's lemma. Even for this law, we could have variants of it involving martingale transforms and subsequences, but we will not go into their detailed presentation.

Theorem 5.3.4 *Consider the computable basic martingale*

$$M_n = \begin{bmatrix} \sum_{i=1}^n (X_i - e_i) \\ \sum_{i=1}^n ((X_i - e_i)^2 - d_i) \end{bmatrix},$$

where X is a computable stochastic sequence, e is a computable predictable sequence and d is a computable non-negative predictable sequence. Let b be a computable non-decreasing positive predictable sequence. For every M-typical sequence ξ, for which $\lim_{n\to\infty} b(\xi^n) = \infty$,

$$\sum_{i=1}^{\infty} \frac{d(\xi^i)}{b^2(\xi^i)} < \infty \implies \lim_{n\to\infty} \frac{1}{b(\xi^n)} \sum_{i=1}^{n} (X(\xi^i) - e(\xi^i)) = 0.$$

Proof. Consider the computable M-martingale

$$S_n = \sum_{i=1}^{n} \frac{(X_i - e_i)}{b_i},$$

and the computable stochastic sequence

$$\begin{aligned} S_n^2 &= \Big(\sum_{i=1}^{n} \frac{(X_i - e_i)}{b_i}\Big)^2 \\ &= \Big[\sum_{i=1}^{n} \frac{((X_i - e_i)^2 - d_i)}{b_i^2} + 2\sum_{i>j}^{n} \frac{(X_i - e_i)}{b_i}\frac{(X_j - e_j)}{b_j}\Big] + \sum_{i=1}^{n} \frac{d_i}{b_i^2}. \end{aligned}$$

Since the term in square brackets is a computable M-martingale, S^2 is a computable M-submartingale. Let ξ be an M-typical sequence such that $\lim_{n\to\infty} b(\xi^n) = \infty$. By Lemma 5.2.2, applied to the computable M-martingale S,

$$\sum_{i=1}^{\infty} \frac{d(\xi^i)}{b^2(\xi^i)} < \infty \implies \sum_{i=1}^{\infty} \frac{(X(\xi^i) - e(\xi^i))}{b(\xi^i)} \text{ converges.}$$

Then, by Kronecker's lemma (Révész, 1968, Theorem 1.2.2), applied to this last series of real numbers,

$$\sum_{i=1}^{\infty} \frac{(X(\xi^i) - e(\xi^i))}{b(\xi^i)} \text{ converges} \implies \lim_{n\to\infty} \frac{1}{b(\xi^n)} \sum_{i=1}^{n} (X(\xi^i) - e(\xi^i)) = 0.$$

Q.E.D.

5.4 A Calibration Theorem

Consider the following sequential situation. On each day $(i-1)$, $i = 1, 2, \ldots$, a forecaster gives his probability p_i of an event A_i that will become known on the following day. Denoting with X_i the indicator of A_i, it is assumed that $p_i = P(A_i|\mathcal{F}_{i-1}) = \mathrm{E}(X_i|\mathcal{F}_{i-1})$, where $A_i \in \mathcal{F}_i$. That is, that the issued probability forecasts are the appropriate conditional probabilities with respect to a fixed probability distribution P defined over a σ-algebra $\mathcal{F} = \bigcup_{i=0}^{\infty} \mathcal{F}_i$, where $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. Let V_i be a sequence of indicator variables, such that V_i is $\mathcal{F}_{i-1}$-measurable, representing a selection rule which picks out day i if $V_i = 1$, and otherwise if $V_i = 0$, and let

$$\nu_n = \sum_{i=1}^{n} V_i, \qquad \overline{a}'_n = \frac{1}{\nu_n} \sum_{i=1}^{n} V_i X_i, \qquad \overline{p}'_n = \frac{1}{\nu_n} \sum_{i=1}^{n} V_i p_i.$$

Then Dawid (1982), using a martingale argument, showed that, with P-probability one, if $\nu_n \to \infty$, as $n \to \infty$, then $\overline{a}'_n - \overline{p}'_n \to 0$, as $n \to \infty$.

In our algorithmic framework, this result is embodied in the following corollary.

Corollary 5.4.1 *Consider the computable basic martingale*

$$M_n = \sum_{i=1}^{n}(X_i - p_i),$$

where X is a binary computable stochastic sequence with values in $\{0,1\}$, and p is a computable predictable sequence with values in $[0,1]$. Let V be a computable predictable sequence with values in $\{0,1\}$ and let $\nu_n = \sum_{i=1}^{n} V_i$. Then for every M-typical sequence ξ such that $\lim_{n\to\infty} \nu(\xi^n) = \infty$,

$$\lim_{n\to\infty} \frac{1}{\nu(\xi^n)} \sum_{i=1}^{n} V(\xi^i)(X(\xi^i) - p(\xi^i)) = 0.$$

Proof. Let ξ be an M-typical sequence. Consider the computable stochastic sequence

$$N_n = \begin{bmatrix} \sum_{i=1}^{n} V_i(X_i - p_i) \\ \sum_{i=1}^{n} V_i^2((X_i - p_i)^2 - p_i(1-p_i)) \end{bmatrix},$$

and note that

$$\sum_{i=1}^{n}((X_i - p_i)^2 - p_i(1-p_i)) = \sum_{i=1}^{n}(1-2p_i)(X_i - p_i),$$

is a computable M-martingale. Then ξ is also N-typical and, if $\lim_{n\to\infty} \nu(\xi^n) = \infty$, by applying Theorem 5.3.4 taking N as the basic martingale and $b_n = \nu_n$, since

$$\sum_{i=1}^{\infty} \frac{V^2(\xi^i)p(\xi^i)(1-p(\xi^i))}{\nu^2(\xi^i)} \leq \sum_{j=1}^{\infty} \frac{0.25}{j^2} < \infty,$$

we have that

$$\lim_{n\to\infty} \frac{1}{\nu(\xi^n)} \sum_{i=1}^{n}(V(\xi^i)X(\xi^i) - V(\xi^i)p(\xi^i)) = 0.$$

Q.E.D.

5.5 A Classical Refinement

In this section we will state a result which, even though a direct consequence of the variants of Section 5.3, can be seen as refining the strong law of large numbers towards the law of the iterated logarithm. Let us first consider an example. Take Theorem 5.3.2 with $d_i = 1$, $i = 1, 2, \ldots$, and $\{n_k\} = \lfloor k^{1.5} \rfloor$. Then for every M-typical sequence this theorem says that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{\lfloor k^{1.5} \rfloor} X_i = 0,$$

which is equivalent to

$$\lim_{J\to\infty} \frac{1}{J^{2/3}} \sum_{i=1}^{J} X_i = \lim_{J\to\infty} J^{1/3} \frac{1}{J} \sum_{i=1}^{J} X_i = 0,$$

and which is strictly stronger than the assertion that we could derive from the strong law of large numbers of Theorem 5.2.1.

On these lines, a more general result is given by the following corollary.

Corollary 5.5.1 *Consider the computable basic martingale*

$$M_n = \begin{bmatrix} \sum_{i=1}^{n} (X_i - e_i) \\ \sum_{i=1}^{n} ((X_i - e_i)^2 - d_i) \end{bmatrix},$$

where X is a computable stochastic sequence, e is a computable predictable sequence, and d is a computable non-negative predictable sequence such that $|d_i| \leq k$. For every M-typical sequence ξ, and any rational $\varepsilon > 0$,

$$n^{\frac{1}{2}-\varepsilon} \left[\frac{\sum_{i=1}^{n} (X(\xi^i) - e(\xi^i))}{n} \right] = \frac{\sum_{i=1}^{n} (X(\xi^i) - e(\xi^i))}{n^{\frac{1}{2}+\varepsilon}} \longrightarrow 0,$$

as $n \to \infty$.

Proof. Since

$$\sum_{i=1}^{\infty} \frac{d(\xi^i)}{b^2(\xi^i)} \leq k \sum_{i=1}^{\infty} \frac{1}{i^{1+2\varepsilon}} < \infty,$$

and $n^{\frac{1}{2}+\varepsilon}$ is computable, by an application of Theorem 5.3.4 with $b_n = n^{\frac{1}{2}+\varepsilon}$ we have the result. **Q.E.D.**

This refinement of the strong law of large numbers can be seen as an analogue of the classical result of Marcinkiewicz and Zygmund (1937a), proved in Kolmogorov's axiomatics for independent and identically distributed random variables. For a statement of this classical result more similar to our corollary, see however Révész (1968, Theorem 2.8.1).

5.6 The Law of the Iterated Logarithm

In the classical probability setting, let $X_1, X_2, \ldots$ be a sequence of independent, not necessarily identically distributed, random variables with $\mathrm{E}(X_i) = 0$ and finite variance, and let $S_n = \sum_{i=1}^{n} X_i$. Kolmogoroff (1929) proved that, if, as $n \to \infty$, $V_n = \mathrm{Var}(S_n) \to \infty$, and

$$|X_n| \leq \varepsilon_n \sqrt{\frac{V_n}{\ln \ln V_n}},$$

almost surely, for some constants $\varepsilon_n \to 0$, then

$$\limsup_{n \to \infty} \frac{S_n}{\sqrt{2 V_n \ln \ln V_n}} = 1,$$

almost surely. This remarkable result provided the best possible refinement of the strong law of large numbers. Later it was noted by Marcinkiewicz and Zygmund (1937b) that if the constants $\varepsilon_n \to 0$ are replaced by a constant $\varepsilon > 0$ the conclusion fails. Hartman and Wintner (1941) showed that, for $X_1, X_2, \ldots$, independent and identically distributed random variables, such that $\mathrm{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$,

$$\limsup_{n \to \infty} \frac{S_n - n\mu}{\sigma^2 \sqrt{2n \ln \ln n}} = 1,$$

almost surely. Results of this kind are referred to in the literature as laws of the iterated logarithm. Under Kolmogorov's probability axiomatics, martingale versions of Kolmogorov's, and Hartman and Wintner's law of the iterated logarithm were obtained by Stout (1970a, 1970b). (See also, for another martingale version of Kolmogorov's law of the iterated logarithm, Stout (1974, Theorem 5.4.1).)

In this section, we will consider a version for M-typical sequences of the upper half of the law of the iterated logarithm in the case of a binary basic martingale. In

proving our result, we follow the proof of the martingale extension of the upper half of Kolmogorov's law of the iterated logarithm given by Vovk (1990a, 1990b, 1991) in his prequential probability framework, which is based on an idea originally due to Ville (1939, ch. V, 1re Section, § 3).

Before considering the theorem itself, we note that, when each observation is binary, specifying a univariate basic martingale is essentially equivalent to specifying a full probability distribution P over Ω^∞. For a univariate binary computable basic martingale $M_n = \sum_{i=1}^n X_i$, $X_i \in \{a_i, b_i\}$, where a and b are a negative and a positive computable predictable sequence respectively, the higher powers of X_i can all be given by

$$\begin{aligned} X_i^2 &= -a_i b_i + (a_i + b_i) X_i, \\ X_i^3 &= -a_i b_i (a_i + b_i) + (a_i^2 + a_i b_i + b_i^2) X_i, \\ &\vdots \end{aligned}$$

and any martingale-like property involving these higher powers can be written as a martingale transform. For instance,

$$\sum_{i=1}^n (X_i^2 + a_i b_i) = \sum_{i=1}^n V_i X_i,$$

where the right-hand side is a computable martingale transform involving the computable predictable sequence $V_i = a_i + b_i$. This fact has been used to adapt the proof of Theorem 4 of Vovk (1990a) to prove the following law of the iterated logarithm.

Theorem 5.6.1 *Let us consider the binary computable basic martingale $M_n = \sum_{i=1}^n X_i$, $X_i \in \{a_i, b_i\}$, where a and b are a negative and a positive computable predictable sequence respectively. Let also $V_n = -\sum_{i=1}^n a_i b_i$. Then for every M-typical sequence ξ such that, as $n \to \infty$,*

$$V(\xi^n) \to \infty, \tag{5.1}$$

$$\Delta(\xi^n) = \max_\omega |X(\xi^{n-1} * \omega)| = o\left(\sqrt{\frac{V(\xi^n)}{\ln \ln V(\xi^n)}}\right), \tag{5.2}$$

we have that

$$\limsup_{n \to \infty} \frac{M(\xi^n)}{\sqrt{2 V(\xi^n) \ln \ln V(\xi^n)}} \leq 1.$$

Proof. Let $\xi \in \Omega^\infty$ be an M-typical sequence for which condition (5.1) and (5.2) are satisfied. We shall prove that, for an arbitrarily small $\varepsilon > 0$, from some n on

$$M(\xi^n) \leq (1+\varepsilon)\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}.$$

For a rational $\alpha > 0$, define the non-negative stochastic sequence $S^{(\alpha)}$ given by $S_0^{(\alpha)} = 1$ and

$$S_n^{(\alpha)} = \prod_{i=1}^{n} \frac{e^{\alpha X_i}}{\frac{1}{b_i - a_i}(b_i e^{\alpha a_i} - a_i e^{\alpha b_i})} = \prod_{i=1}^{n} \left(1 + \frac{e^{\alpha b_i} - e^{\alpha a_i}}{b_i e^{\alpha a_i} - a_i e^{\alpha b_i}} X_i\right),$$

for $n = 1, 2, \ldots$. It is easy to check that

$$S_n^{(\alpha)} = 1 + \sum_{i=1}^{n} U_i^{(\alpha)} X_i, \quad \text{where} \quad U_i^{(\alpha)} = S_{i-1}^{(\alpha)} \frac{e^{\alpha b_i} - e^{\alpha a_i}}{b_i e^{\alpha a_i} - a_i e^{\alpha b_i}},$$

and so, that $S^{(\alpha)}$ is an M-martingale. Moreover, since a and b are two computable sequences, the predictable sequence $U^{(\alpha)}$ and the M-martingale $S^{(\alpha)}$ are also computable.

Now, let $\varepsilon > 0$ be arbitrarily small, let $\delta > 0$ be a fixed rational small compared to ε, and consider the non-negative stochastic sequence

$$S_n = c_1 \sum_{k=1}^{\infty} \frac{1}{k^{1+\delta}} S_n^{(\alpha(k))},$$

where $\alpha(k)$ are rational approximations, to a given precision, to the computable function

$$\sqrt{2(1+\delta)^{-k}\ln k},$$

and $c_1 > 0$ is a rational chosen so that to ensure that $0 < S_0 \leq 1$. It is easy to check that, since for every $k = 1, 2, \ldots$, $S^{(\alpha(k))}$ is an M-martingale, also S is an M-martingale. As far as the computability of S is concerned, note first of all that both the quantities $\sqrt{2(1+\delta)^{-k}\ln k}$ and $k^{-(1+\delta)}$ are computable functions in k. Then, given an algorithm computing the function $\sqrt{2(1+\delta)^{-k}\ln k}$, whatever the precision of its rational approximations $\alpha(k)$, the sequence of stochastic sequences defined by

$$\frac{1}{k^{1+\delta}} S_n^{(\alpha(k))}, \qquad k = 1, 2, \ldots,$$

is a computable sequence of computable functions, and since each one of them is non-negative and for every given n the series in k converges, we have that S is computable.

For the M-typical sequence ξ, by Lemma 5.1.1, the limit $\lim_{n\to\infty} S(\xi^n)$ exists and is finite, that is, there exists a constant c_2 such that $S(\xi^n) \leq c_2$, for every $n = 1, 2, \ldots$. So, since S is a sum of non-negative stochastic sequences, for every $k = 1, 2, \ldots$,

$$c_1 \frac{1}{K^{1+\delta}} S^{(\alpha(k))}(\xi^n) \leq c_2, \qquad n = 1, 2, \ldots,$$

or $S^{(\alpha(k))}(\xi^n) \leq c_3 k^{1+\delta}$, $n = 1, 2, \ldots$. (The constants c_1, c_2, c_3 above and c_4 below depend only on δ.)

Let us consider now a sufficiently large n and choose the non-negative M-martingale $S^{(\alpha(k))}$ where $\alpha(k)$ is sufficiently close to $\sqrt{2(1+\delta)^{-k} \ln k}$ and

$$k = \lfloor \log_{1+\delta} V_n \rfloor.$$

Then

$$\ln S_n^{(\alpha(k))} \leq (1+\delta) \ln k + \ln c_3 = (1+\delta) \ln \ln V_n + c_4,$$

and, by the definition of $S^{(\alpha)}$,

$$\ln S_n^{(\alpha)} = \alpha M_n - \sum_{i=1}^{n} \ln \left(\frac{b_i e^{\alpha a_i} - a_i e^{\alpha b_i}}{b_i - a_i} \right),$$

and so the last inequality is equivalent to

$$\alpha M_n \leq \sum_{i=1}^{n} \ln \left(\frac{b_i e^{\alpha a_i} - a_i e^{\alpha b_i}}{b_i - a_i} \right) + (1+\delta) \ln \ln V_n + c_4,$$

where $\alpha = \alpha(k)$. Using the inequalities

$$e^t \leq 1 + t + \frac{t^2}{2} e^{|t|}, \qquad \ln(1+t) \leq t,$$

we obtain

$$\begin{aligned} \alpha M_n &\leq \sum_{i=1}^{n} \left(\frac{b_i \frac{\alpha^2 a_i^2}{2} e^{\alpha |a_i|} - a_i \frac{\alpha^2 b_i^2}{2} e^{\alpha |b_i|}}{b_i - a_i} \right) + (1+\delta) \ln \ln V_n + c_4 \\ &\leq \frac{\alpha^2}{2} + \sum_{i=1}^{n} e^{\alpha \Delta_i} (-a_i b_i) + (1+\delta) \ln \ln V_n + c_4, \end{aligned}$$

and, defining $\Delta_n^* = \max_{i \le n} \Delta_i$,

$$\alpha M_n \le \frac{\alpha^2}{2} V_n e^{\alpha \Delta_n^*} + (1+\delta) \ln \ln V_n + c_4.$$

Now, since $V_n \to \infty$, as $n \to \infty$, if n had been chosen sufficiently large, and the rational $\alpha(k)$ computed with sufficient accuracy,

$$\frac{1}{1+\delta} \sqrt{\frac{2 \ln \ln V_n}{V_n}} \le \alpha(k) \le (1+\delta) \sqrt{\frac{2 \ln \ln V_n}{V_n}}.$$

Also, since $\Delta_n = o((V_n / \ln \ln V_n)^{1/2})$, for sufficiently large n,

$$\Delta_n^* = \max_{i \le n} \Delta_i \le \delta \sqrt{\frac{V_n}{\ln \ln V_n}}.$$

Thus, we finally get

$$\begin{aligned} M_n &\le \frac{\alpha}{2} V_n e^{\alpha \Delta_n^*} + \frac{1}{\alpha}(1+\delta) \ln \ln V_n + \frac{c_4}{\alpha} \\ &\le \sqrt{2 V_n \ln \ln V_n} \left(\frac{1}{2}(1+\delta) e^{(1+\delta)\delta\sqrt{2}} + \frac{1}{2}(1+\delta)^2 + \frac{c_4(1+\delta)}{2 \ln \ln V_n} \right), \end{aligned}$$

and for sufficiently small δ this implies

$$M_n \le \sqrt{2 V_n \ln \ln V_n}\ (1+\varepsilon).$$

Q.E.D.

The following theorem gives the complete statement of the upper half of the law of the iterated logarithm for binary basic martingales.

Theorem 5.6.2 *Let us consider the binary computable basic martingale* $M_n = \sum_{i=1}^n X_i$, $X_i \in \{a_i, b_i\}$, *where* a *and* b *are a negative and a positive computable predictable sequence respectively. Let also* $V_n = -\sum_{i=1}^n a_i b_i$ *and consider an* M*-typical sequence* ξ *such that, as* $n \to \infty$,

$$V(\xi^n) \to \infty, \qquad \Delta(\xi^n) = \max_{\omega} |X(\xi^{n-1} * \omega)| = o\left(\sqrt{\frac{V(\xi^n)}{\ln \ln V(\xi^n)}} \right).$$

Then

$$\limsup_{n \to \infty} \frac{|M(\xi^n)|}{\sqrt{2 V(\xi^n) \ln \ln V(\xi^n)}} \le 1.$$

Proof. By the previous theorem we have

$$\limsup_{n\to\infty} \frac{M(\xi^n)}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \le 1.$$

Consider the stochastic sequences $X' = -X$ and $M' = -M$. Since for any arbitrary M'-martingale R

$$R_n = R_0 + \sum_{i=1}^{n} V_i X_i' = R_0 + \sum_{i=1}^{n} (-V_i) X_i,$$

any lower semicomputable non-negative M'-martingale is also a lower semicomputable non-negative M-martingale, and so, the M-typical sequence ξ is also M'-typical. Since $X_i' \in \{a_i', b_i'\}$, where $a_i' = -a_i > 0$, $b_i' = -b_i < 0$, we have that

$$V_n' = \sum_{i=1}^{n} a_i' b_i' = -\sum_{i=1}^{n} a_i b_i = V_n,$$
$$\Delta'(\xi^n) = \max_{\omega} |X'(\xi^{n-1} * \omega)| = \Delta(\xi^n),$$

and by another application of the previous theorem to M' we also have

$$\limsup_{n\to\infty} \frac{M'(\xi^n)}{\sqrt{2V'(\xi^n)\ln\ln V'(\xi^n)}} = \limsup_{n\to\infty} \frac{-M(\xi^n)}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \le 1.$$

Thus,

$$\limsup_{n\to\infty} \frac{|M(\xi^n)|}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \le 1.$$

Q.E.D.

The next theorem gives a somewhat stronger assertion of the previous law of the iterated logarithm, which is achieved, following Vovk (1990b, Theorem 6), by a slight modification in the proof. Now the condition on the order of magnitude of X_n is weakened.

Theorem 5.6.3 *Let us consider the binary computable basic martingale* $M_n = \sum_{i=1}^{n} X_i$, $X_i \in \{a_i, b_i\}$, *where* a *and* b *are a negative and a positive computable predictable sequence respectively. Let also*

$$V_n = -\sum_{i=1}^{n} a_i b_i, \qquad W_n = \sum_{i=1}^{n} \frac{b_i |a_i|^3 - a_i |b_i|^3}{b_i - a_i},$$

and consider an M-typical sequence ξ such that, as $n \to \infty$,

$$V(\xi^n) \to \infty, \qquad \Delta(\xi^n) = \max_\omega |X(\xi^{n-1} * \omega)| = O\left(\sqrt{\frac{V(\xi^n)}{\ln\ln V(\xi^n)}}\right),$$
$$W(\xi^n) = o\left(\sqrt{\frac{V^3(\xi^n)}{\ln\ln V(\xi^n)}}\right).$$

Then

$$\limsup_{n\to\infty} \frac{|M(\xi^n)|}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \le 1.$$

Proof. Replacing in the previous proof the inequality

$$e^t \le 1 + t + \frac{t^2}{2} e^{|t|},$$

by

$$e^t \le 1 + t + \frac{t^2}{2} + \frac{|t|^3}{6} e^{|t|},$$

we obtain

$$\begin{aligned} \alpha M_n &\le \sum_{i=1}^n \ln\left(\frac{b_i e^{\alpha a_i} - a_i e^{\alpha b_i}}{b_i - a_i}\right) + (1+\delta)\ln\ln V_n + c_4 \\ &\le \frac{\alpha^2}{2} V_n + \frac{\alpha^3}{6} \sum_{i=1}^n e^{\alpha\Delta_i} \frac{b_i|a_i|^3 - a_i|b_i|^3}{b_i - a_i} + (1+\delta)\ln\ln V_n + c_4, \end{aligned}$$

and, for $\Delta_n^* = \max_{i\le n} \Delta_i$, as before,

$$\alpha M_n \le \frac{\alpha^2}{2} V_n + \frac{\alpha^3}{6} e^{\alpha\Delta_n^*} W_n + (1+\delta)\ln\ln V_n + c_4.$$

Then, since $\Delta_n = O((V_n/\ln\ln V_n)^{1/2})$ and $W_n = o((V_n^3/\ln\ln V_n)^{1/2})$, for sufficiently large n,

$$\frac{\Delta_n}{\sqrt{\frac{V_n}{\ln\ln V_n}}} < B, \quad \text{and} \quad W_n \le \delta\sqrt{\frac{V_n^3}{\ln\ln V_n}},$$

for some B. Thus, we get

$$\begin{aligned} M_n &\le \frac{\alpha}{2} V_n + \frac{\alpha^2}{6} e^{\alpha\Delta_n^*} W_n + (1+\delta)\ln\ln V_n + \frac{c_4}{\alpha} \\ &\le \sqrt{2V_n \ln\ln V_n}\left(\frac{1}{2}(1+\delta) + \frac{\sqrt{2}}{6}(1+\delta)^2\delta e^{(1+\delta)B\sqrt{2}} + \frac{1}{2}(1+\delta)^2 + \frac{c_4(1+\delta)}{2\ln\ln V_n}\right), \end{aligned}$$

which for sufficiently small δ implies

$$M_n \le \sqrt{2V_n \ln\ln V_n}\ (1+\varepsilon),$$

that is,

$$\limsup_{n\to\infty} \frac{M(\xi^n)}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \leq 1.$$

Also, considering as before the stochastic sequences $X' = -X$ and $M' = -M$, and that

$$W'_n = \sum_{i=1}^{n} \frac{b'_i|a'_i|^3 - a'_i|b'_i|^3}{b'_i - a'_i} = \sum_{i=1}^{n} \frac{b_i|a_i|^3 - a_i|b_i|^3}{b_i - a_i} = W_n,$$

by applying the obtained result to M' we get

$$\limsup_{n\to\infty} \frac{-M(\xi^n)}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \leq 1,$$

and the conclusion then follows. **Q.E.D.**

5.7 Variants of the Law of the Iterated Logarithm

In the laws of the iterated logarithm of the previous section we considered a binary computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, $X_i \in \{a_i, b_i\}$, where a and b are a negative and a positive computable predictable sequence respectively. Of course, we could have just assumed that $a_i \neq 0$, $b_i \neq 0$, and that, for every i, a_i and b_i are opposite in sign, without wondering about which one is negative and which one is positive. Easily, given an M-typical sequence ξ with respect to a basic martingale M for which the latter assumption is true, we have that ξ is also typical with respect to a computable stochastic sequence for which the former assumption is true.

The situation, however, is completely different if we allow a_i and/or b_i to be zero. In this case, the M-martingales $S^{(\alpha)}$ defined by $S_0^{(\alpha)} = 1$,

$$S_n^{(\alpha)} = \prod_{i=1}^{n} \frac{e^{\alpha X_i}}{\frac{1}{b_i - a_i}(b_i e^{\alpha a_i} - a_i e^{\alpha b_i})} = \prod_{i=1}^{n} \left(1 + \frac{e^{\alpha b_i} - e^{\alpha a_i}}{b_i e^{\alpha a_i} - a_i e^{\alpha b_i}} X_i\right),$$

$n = 1, 2, \ldots$, would not be computable any more, since the denominators could be equal to zero. To guarantee the existence of an algorithm giving rational approximations to $S^{(\alpha)}$ to any desired precision we would need to treat separately the case in which a_i and b_i are equal to zero. This will be possible by considering the following stronger computability assumption.

Definition 5.7.1 *A stochastic sequence* $S:\Omega^* \to \mathbf{R}$ *is strongly computable if there exists an algorithm which when fed with* (ξ^n, m) *yields the symbol* $\oslash$ *if* $S(\xi^n) = 0$ *and an approximation to* 2^{-m} *if* $S(\xi^n) \neq 0$.

With this definition we can now give the following result.

Theorem 5.7.1 *Let us consider the binary computable basic martingale* $M_n = \sum_{i=1}^n X_i$, *where* $X_i \in \{a_i, b_i\}$, $a_i \leq 0$, $b_i \geq 0$, *is a strongly computable stochastic sequence. Let also* $V_n = -\sum_{i=1}^n a_i b_i$. *Then for every* M*-typical sequence* ξ *such that, as* $n \to \infty$, *condition (5.1) and (5.2) are true, we have that*

$$\limsup_{n\to\infty} \frac{M(\xi^n)}{\sqrt{2V(\xi^n)\ln\ln V(\xi^n)}} \leq 1.$$

Proof. The proof is the same as in Theorem 5.6.1, but for the following note. We set

$$\frac{1}{b_i - a_i}(b_i e^{\alpha a_i} - a_i e^{\alpha b_i}) = 1,$$

when either a_i or b_i, or both, are equal to zero, so that $S^{(\alpha)}$ is now defined as $S_0^{(\alpha)} = 1$,

$$S_n^{(\alpha)} = \begin{cases} S_{n-1}^{(\alpha)} \dfrac{e^{\alpha X_n}}{\frac{1}{b_n - a_n}(b_n e^{\alpha a_n} - a_n e^{\alpha b_n})}, & a_n < 0,\ b_n > 0, \\ \\ S_{n-1}^{(\alpha)}\ e^{\alpha X_n}, & a_n,\ b_n, \text{ or both, are zero,} \end{cases}$$

$n = 1, 2, \ldots$. In this way, due to the strong computability of X, the M-martingales $S^{(\alpha)}$ are still computable and the previous proof remains perfectly valid. **Q.E.D.**

Let us note that assuming X to be a strongly computable stochastic sequence is similar to assume, as in Vovk (1988a), under Kolmogorov's axiomatics, a strongly computable probability distribution P, in the sense that the set $P^{-1}\{0\}$ is decidable.

Making the same stronger computability assumption about X, we similarly have that the conclusions of Theorem 5.6.2 and Theorem 5.6.3 are still valid when $a_i \leq 0$, $b_i \geq 0$. Then we can have the following result for a generic martingale transform.

Theorem 5.7.2 *Let us consider the binary computable basic martingale* $M_n = \sum_{i=1}^n X_i$, *where* $X_i \in \{a_i, b_i\}$, $a_i \leq 0$, $b_i \geq 0$, *is a strongly computable stochastic sequence, and a strongly computable predictable sequence* U. *Let also* $V_n = -\sum_{i=1}^n U_i^2 a_i b_i$, *and consider an* M*-typical sequence* ξ *such that, as* $n \to \infty$,

$$V(\xi^n) \to \infty, \qquad \Delta(\xi^n) = \max_\omega |U(\xi^n) X(\xi^{n-1} * \omega)| = o\left(\sqrt{\frac{V(\xi^n)}{\ln \ln V(\xi^n)}}\right).$$

Then

$$\limsup_{n \to \infty} \frac{|\sum_{i=1}^n U(\xi^i) X(\xi^i)|}{\sqrt{2V(\xi^n) \ln \ln V(\xi^n)}} \leq 1.$$

Proof. Let $N_n = \sum_{i=1}^n U_i X_i$. Since N is a computable M-martingale, the M-typical sequence ξ is also N-typical, and the result follows by applying the law of the iterated logarithm, for $a_i \leq 0$ and $b_i \geq 0$, following from Theorem 5.6.2, taking N as the basic martingale. **Q.E.D.**

5.8 Sampled Martingales

In this section we will consider a strong law of large numbers which, even if it seems to be similar to the previous variants for subsequences of Section 5.3, has nevertheless some intrinsic algorithmic differences. An application of subsequent Lemma 5.8.3 will be seen in the next chapter.

Let us introduce the following notation. Fix a filtered space $(\Omega^\infty, (\mathcal{F}_0, \mathcal{F}_1, \ldots), \mathcal{F})$, with $\mathcal{F}_0 = \{\emptyset, \Omega^\infty\}$, where Ω is a subset of $\mathbf{Q}$ and the family of σ-algebras $(\mathcal{F}_0, \mathcal{F}_1, \ldots)$ is the filtration generated by the cylinder sets in Ω^∞. Also, fix a subsequence $\{n_k\}$ of $\{n\}$, and consider the induced filtered space $(\Omega^\infty, (\mathcal{F}_0, \mathcal{F}_{n_1}, \mathcal{F}_{n_2}, \ldots), \mathcal{F})$. For a given basic martingale $M_n = \sum_{i=1}^n X_i$, on the filtered space $(\Omega^\infty, (\mathcal{F}_0, \mathcal{F}_1, \ldots), \mathcal{F})$, we consider, on the induced filtered space $(\Omega^\infty, (\mathcal{F}_0, \mathcal{F}_{n_1}, \mathcal{F}_{n_2}, \ldots), \mathcal{F})$, the induced basic martingale M_{n_k}. For $\xi \in \Omega^\infty$ and $Y_j = X_{n_{j-1}+1} + \cdots + X_{n_j}$, we have

$$M_{n_k}(\xi^{n_k}) = \sum_{j=1}^k Y_j(\xi^{n_j}), \qquad Y_j(\xi^{n_j}) = \sum_{i=n_{j-1}+1}^{n_j} X_i(\xi^i).$$

We define M_{n_k}-martingales, M_{n_k}-submartingales, compensators and other quantities in an obvious manner. For instance, an M_{n_k}-martingale S is an $\mathcal{F}_{n_k}$-stochastic sequence (measurable with respect to $\mathcal{F}_{n_k}$) of the form

$$S_k = c + \sum_{j=1}^{k} V_j \cdot (M_{n_j} - M_{n_{j-1}}),$$

where c is a real number and V_j is an $\mathcal{F}_{n_k}$-predictable sequence.

With this terminology, we can now give our strong law of large numbers for sampled martingales following the steps of the proof of Theorem 5.2.1.

Lemma 5.8.1 *Consider a computable basic martingale $M_n = \sum_{i=1}^{n} X_i$ such that for every $n \in \mathbf{N}$, $\min(M_n|\mathcal{F}_{n-1}) \leq M_{n-1} \leq \max(M_n|\mathcal{F}_{n-1})$. Also, let $\{n_k\}$ be a computable subsequence and ξ be an M-typical sequence. If S is a computable non-negative M_{n_k}-martingale*

$$S_k = S_0 + \sum_{j=1}^{k} V_j \cdot (M_{n_j} - M_{n_{j-1}}),$$

with V_j a computable $\mathcal{F}_{n_k}$-predictable sequence and $S_0 > 0$, then $\lim_{k\to\infty} S_k(\xi^{n_k})$ exists and is finite.

Proof. Consider the $\mathcal{F}_n$-stochastic sequence R defined by

$$R_n = S_0 + \sum_{i=1}^{n} U_i \cdot X_i,$$

where $U_i = V_j$, $n_{j-1} < i \leq n_j$. This stochastic sequence is a computable M-martingale such that $R_0 > 0$, $R_{n_k} = S_k$, for all $k \in \mathbf{N}$. Since $\min(M_n|\mathcal{F}_{n-1}) \leq M_{n-1} \leq \max(M_n|\mathcal{F}_{n-1})$, for all $n \in \mathbf{N}$, also $\min(R_n|\mathcal{F}_{n-1}) \leq R_{n-1} \leq \max(R_n|\mathcal{F}_{n-1})$, for all $n \in \mathbf{N}$. So, from $R_{n_k} \geq 0$, for all k, we have that $R_{n_{k-1}}, R_{n_{k-2}}, \ldots$ are all non-negative, that is, that R is non-negative. Thus, by Lemma 5.1.1, the limit $\lim_{n\to\infty} R_n(\xi^n)$ exists and is finite, and since $R_{n_k} = S_k$, also the limit $\lim_{k\to\infty} S_k(\xi^{n_k})$ exists and is finite. **Q.E.D.**

Lemma 5.8.2 *Let M and $\{n_k\}$ be as in Lemma 5.8.1. If T is a computable non-negative M_{n_k}-submartingale, A is one of its computable compensators and the computable M_{n_k}-martingale $S = T - A$ is such that*

$$S_k = S_0 + \sum_{j=1}^{k} W_j \cdot (M_{n_j} - M_{n_{j-1}}),$$

for some computable $\mathcal{F}_{n_k}$-predictable sequence W_j, then for every M-typical sequence ξ,

$$A_\infty(\xi) < \infty \implies T_k(\xi^{n_k}) \text{ converges.}$$

Proof. The proof is the same as in Lemma 5.2.1. We just note that the stochastic sequence

$$S_k^{(C)} = 1 + \sum_{j=1}^{k} \frac{V_j}{S_0 + C}(S_j - S_{j-1}) = 1 + \sum_{j=1}^{k} \frac{V_j}{S_0 + C} W_j \cdot (M_{n_j} - M_{n_{j-1}}),$$

is a computable non-negative M_{n_k}-martingale, with $S_0^{(C)} = 1$, where $V_j W_j/(S_0 + C)$ is a computable $\mathcal{F}_{n_k}$-predictable sequence. Thus, by Lemma 5.8.1 the limit $\lim_{k\to\infty} S_k^{(C)}(\xi^{n_k})$ exists and is finite. **Q.E.D.**

Lemma 5.8.3 *Let M and $\{n_k\}$ be as in Lemma 5.8.1. If S is a computable M_{n_k}-martingale (obtained from M with a computable $\mathcal{F}_{n_k}$-predictable sequence), S^2 is a computable M_{n_k}-submartingale, A is one of its computable compensators ($S^2 - A$ is an M_{n_k}-martingale which can be obtained from M with a computable $\mathcal{F}_{n_k}$-predictable sequence), and ξ is an M-typical sequence, then*

$$A_\infty(\xi) < \infty \implies S_k(\xi^{n_k}) \text{ converges.}$$

Proof. The proof is the same as in Lemma 5.2.2. Here, we just note that

$$S_k^2 - A_k = N_k = N_0 + \sum_{j=1}^{k} V_j \cdot (M_{n_j} - M_{n_{j-1}}),$$

for some computable $\mathcal{F}_{n_k}$-predictable sequence V_j, and that

$$\begin{aligned}
(S_k + 1)^2 - A_k &= 1 + N_k + 2S_k \\
&= 1 + N_0 + \sum_{j=1}^{k} V_j \cdot (M_{n_j} - M_{n_{j-1}}) + 2\Big[S_0 + \sum_{j=1}^{k} W_j \cdot (M_{n_j} - M_{n_{j-1}})\Big] \\
&= (1 + N_0 + 2S_0) + \sum_{j=1}^{k} (V_j + 2W_j) \cdot (M_{n_j} - M_{n_{j-1}}),
\end{aligned}$$

for some computable $\mathcal{F}_{n_k}$-predictable sequence W_j. **Q.E.D.**

Theorem 5.8.1 *Let M be a computable basic martingale such that $\min(M_n|\mathcal{F}_{n-1}) \leq M_{n-1} \leq \max(M_n|\mathcal{F}_{n-1})$, for all $n \in \mathbf{N}$, and let $\sum_{i=1}^{n} X_i$ be a component of M, and $\{n_k\}$ be a computable subsequence. Assume that*

$$\sum_{j=1}^{k}(Y_j^2 - d_j), \quad Y_j = X_{n_{j-1}+1} + \cdots + X_{n_j},$$

is a computable M_{n_k}-martingale (which can be obtained from M with a computable $\mathcal{F}_{n_k}$-predictable sequence) where d is a computable non-negative $\mathcal{F}_{n_k}$-predictable sequence. Then for every M-typical sequence ξ,

$$\sum_{j=1}^{\infty}\frac{d_j}{j^2} < \infty \implies \lim_{k\to\infty}\frac{1}{k}\sum_{j=1}^{k} Y_j = 0.$$

Proof. Consider the computable M_{n_k}-martingale

$$S_k = \sum_{j=1}^{k}\frac{Y_j}{j},$$

and

$$S_k^2 = \Big(\sum_{j=1}^{k}\frac{Y_j}{j}\Big)^2 = \Big[\sum_{j=1}^{k}\frac{(Y_j^2 - d_j)}{j^2} + 2\sum_{j>r}^{k}\frac{Y_j}{j}\frac{Y_r}{r}\Big] + \sum_{j=1}^{k}\frac{d_j}{j^2}.$$

Since S_k^2 is a computable M_{n_k}-submartingale, by Lemma 5.8.3,

$$\sum_{j=1}^{\infty}\frac{d_j}{j^2} < \infty \implies \sum_{j=1}^{\infty}\frac{Y_j}{j} \text{ converges},$$

and by Kronecker's lemma we have the result. **Q.E.D.**

5.9 Discussion

Let us conclude this chapter by stressing some points about the results we have just presented. In Kolmogorov's probability axiomatics, the reformulation of almost sure results in algorithmic terms, by stating them for every random sequence, leads to an essential strengthment of the corresponding non-algorithmic theorems. To see

an example of this strengthment in our purely martingale framework, just take the case of the strong law of large numbers. Define the following two subsets of Ω^∞

$$K = \Big\{\xi : \sum_{i=1}^{\infty} \frac{d(\xi^i)}{i^2} < \infty\Big\}, \qquad L = \Big\{\xi : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X(\xi^i) = 0\Big\},$$

and consider the set T_M of all M-typical sequences. Then whereas the strong law of large numbers of Theorem 5.2.1 says that the set $T_M \cap K \cap (\Omega^\infty \backslash L)$ is empty, its non-algorithmic counterpart (Vovk, 1993c, Theorem 2) would just say that this set is M-null.

Another key feature of the algorithmic approach, in whatever probability framework, is that it naturally suggests results based on 'local conditions'. To show this point, consider, in Kolmogorov's probability axiomatics, a scalar local martingale $M_n = \sum_{i=1}^{n} X_i$, with respect to a filtration $(\mathcal{F}_n)_{n\geq 0}$, where $\mathrm{E}(X_i|\mathcal{F}_{i-1}) = 0$, $\mathrm{E}(X_i^2|\mathcal{F}_{i-1}) = \sigma_i^2$, and σ^2 is a predictable sequence. Then, for a predictable subsequence $\{n_k\}$, the sum $M_{n_k} = \sum_{j=1}^{k} Y_j$, where $Y_j = X_{n_{j-1}+1} + \cdots + X_{n_j}$, is a sampled martingale with $\mathrm{E}(Y_j|\mathcal{F}_{j-1}) = 0$ and

$$\mathrm{E}(Y_j^2|\mathcal{F}_{j-1}) = \sum_{i=n_{j-1}+1}^{n_j} \mathrm{E}(X_i^2|\mathcal{F}_{j-1}) = \sum_{i=n_{j-1}+1}^{n_j} \mathrm{E}(\sigma_i^2|\mathcal{F}_{j-1}),$$

and by the generalization of Kolmogorov's strong law of large numbers, considered at the beginning of Section 5.2, we have that,

$$\sum_{j=1}^{\infty} \frac{\mathrm{E}(Y_j^2|\mathcal{F}_{j-1})}{j^2} < \infty \implies \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} Y_j = 0,$$

almost surely. On the other hand, in our purely martingale algorithmic framework, the strong law of large numbers of Theorem 5.3.2 states that, for every M-typical sequence ξ,

$$\sum_{j=1}^{\infty} \frac{d_{n_{j-1}+1} + \cdots + d_{n_j}}{j^2} < \infty \implies \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} Y_j = 0,$$

and it is clearly evident, by comparing the latter condition on the cumulative variance with the former, that this last statement has a specific local character. In fact, this last condition involves only the conditional variances given the past of the martingale differences X_i, and not the conditional variances of the sampled martingale differences Y_j. Note, however, that the statement of the strong law of large

numbers for sampled martingales of Theorem 5.8.1 is based on a condition which has the same non-local character of the previous non-algorithmic formulation.

As far as the law of the iterated logarithm is concerned, let us remember that, in the usual probability axiomatics with respect to a probability distribution P, Vovk (1988b) proved it for chaotic sequences, whereas Vovk (1988a) proved it for typical sequences, using, in this latter result, martingale ideas similar to those used in the proof, in the prequential probability framework, of the law of the iterated logarithm of Vovk (1990a). We did not attempt to prove any lower half of the law of the iterated logarithm for M-typical sequences. Vovk (1990b, Theorem 7) proved, in the prequential probability framework, a variant of Kolmogorov's lower half of the law of the iterated logarithm, while Vovk (1991, Section 10) noted that the lower half of the law of the iterated logarithm seemingly cannot be formulated without loss in his 'finitary' prequential probability framework.

Chapter 6

Distributions of Values and Strong Central Limit Theorem

In this chapter we deal with the problem of characterizing the distribution of values corresponding to a single M-typical sequence in the case of stochastic sequences obtained from the basic martingale M. From the point of view of applications, the cases considered are fairly elementary and the basic martingale M will be essentially a coin-tossing process. In Section 6.1, we will consider the distributions of values for two stochastic sequences obtained from a symmetric Bernoulli stochastic sequence, namely a moving average and a first-order autoregressive stochastic sequence. Then in Sections 6.2, 6.3 and 6.4 we consider the strong central limit theorem, concerning the distribution of values, in logarithmic density, or in suitable subsequences, of a standardized sum statistic.

6.1 Distributions of Values

In this section we will consider the distributions of values, corresponding to a single M-typical sequence, of a martingale difference sequence, a moving average stochastic sequence, and a first-order autoregressive stochastic sequence. These results could be seen as the analogue of the strong consistency, in Kolmogorov's probability axiomatics, of the empirical distribution functions of the corresponding stochastic

processes. Throughout this section we always consider a scalar binary basic martingale difference sequence with only two possible fixed values, for example $\{-1,1\}$, so representing just a simple coin-tossing process. If the values of the scalar binary basic martingale difference sequence were not fixed (as in the case of the law of the iterated logarithm of Section 5.6) we could not guarantee, in general, the existence of a distribution of values for any given M-typical sequence. For instance, no sequence, M-typical or not, admits a distribution of values when the binary basic martingale difference sequence is such that $X_i \in \{-1,1\}$ for $i = 1,2$, $X_i \in \{-2,2\}$ for $i = (2+1),\ldots,(2+2^2)$, $X_i \in \{-1,1\}$ for $i = (2+2^2+1),\ldots,(2+2^2+2^3)$, and so forth. Note also that, if we were to deal with ternary observations, we would have to consider, to characterize distributions of values, multivariate basic martingales.

Let us start by considering the simple case of a scalar binary basic martingale difference sequence. Let $M_n = \sum_{i=1}^n X_i$ be a computable basic martingale such that $X_i \in \{a,b\}$, and $a < 0$, $b > 0$ are rational numbers. For every $z \in \mathbf{R}$, we have that

$$I_{(-\infty,z]}(X_i) = \begin{cases} 0, & z < a, \\ \frac{1}{b-a}(b - X_i), & a \leq z < b, \\ 1, & z \geq b, \end{cases}$$

and so, that

$$F_n(z) = \begin{cases} 0, & z < a, \\ \frac{1}{n}\sum_{i=1}^n I_{(-\infty,z]}(X_i) = \frac{b}{b-a} - \frac{1}{b-a}\frac{1}{n}\sum_{i=1}^n X_i, & a \leq z < b, \\ 1, & z \geq b. \end{cases}$$

Then, since any M-typical sequence ξ is also N-typical, where N is the computable stochastic sequence defined by

$$N_n = \begin{bmatrix} \sum_{i=1}^n X_i \\ (a+b)\sum_{i=1}^n X_i \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n (X_i^2 + ab) \end{bmatrix},$$

we have, by an application of the strong law of large numbers of Theorem 5.2.1 taking N as the basic martingale, that, for any M-typical sequence ξ, $F_n(z)(\xi^n) \to b/(b-a)$, $a \leq z < b$, as $n \to \infty$. We will examine now the case in which $a = -1$ and $b = 1$.

6.1.1 A Symmetric Bernoulli Stochastic Sequence

Let us consider the symmetric Bernoulli stochastic sequence given by a computable basic martingale $M_n = \sum_{i=1}^{n} X_i$ with $X_i \in \{-1, 1\}$. Let ξ be an M-typical sequence, $N \gg q$, $q \in \mathbf{N}$, and consider the following auxiliary finite probability space. Define the $2^{q+1} + 1$ sets

$$\begin{aligned}
E_{-\cdots-} &= \{n : n \le N,\ X_n = -1, X_{n-1} = -1, \ldots, X_{n-q} = -1\},\\
E_{-\cdots+} &= \{n : n \le N,\ X_n = -1, X_{n-1} = -1, \ldots, X_{n-q} = +1\},\\
&\vdots\\
E_{+\cdots+} &= \{n : n \le N,\ X_n = +1, X_{n-1} = +1, \ldots, X_{n-q} = +1\},\\
E_0 &= \{n : n \le N,\ X_m = 0 \text{ for some } m = n, n-1, \ldots, n-q\},
\end{aligned}$$

where $X_0 = X_{-1} = \cdots = X_{-q} = 0$. These sets are disjoint, their union is equal to the set $\{1, 2, \ldots, N\}$ of the first N natural numbers, and some of them could well be empty. We regard these sets formally. They are, empty or not, the 'points' of a finite space Λ_{q+1}. On this space we consider the algebra $\mathcal{A}_{q+1}$ generated by the above sets, and the measure μ giving equal probability $1/2^{q+1}$ to all points $E \in \Lambda_{q+1}$, but to E_0 to which it gives probability zero.

On the probability space $(\Lambda_{q+1}, \mathcal{A}_{q+1})$, define also the frequency measure ν_N given by

$$\begin{aligned}
\nu_N(E_{-\cdots-}) &= \frac{1}{N}\#\{n : n \le N,\ X_n = -1, X_{n-1} = -1, \ldots, X_{n-q} = -1\},\\
\nu_N(E_{-\cdots+}) &= \frac{1}{N}\#\{n : n \le N,\ X_n = -1, X_{n-1} = -1, \ldots, X_{n-q} = +1\},\\
&\vdots\\
\nu_N(E_{+\cdots+}) &= \frac{1}{N}\#\{n : n \le N,\ X_n = +1, X_{n-1} = +1, \ldots, X_{n-q} = +1\},\\
\nu_N(E_0) &= \frac{1}{N}\#\{n : n \le N,\ X_m = 0 \text{ for some } m = n, n-1, \ldots, n-q\}.
\end{aligned}$$

Lemma 6.1.1 *Let M be the symmetric Bernoulli stochastic sequence. Then, for any M-typical sequence ξ, the measures ν_N tend to the measure μ, uniformly over $\mathcal{A}_{q+1}$, as $N \to \infty$.*

Proof. To show this, we will consider in turn the sets E_-, E_+, then the sets E_{--}, E_{-+}, E_{+-}, E_{++}, and so forth up to the sets $E \in \Lambda_{q+1}$. Let us start with the sets E_- and E_+. Since $X_n \in \{-1,1\}$, we have that $\sum_{n=1}^N X_n = \#\{n : n \leq N,\ X_n = 1\} - \#\{n : n \leq N,\ X_n = -1\}$. So, since we can write

$$\begin{aligned}\nu_N(E_+) &= \frac{1}{N}\#\{n : n \leq N,\ X_n = 1\} = \frac{1}{2} + \frac{1}{N}\sum_{n=1}^N \frac{1}{2}X_n,\\ \nu_N(E_-) &= 1 - \nu_N(E_+) = \frac{1}{2} - \frac{1}{N}\sum_{n=1}^N \frac{1}{2}X_n,\end{aligned}$$

by the strong law of large numbers of Theorem 5.2.1 we have that, for the M-typical sequence ξ,

$$\nu_N(E_+) \to 1/2, \qquad \nu_N(E_-) \to 1/2,$$

as $N \to \infty$.

Let us consider the sets E_{++}, E_{-+}, E_{+-} and E_{--}. We define the computable predictable sequences

$$V_n^+ = \begin{cases} 1, & X_{n-1} = 1,\\ 0, & \text{otherwise,}\end{cases} \qquad V_n^- = \begin{cases} 1, & X_{n-1} = -1,\\ 0, & \text{otherwise.}\end{cases}$$

By using the predictable sequence V_n^+, similarly as before, we have that

$$\frac{1}{N}\sum_{n=1}^N V_n^+ X_n = \frac{1}{N}\Big(\#\{n : n \leq N, X_n = 1, X_{n-1} = 1\} - \#\{n : n \leq N, X_n = -1, X_{n-1} = 1\}\Big),$$

and so

$$\begin{aligned}\nu_N(E_{++}) &= \frac{1}{N}\#\{n : n \leq N, X_n = 1, X_{n-1} = 1\} = \frac{1}{4} + O\Big(\frac{1}{N}\Big) + \frac{1}{N}\sum_{n=1}^N \Big(\frac{1}{2}V_n^+ + \frac{1}{4}\Big)X_n,\\ \nu_N(E_{-+}) &= \frac{1}{N}\#\{n : n \leq N, X_n = -1, X_{n-1} = 1\} = \frac{1}{4} + O\Big(\frac{1}{N}\Big) + \frac{1}{N}\sum_{n=1}^N \Big(-\frac{1}{2}V_n^+ + \frac{1}{4}\Big)X_n.\end{aligned}$$

Also, by considering the predictable sequence V_n^-, similar expressions can be written for $\nu_N(E_{+-})$ and $\nu_N(E_{--})$, and by the strong law of large numbers for martingale transforms of Theorem 5.3.1 we have that, for the M-typical sequence ξ,

$$\nu_N(E_{++}) \to 1/4, \qquad \nu_N(E_{-+}) \to 1/4, \qquad \nu_N(E_{+-}) \to 1/4, \qquad \nu_N(E_{--}) \to 1/4,$$

as $N \to \infty$.

For the sets E_{+++}, E_{++-}, E_{+-+}, E_{+--}, E_{-++}, E_{-+-}, E_{--+}, E_{---}, by using the computable predictable sequences V_n^{++}, V_n^{+-}, V_n^{-+}, V_n^{--}, defined by

$$V_n^{++} = \begin{cases} 1, & X_{n-1}=1, X_{n-2}=1, \\ 0, & \text{otherwise}, \end{cases} \qquad V_n^{+-} = \begin{cases} 1, & X_{n-1}=1, X_{n-2}=-1, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{etc.},$$

we have, for instance, that

$$\begin{aligned} \nu_N(E_{+++}) &= \frac{1}{N}\#\{n : n \le N,\ X_n=1, X_{n-1}=1, X_{n-2}=1\} \\ &= \frac{1}{8} + O\Big(\frac{1}{N}\Big) + \frac{1}{N}\sum_{n=1}^{N}\Big(\frac{1}{2}V_n^{++} + \frac{1}{4}V_n^{+} + \frac{1}{8}\Big)X_n, \end{aligned}$$

and so, for the M-typical sequence ξ, $\nu_N(E_{+++}) \to 1/8$, as $N \to \infty$. And similarly for the other sets.

Proceeding in this way, we can show that, for the M-typical sequence ξ, for any set $E \in \Lambda_{q+1}$, but for E_0 for which $\nu_N(E_0) \to 0$, as $N \to \infty$, $\nu_N(E) \to 1/2^{q+1}$, as $N \to \infty$. Also, since Λ_{q+1} is finite, $\nu_N(A) \to \mu(A)$, as $N \to \infty$, uniformly for all sets A in the algebra $\mathcal{A}_{q+1}$, and thus, the desired result is proved. **Q.E.D.**

6.1.2 A Moving Average Stochastic Sequence

Let us consider now a simple moving average stochastic sequence. Let $M_n = \sum_{i=1}^{n} X_i$ be the symmetric Bernoulli stochastic sequence, and consider the computable stochastic sequence defined by

$$Y_n = X_n + \beta_1 X_{n-1} + \beta_2 X_{n-2} + \cdots + \beta_q X_{n-q}, \qquad n = 1, 2, \ldots, \tag{6.1}$$

where $X_0 = X_{-1} = \cdots = X_{-q} = 0$, and $\beta_1, \beta_2, \ldots, \beta_q$ are rational numbers.

Theorem 6.1.1 *For the moving average stochastic sequence (6.1), for any M-typical sequence ξ, as $N \to \infty$,*

$$\frac{1}{N}\#\{n : n \le N,\ Y(\xi^n) \le z\} \to \varphi(z),$$

uniformly in $z \in \mathbf{R}$, where $\varphi(z) = \Pr(Z_0 + Z_1 + \cdots + Z_q \le z)$, and $Z_0, Z_1, \ldots, Z_q$ are independent discrete random variables with $\Pr(Z_j = \beta_j) = \Pr(Z_j = -\beta_j) = 1/2$, $(\beta_0 = 1)$, $j = 0, 1, \ldots, q$.

Proof. Consider $(\Lambda_{q+1}, \mathcal{A}_{q+1})$, μ and ν_N as defined in Section 6.1.1. On the probability space $(\Lambda_{q+1}, \mathcal{A}_{q+1})$, define the random variables $Z_0, Z_1, \ldots, Z_q$ given by

$$Z_0(E) = \begin{cases} 1, & E = E_{+\cdots}, \\ 0, & E = E_0, \\ -1, & E = E_{-\cdots}, \end{cases} \quad , \cdots , \quad Z_q(E) = \begin{cases} \beta_q, & E = E_{\cdots+}, \\ 0, & E = E_0, \\ -\beta_q, & E = E_{\cdots-}. \end{cases}$$

With respect to the measure μ, these random variables are independent and such that $\mu(Z_j = \beta_j) = \mu(Z_j = -\beta_j) = 1/2$, $j = 0, 1, \ldots, q$.

Then since, for every $z \in \mathbf{R}$, we have that

$$\begin{aligned} \frac{1}{N}\#\{n{:}n \le N,\ Y(\xi^n) \le z\} &= \frac{1}{N}\#\Big\{n{:}n \le N,\ n \in \text{ some } E \text{ s.t. } \sum_{j=0}^{q} Z_j(E) \le z\Big\} \\ &= \nu_N\Big\{E{:}E \in \Lambda_{q+1}, \sum_{j=0}^{q} Z_j(E) \le z\Big\} \\ &= \nu_N\Big(\sum_{j=0}^{q} Z_j(E) \le z\Big), \end{aligned}$$

the desired result is proved since by Lemma 6.1.1 the measures ν_N tend to the measure μ, as $N \to \infty$, uniformly over $\mathcal{A}_{q+1}$. **Q.E.D.**

6.1.3 A First-Order Autoregressive Stochastic Sequence

Let $M_n = \sum_{i=1}^{n} X_i$ be the symmetric Bernoulli stochastic sequence, and consider the computable stochastic sequence defined by

$$Y_n = \alpha Y_{n-1} + X_n, \qquad n = 1, 2, \ldots, \tag{6.2}$$

where $Y_0 = 0$, and α, with $|\alpha| < 1$, is a rational number.

Let us remark that, in the following theorem, we will use a technique, which relies on the properties of Lèvy's metric, adapted from the proof of Theorem 5.2 of Elliott (1979), in which it is studied the limiting distribution of the values of additive arithmetic functions.

Theorem 6.1.2 *For the first-order autoregressive stochastic sequence (6.2), for any M-typical sequence ξ, as $N \to \infty$,*

$$\frac{1}{N}\#\{n : n \le N,\ Y(\xi^n) \le z\} \to \varphi(z),$$

uniformly in $z \in \mathbf{R}$, *where*

$$\varphi(z) = \lim_{r\to\infty} \varphi_r(z), \qquad \varphi_r(z) = \Pr(Z_1 + Z_2 + \cdots + Z_r \le z),$$

and $Z_1, Z_2, \ldots, Z_r$ *are independent discrete random variables with* $\Pr(Z_j = \alpha^{j-1}) = \Pr(Z_j = -\alpha^{j-1}) = 1/2$, $j = 1, 2, \ldots, r$.

Proof. Let ξ be an M-typical sequence, and consider, for large N, the computable stochastic sequences

$$Z_n^{(1)} = X_n, \quad Z_n^{(2)} = \alpha X_{n-1}, \ldots, \quad Z_n^{(j)} = \alpha^{j-1} X_{n-j+1}, \ldots, \quad Z_n^{(N)} = \alpha^{N-1} X_{n-N+1},$$

where $X_0 = 0$, $X_{-1} = 0, \ldots$. For any arbitrary r, $1 \le r \le N$, we split Y_n as follows,

$$Y_n = \sum_{j=1}^{N} Z_n^{(j)} = \sum_{j=1}^{r} Z_n^{(j)} + \sum_{j=r+1}^{N} Z_n^{(j)} = Y_n'(r) + Y_n''(r),$$

say, and to show the theorem, we just have to show that (see the remark following Lemma 1.7 of Elliott (1979))

$$(i) \qquad \frac{1}{N}\#\{n : n \le N,\ Y'(r)(\xi^n) \le z\} \to \varphi(z),$$

$$(ii) \qquad \frac{1}{N}\#\{n : n \le N,\ |Y''(r)(\xi^n)| > \epsilon\} \to 0, \qquad \text{for all } \epsilon > 0,$$

for an $r = r(N) \to \infty$, as $N \to \infty$.

Let us show (i). For every fixed r, $Y_n'(r)$ is a moving average stochastic sequence of order $q = r - 1$, where $\beta_j = \alpha^j$, $j = 0, 1, \ldots, r-1$, and we have already seen that

$$\frac{1}{N}\#\{n : n \le N,\ Y'(r)(\xi^n) \le z\} \to \varphi_r(z),$$

uniformly in $z \in \mathbf{R}$, as $N \to \infty$. Then we can certainly find a sequence $\{\epsilon_r\}$, $\epsilon_r > 0$, $\epsilon_r \to 0$, as $r \to \infty$, and integers $\{N_r\}$, such that, for every $z \in \mathbf{R}$,

$$\left|\frac{1}{N}\#\{n : n \le N,\ Y'(r)(\xi^n) \le z\} - \varphi_r(z)\right| < \epsilon_r,$$

for every $N \ge N_r$. Thus, there exists a function $r = r(N)$ which tends to infinity, as $N \to \infty$, slowly enough to ensure that

$$\frac{1}{N}\#\{n : n \le N,\ Y'(r)(\xi^n) \le z\} \to \varphi(z),$$

uniformly in $z \in \mathbf{R}$, as $N \to \infty$.

Let us now show (ii). Note that $|Y_n''(r)| = 0$, $n = 1, 2, \ldots, r$, and that

$$Y_n''(r) = \sum_{j=r+1}^{N} Z_n^{(j)} = \sum_{i=1}^{n-r} \alpha^{n-i} X_i,$$

$n = r+1, r+2, \ldots, N$, that is, that

$$|Y_n''(r)| = \left| \sum_{i=1}^{n-r} \alpha^{n-i} X_i \right| \leq \sum_{i=1}^{n-r} \alpha^{n-i} < \alpha^r \frac{1}{1-\alpha}.$$

So, for every $\epsilon > 0$, there is an r^* such that for every $r \geq r^*$, $|Y_n''(r)| < \epsilon$, and $1/N\#\{n : n \leq N,\ |Y''(r)(\xi^n)| > \epsilon\} = 0$. That is, (ii) is true for any $r = r(N) \to \infty$, as $N \to \infty$. **Q.E.D.**

6.2 Schatte's Strong Central Limit Theorem

Assuming Kolmogorov's axiomatics, let $X_1, X_2, \ldots$ be a sequence of independent identically distributed (i.i.d.) random variables, with $\mathrm{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$, on a probability space $(\Omega^\infty, \mathcal{F}, P)$, and let $S_n = X_1 + X_2 + \cdots + X_n$. Then, by the classical central limit theorem, we know that the distribution of the standardized sum $S_n/\sqrt{n}$ converges to the standard normal distribution $\Phi(z)$, as $n \to \infty$, but, of course, we do not know anything about the distribution of values of a single realization of the sequence of random variables $S_n/\sqrt{n}$. Indeed, Schatte (1988) showed that, owing to the strong serial dependence in the sequence $(S_n/\sqrt{n})$ (see for a depiction Figure 6.1), the probability is zero that the arithmetic means of the sequence $I_{(-\infty,z]}(S_n/\sqrt{n})$ converge, as $n \to \infty$.

On the other hand, by considering 'logarithmic means' instead of arithmetic means, Schatte proved the following result.

Theorem 6.2.1 (Schatte, 1988, Theorem 2) *Let $X_1, X_2, \ldots$ be i.i.d. random variables such that $\mathrm{E}(X_i)=0$, $\mathrm{Var}(X_i)=1$, $\mathrm{E}(|X_i|^3) < \infty$, and $S_n = X_1 + X_2 + \cdots + X_n$. Then*

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n} I_{(-\infty,z]}\left(\frac{S_n}{\sqrt{n}}\right) = \Phi(z), \tag{6.3}$$

almost surely, for all $z \in (-\infty, \infty)$, *where* $\Phi(z)$ *is the standard normal distribution function.*

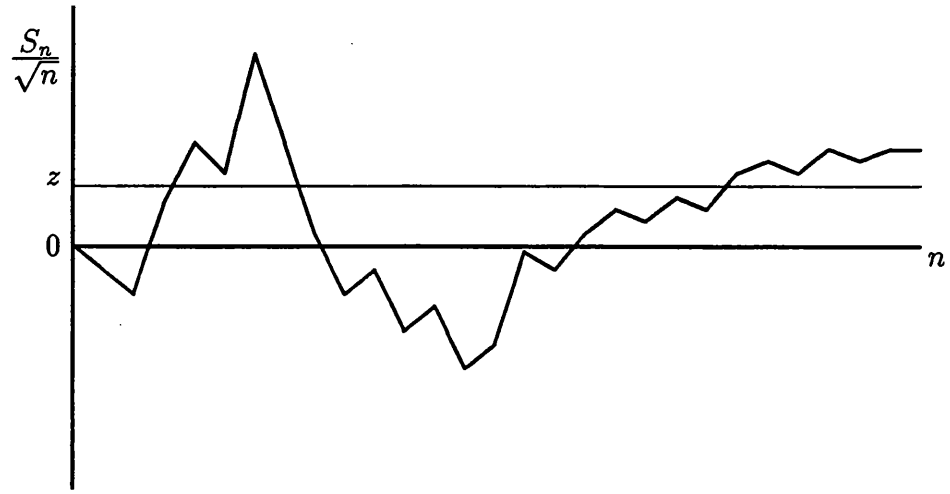


Figure 6.1: Typical realization of the standardized statistic $S_n/\sqrt{n}$.

Results of this kind, known as strong central limit theorems, were first shown independently by Schatte (1988) and by Brosamler (1988), this last for i.i.d. random variables having finite $(2+\delta)$th moments, and later by Lacey and Philipp (1990), assuming only finite variances. Schatte (1988) showed that the dependence among the $(S_n/\sqrt{n})$ can also be overcome by considering a suitable thinning of the original sequence of values, giving the following strong central limit theorem for subsequences.

Theorem 6.2.2 (Schatte, 1988, Theorem 3) *Let* $X_1, X_2, \ldots$ *be i.i.d. random variables such that* $\mathrm{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$, $\mathrm{E}(|X_i|^3) < \infty$, *and* S_n *be as before. Then for* $n_k = \lfloor c^k \rfloor$, $c > 1$, *where* $\lfloor c^k \rfloor$ *is the integer part of* c^k, *we have that*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} I_{(-\infty,z]}\left(\frac{S_{n_k}}{\sqrt{n_k}}\right) = \Phi(z),$$

almost surely, for all $z \in (-\infty, \infty)$.

The strong central limit theorem has been extended, still in the classical Kolmogorov probability axiomatics and a non-algorithmic framework, to a variety of

other stochastic processes. However, these results are usually proved by extending the strong central limit theorem for a sequence of i.i.d. random variables to the desired stochastic process, using some invariance principle. In particular, a direct martingale-based proof of this result does not yet seem available.

In the case of the strong central limit theorem for subsequences, a more direct argument is provided, for a martingale process, by the following reasoning. Let $S_0, S_1, \ldots$ be a martingale process, with respect to the filtration $(\mathcal{F}_n)_{n\geq 0}$ on the probability space $(\Omega^\infty, \mathcal{F}, P)$, such that the differences $X_n = S_n - S_{n-1}$ satisfy $\mathrm{E}(X_n|\mathcal{F}_{n-1}) = 0$ and $\mathrm{Var}(X_n|\mathcal{F}_{n-1}) = 1$. Fix a constant $\alpha \in (0,1)$, and consider the equality

$$\frac{S_{n_k}}{\sqrt{n_k}} = \frac{\sqrt{n_{k-1}}}{\sqrt{n_k}} \frac{S_{n_{k-1}}}{\sqrt{n_{k-1}}} + \frac{X_{n_{k-1}+1} + \cdots + X_{n_k}}{\sqrt{n_k}},$$

where $\{n_k\}$ is such that $n_1 = 1$, and, for $k = 2, 3, \ldots$, n_k is the first integer for which

$$\frac{\sqrt{n_{k-1}}}{\sqrt{n_k}} \leq \alpha.$$

Defining $T_{n_k} = S_{n_k}/\sqrt{n_k}$ and $U_{n_k} = (X_{n_{k-1}+1} + \cdots + X_{n_k})/\sqrt{n_k}$, this equality can approximately be rewritten as

$$T_{n_k} = \alpha T_{n_{k-1}} + U_{n_k},$$

where $\mathrm{E}(U_{n_k}|\mathcal{F}_{n_{k-1}}) = 0$ and $\mathrm{Var}(U_{n_k}|\mathcal{F}_{n_{k-1}}) = (n_k - n_{k-1})/n_k \approx 1 - \alpha^2$. Thus, under some mild conditions assuring that

$$T_n = \frac{S_n}{\sqrt{n}} \xrightarrow{L} \mathcal{N}(0,1),$$

as $n \to \infty$, T_{n_k} behaves approximately as a Gaussian autoregressive process of first-order with a standard normal stationary distribution, and so, by the classical ergodic theorem, the empirical distribution function

$$\frac{1}{K} \sum_{k=1}^{K} I_{(-\infty,z]}\left(\frac{S_{n_k}}{\sqrt{n_k}}\right) \longrightarrow \Phi(z),$$

almost surely, as $k \to \infty$.

This argument, even if it does not provide any formal proof, apart from giving a clear intuition of the result, also points out the fundamental role which seems to be

played by the ergodic theorem, by means of which it is possible to attain an almost sure convergence from a convergence in distribution.

Now, unlike many other almost sure results, neither the ergodic theorem nor the strong central limit theorem, which are both about an asymptotic regularity which is valid with probability one, that is, which is valid for almost every single infinite realization, have yet been proved in an algorithmic framework for some random (typical or chaotic) sequences (Vovk, 1995b). And, indeed, it might be that the ergodic theorem does not hold for some random sequences. On the other hand, as far as M-typical sequences are concerned, we will tackle the problem of proving the strong central limit theorem in our purely martingale algorithmic framework in the next two sections.

6.3 Towards the Strong Central Limit Theorem

In this and the next section we will consider the problem of proving the strong central limit theorem for M-typical sequences in the case of a computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, with $X_i \in \{-1, 1\}$. Here, we will start by considering some properties about the indicator function of the standardized martingale $M_n/\sqrt{n}$. We will end this section by showing two results. One about the disjoint sums of the basic martingale M, and the other about a simplified version of the strong central limit theorem for subsequences.

Let us first consider the indicator function $I_{(-\infty,z]}(M_n/\sqrt{n})$ in the case $z = 0$, for which it reduces to $I_{(-\infty,0]}(M_n/\sqrt{n}) = I_{(-\infty,0]}(M_n)$. For $n = 1$, we simply have

$$I_{(-\infty,0]}(X_1) = \frac{1}{2} - \frac{X_1}{2}.$$

For $n = 2$, by setting $I_i = I_{(-\infty,0]}(X_i)$, and by considering the following table, in

which all possible combinations of X_1, X_2 are considered,

X_1	X_2	$I_{(-\infty,0]}(X_1+X_2)$	I_1	I_2	$I_{(-\infty,0]}(X_1+X_2)$
1	1	0	0	0	—
1	−1	1	0	1	$(1-I_1)I_2$
−1	1	1	1	0	$I_1(1-I_2)$
−1	−1	1	1	1	I_1I_2

it is easy to check that

$$I_{(-\infty,0]}(X_1+X_2) = (1-I_1)I_2 + I_1(1-I_2) + I_1I_2 = \frac{3}{4} - \frac{X_1}{4} - \frac{X_2}{4} - \frac{X_1X_2}{4}.$$

In the same way, we would also have, for $n = 3$,

$$\begin{aligned} I_{(-\infty,0]}(X_1+X_2+X_3) &= I_1I_2I_3 + I_1I_2(1-I_3) + I_1(1-I_2)I_3 + (1-I_1)I_2I_3 \\ &= \frac{1}{2} - \frac{X_1}{4} - \frac{X_2}{4} - \frac{X_3}{4} + \frac{X_1X_2X_3}{4}, \end{aligned}$$

for $n = 4$,

$$\begin{aligned} &I_{(-\infty,0]}(X_1+X_2+X_3+X_4) \\ &= I_1I_2I_3I_4 + I_1I_2I_3(1-I_4) + I_1I_2(1-I_3)I_4 + I_1I_2(1-I_3)(1-I_4) + I_1(1-I_2)I_3I_4 \\ &\quad + I_1(1-I_2)I_3(1-I_4) + I_1(1-I_2)(1-I_3)I_4 + (1-I_1)I_2I_3I_4 + (1-I_1)I_2I_3(1-I_4) \\ &\quad + (1-I_1)I_2(1-I_3)I_4 + (1-I_1)(1-I_2)I_3I_4 \\ &= \frac{11}{16} - \frac{3}{16}(X_1+X_2+X_3+X_4) - \frac{1}{16}(X_1X_2+X_1X_3+X_1X_4+X_2X_3+X_2X_4+X_3X_4) \\ &\quad + \frac{1}{16}(X_1X_2X_3+X_1X_2X_4+X_1X_3X_4+X_2X_3X_4) + \frac{3}{16}X_1X_2X_3X_4, \end{aligned}$$

and so forth, for all $n \in \mathbf{N}$. Here, just note that the sum of the coefficients of the expression in the X_i is zero, for every $n \in \mathbf{N}$, since it is the value of the expression when all the X_i are equal to one.

It is easy to see that all these indicator functions can be written as

$$\begin{aligned} I_{(-\infty,0]}(X_1) &= \frac{1}{2} + V_{11}X_1, \\ I_{(-\infty,0]}(X_1+X_2) &= \frac{3}{4} + V_{21}X_1 + V_{22}X_2, \\ I_{(-\infty,0]}(X_1+X_2+X_3) &= \frac{4}{8} + V_{31}X_1 + V_{32}X_2 + V_{33}X_3, \end{aligned}$$

$$I_{(-\infty,0]}(X_1+X_2+X_3+X_4) = \frac{11}{16} + V_{41}X_1 + V_{42}X_2 + V_{43}X_3 + V_{44}X_4,$$

$$\vdots$$

where the V_{ni} depend only on $X_1, X_2, \ldots, X_{i-1}$, and that, noting that the predictable sequences defined in the proof of Lemma 6.1.1 satisfy

$$V_n^- = I_{n-1}, \quad V_n^+ = 1 - I_{n-1},$$
$$V_n^{--} = V_n^- V_{n-1}^- = I_{n-1}I_{n-2}, \quad V_n^{+-} = V_n^+ V_{n-1}^- = (1 - I_{n-1})I_{n-2}, \quad \text{etc.},$$

we can also write

$$I_{(-\infty,0]}(M_1) = \frac{1}{2} - \frac{1}{2}X_1,$$
$$I_{(-\infty,0]}(M_2) = \frac{3}{4} - \frac{1}{4}X_1 - \frac{1}{2}V_2^+ X_2,$$
$$I_{(-\infty,0]}(M_3) = \frac{4}{8} - \frac{1}{8}2X_1 - \frac{1}{4}X_2 - \frac{1}{2}(V_3^{+-} + V_3^{-+})X_3,$$
$$I_{(-\infty,0]}(M_4) = \frac{11}{16} - \frac{1}{16}3X_1 - \frac{1}{8}(V_2^- + 2V_2^+)X_2 - \frac{1}{4}(V_3^{+-} + V_3^{-+} + V_3^{++})X_3$$
$$- \frac{1}{2}(V_4^{++-} + V_4^{+-+} + V_4^{-++})X_4.$$

$$\vdots$$

In this way, for every $n \in \mathbf{N}$, the indicator function of M_n, for $z = 0$, can be written as a polynomial expansion in the X_i of the form

$$I_{(-\infty,0]}(M_n) = \varphi_n + V_{n1}X_1 + \cdots + V_{nn}X_n,$$

where φ_n is a rational number in $(1/2,\ 1]$, such that $\varphi_n \to 1/2$, as $n \to \infty$, and the V_{ni} are computable quantities which depend only on $X_1, X_2, \ldots, X_{i-1}$.

Consider now the case of a general $z \in \mathbf{R}_c$. Similarly to before, for every $n \in \mathbf{N}$, the indicator function $I_{(-\infty,z]}(M_n/\sqrt{n})$ can be written as the sum of m products of n terms, where m is the number of paths of length n such that $M_n \leq \sqrt{n}z$, and the n terms are of the form $(1/2 \pm X_i/2)$. That is, we can write

$$I_{(-\infty,z]}\left(\frac{M_n}{\sqrt{n}}\right) = \varphi_n(z) + V_{n1}(z)X_1 + \cdots + V_{nn}(z)X_n,$$

where $\varphi_n(z)$ is a rational number which does not depend on the X_i, and the $V_{ni}(z)$ are computable quantities which depend on the X_i only through $X_1, X_2, \ldots, X_{i-1}$.

By the structure of this polynomial expansion, it is easy to see that $\varphi_n(z)$ is just m/n, that is, the proportion of paths of length n for which $M_n \leq \sqrt{n}z$. And, by introducing on Ω^∞ the auxiliary uniform Bernoulli measure μ, since

$$\varphi_n(z) = \mu\Big(\frac{M_n}{\sqrt{n}} \leq z\Big),$$

we also have that $\varphi_n(z) \to \Phi(z)$, as $n \to \infty$, by the De Moivre–Laplace central limit theorem (Feller, 1968, Theorem VII.3.2). Let us note that here and later the measure μ plays only an auxiliary role and in our purely martingale framework it does not have any probabilistic interpretation. Indeed, it just provides a convenient way to express an underlying combinatorial argument.

With the help of the auxiliary measure μ, the quantities $V_{ni}(z)$ can be evaluated by considering the proportion of paths of length n for which $M_n \leq \sqrt{n}z$, given the initial realizations $(X_1, X_2, \ldots, X_i)$. In fact, we have

$$\begin{aligned}
\mu\Big(\frac{M_n}{\sqrt{n}} \leq z \Big| X_1, \ldots, X_i\Big) &= \mu\Big(I_{(-\infty,z]}\Big(\frac{M_n}{\sqrt{n}}\Big) = 1 \Big| X_1, \ldots, X_i\Big) \\
&= \mathrm{E}\Big(I_{(-\infty,z]}\Big(\frac{M_n}{\sqrt{n}}\Big) \Big| X_1, \ldots, X_i\Big) \\
&= \varphi_n(z) + V_{n1}(z)X_1 + \cdots + V_{ni}(z)X_i,
\end{aligned}$$

where the expectation $\mathrm{E}(\cdot)$ is taken with respect to μ, since, under μ, the X_i are independent random variables with $\mathrm{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 1$. And so, by considering the two conditioning sets $(X_1, \ldots, X_{i-1})$ and $(X_1, \ldots, X_{i-1}, X_i = 1)$, we immediately have

$$V_{ni}(z) = -\mu\Big(\frac{M_n}{\sqrt{n}} \leq z \Big| X_1, \ldots, X_{i-1}\Big) + \mu\Big(\frac{M_n}{\sqrt{n}} \leq z \Big| X_1, \ldots, X_{i-1}, X_i = 1\Big).$$

By rewriting this last expression as

$$V_{ni}(z) = -\mu(M_{n-i+1} \leq z^*) + \mu(M_{n-i} \leq z^* - 1),$$

where $z^* = z\sqrt{n} - (X_1 + \cdots + X_{i-1})$, it is possible to see that we can bound $V_{ni}(z)$ by

$$-\frac{1}{2}\mu(M_{n-i} = 0), \quad \text{or} \quad -\frac{1}{2}\mu(M_{n-i} = 1),$$

depending on whether $(n-i)$ is even or odd, which, by using Stirling's formula (Feller, 1968, Section II.9), can both be approximated by

$$-\frac{1}{\sqrt{2\pi(n-i)}}, \tag{6.4}$$

as $n \to \infty$.

6.3.1 Disjoint Sums

Let us consider now the following result which can be seen as a first step towards the strong central limit theorem for subsequences.

Theorem 6.3.1 *Consider the computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, with $X_i \in \{-1, 1\}$, and an arbitrary computable subsequence $\{n_k\}$, such that $l_j = n_j - n_{j-1} \to \infty$, as $j \to \infty$. Then, for every M-typical sequence ξ,*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K} I_{(-\infty,z]}\left(\frac{Y_j(\xi^{n_j})}{\sqrt{n_j - n_{j-1}}}\right) = \Phi(z),$$

for all $z \in \mathbf{R}_c$, where $\mathbf{R}_c$ is the set of all computable real numbers, $\Phi(z)$ is the standard normal distribution function, and $Y_j = X_{n_{j-1}+1} + \cdots + X_{n_j}$.

Proof. Let us fix a $z \in \mathbf{R}_c$. By using the previous properties about the indicator function, we can write

$$\frac{1}{K}\sum_{j=1}^{K} I_{(-\infty,z]}\left(\frac{Y_j}{\sqrt{n_j - n_{j-1}}}\right) = \frac{1}{K}\sum_{j=1}^{K} \varphi_{l_j} + \frac{1}{K}\sum_{j=1}^{K}(V_{n_{j-1}+1}X_{n_{j-1}+1} + \cdots + V_{n_j}X_{n_j}),$$

where V is a computable predictable sequence, every V_n, $n_{j-1}+1 \le n \le n_j$, depends only on $X_{n_{j-1}+1} + \cdots + X_n$, and the φ_{l_j} are rational numbers in [0, 1], such that $\varphi_{l_j} \to \Phi(z)$, as $l_j \to \infty$. Since, by the properties of the arithmetic mean, the first average on the right-hand side tends to $\Phi(z)$, as $K \to \infty$, we have just to show that the second average, on the right-hand side, tends to zero, as $K \to \infty$.

Let ξ be an M-typical sequence, and note that, since $S_n = \sum_{i=1}^{n} V_i X_i$ is a computable M-martingale, the sequence ξ is also S-typical. Consider the computable S_{n_k}-martingale

$$R_K = \sum_{j=1}^{K} \frac{1}{j}(S_{n_j} - S_{n_{j-1}}) = \sum_{j=1}^{K} \frac{1}{j}(V_{n_{j-1}+1}X_{n_{j-1}+1} + \cdots + V_{n_j}X_{n_j}),$$

and also the computable stochastic sequence

$$R_K^2 = \Big[\sum_{j=1}^{K} \frac{(S_{n_j} - S_{n_{j-1}})^2 - d_j)}{j^2} + 2\sum_{j>h}^{K} \frac{S_{n_j} - S_{n_{j-1}}}{j} \frac{S_{n_h} - S_{n_{h-1}}}{h}\Big] + \sum_{j=1}^{K} \frac{d_j}{j^2},$$

where d is a computable non-negative predictable sequence defined by

$$d_j = \varphi_{l_j} - \varphi_{l_j}^2.$$

By noting that $I_{(-\infty,z]}(\cdot) = (I_{(-\infty,z]}(\cdot))^2$, we can see that $1 - 2\varphi_{l_j}$ is a computable predictable sequence, and that

$$\begin{aligned}
\sum_{j=1}^{K}[(S_{n_j} - S_{n_{j-1}})^2 - d_j] &= \sum_{j=1}^{K}\Big[(I_{(-\infty,z]}(Y_j/\sqrt{n_j - n_{j-1}}) - \varphi_{l_j})^2 - d_j\Big] \\
&= \sum_{j=1}^{K}[1 - 2\varphi_{l_j}]\Big[I_{(-\infty,z]}(Y_j/\sqrt{n_j - n_{j-1}}) - \varphi_{l_j}\Big] \\
&= \sum_{j=1}^{K}[1 - 2\varphi_{l_j}](S_{n_j} - S_{n_{j-1}}).
\end{aligned}$$

So, considering the right-hand side of R_K^2, since both sums in square brackets are computable S_{n_k}-martingales, the whole sum in square brackets is a computable S_{n_k}-martingale, and $\sum_{j=1}^{K} d_j/j^2$ is a computable compensator of the computable S_{n_k}-submartingale R_K^2. Then, since

$$\sum_{j=1}^{\infty} \frac{d_j}{j^2} = \sum_{j=1}^{\infty} \frac{\varphi_{l_j} - \varphi_{l_j}^2}{j^2} < \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty,$$

by Lemma 5.8.3, applied to R_K and R_K^2, we have that R_K converges, and, by Kronecker's lemma, that

$$\lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K}(V_{n_{j-1}+1}X_{n_{j-1}+1} + \cdots + V_{n_j}X_{n_j}) = 0.$$

Q.E.D.

6.3.2 Signs in Subsequences

We consider a simplified version of the strong central limit theorem for subsequences in which the standardized sum $M_n/\sqrt{n}$ is replaced with a statistic which permits use of the argument employed in Theorem 6.1.2.

For every fixed rational number $\alpha \in (0,1)$, let $\{n_k\}$ be the computable subsequence such that $n_1 = 1$, and, for $k = 2, 3, \ldots$, n_k is the first integer for which

$$\frac{\sqrt{n_{k-1}}}{\sqrt{n_k}} \leq \alpha. \tag{6.5}$$

Theorem 6.3.2 *Consider the computable basic martingale $M_n = \sum_{i=1}^n X_i$, with $X_i \in \{-1, 1\}$, and, for a fixed rational number $\alpha \in (0,1)$, the computable subsequence $\{n_k\}$ given by (6.5). Define*

$$W_j = \begin{cases} -1, & Y_j < 0, \\ 0, & Y_j = 0, \\ 1, & Y_j > 0, \end{cases}$$

where $Y_j = X_{n_{j-1}+1} + \cdots + X_{n_j}$, and consider the weighted sum statistic

$$L_k = \frac{1}{\sqrt{n_k}} \sum_{j=1}^{k} \sqrt{n_j} W_j.$$

Then, for any M-typical sequence ξ, as $K \to \infty$,

$$\frac{1}{K} \#\{k : k \leq K,\ L_k(\xi^{n_k}) \leq z\} \to \varphi(z),$$

uniformly in $z \in \mathbf{R}$, where $\varphi(z)$ is as defined in Theorem 6.1.2.

Proof. For $r \in \mathbf{N}$, $1 \leq r \leq K$, we have

$$L_k = \sum_{j=1}^{k} \frac{\sqrt{n_j}}{\sqrt{n_k}} W_j = \sum_{j=1}^{k-r} \frac{\sqrt{n_j}}{\sqrt{n_k}} W_j + \sum_{j=k-r+1}^{k} \frac{\sqrt{n_j}}{\sqrt{n_k}} W_j = L_k''(r) + L_k'(r),$$

say. Then it is enough to show that, for any M-typical sequence ξ,

$$(i) \quad \frac{1}{K} \#\{k : k \leq K,\ L_k'(r)(\xi^{n_k}) \leq z\} \to \varphi(z),$$

$$(ii) \quad \frac{1}{K} \#\{k : k \leq K,\ |L_k''(r)(\xi^{n_k})| > \epsilon\} \to 0, \qquad \text{for all } \epsilon > 0,$$

for an $r = r(K) \to \infty$, as $K \to \infty$.

Let us prove (i). For a fixed r, consider the truncated weighted sum statistic

$$\overline{L}_k'(r) = \sum_{j=k-r+1}^{k} \alpha^{k-j} W_j.$$

Following the argument used in Theorem 6.1.2, and using the result for disjoint partial sums of Theorem 6.3.1, we have that, for any M-typical sequence ξ,

$$\frac{1}{K}\#\{k : k \leq K,\ \overline{L}'_k(r)(\xi^{n_k}) \leq z\} \to \varphi_r(z),$$

uniformly in $z \in \mathbf{R}$, as $K \to \infty$, where $\varphi_r(z) = \Pr(Z_1 + Z_2 + \cdots + Z_r \leq z)$, and $Z_1, Z_2, \ldots, Z_r$ are independent discrete random variables with $\Pr(Z_j = \alpha^{j-1}) = \Pr(Z_j = -\alpha^{j-1}) = 1/2$, $j = 1, 2, \ldots, r$. Then, since $\sqrt{n_j}/\sqrt{n_k} \to \alpha_{k-j}$, $(j = k-r+1, k-r+2, \ldots, k)$, as $k \to \infty$, for every M-typical sequence ξ, we also have that $L'_k(r)(\xi^{n_k}) \to \overline{L}'_k(r)(\xi^{n_k})$, as $k \to \infty$, and that

$$\frac{1}{K}\#\{k : k \leq K,\ L'_k(r)(\xi^{n_k}) \leq z\} \to \frac{1}{K}\#\{k : k \leq K,\ \overline{L}'_k(r)(\xi^{n_k}) \leq z\} \to \varphi_r(z),$$

as $K \to \infty$. Thus, since this is valid for every fixed r, it is also valid for some $r = r(K) \to \infty$, as $K \to \infty$, and so (i) is proved.

To show (ii), note that, $|L''_k(r)| = 0$, $k = 1, 2, \ldots, r$, and

$$|L''_k(r)| \leq \sum_{j=1}^{k-r} \frac{\sqrt{n_j}}{\sqrt{n_k}} \leq \sum_{j=1}^{k-r} \alpha^{k-j} < \alpha^r \frac{1}{1-\alpha},$$

$k = r+1, r+2, \ldots, K$. That is, (ii) is true for any $r = r(K) \to \infty$, as $K \to \infty$. So, the result is proved. **Q.E.D.**

Note that, as compared with the statement of the central limit theorem for subsequences, we have replaced the standardized sum

$$\frac{1}{\sqrt{n_k}} \sum_{j=1}^{k} Y_j, \quad \text{with} \quad \frac{1}{\sqrt{n_k}} \sum_{j=1}^{k} \sqrt{n_j} W_j,$$

that is, we substituted Y_j with the easier binary quantity $\sqrt{n_j} W_j$. Also the distribution of values obtained, as in the case of the first-order autoregressive stochastic sequence, is not standard normal.

6.4 The Strong Central Limit Theorem

We will present here some strong central limit theorem type results for M-typical sequences in the case of a computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, with $X_i \in \{-1, 1\}$. In Section 6.4.1 we use logarithmic averages, while in Section 6.4.2 we give results for subsequences.

6.4.1 Logarithmic Averages

The following result is a version for M-typical sequences of the strong central limit theorem for logarithmic averages, in the case in which the basic martingale M represents a coin-tossing process. Note that, for finite N, the logarithmic average considered by Schatte's strong central limit theorem in (6.3) is not a distribution function, being greater than one for z large enough. However, it tends to a distribution function, as $N \to \infty$.

Theorem 6.4.1 *Consider the computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, with $X_i \in \{-1, 1\}$. Then, for any M-typical sequence ξ,*

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n} I_{(-\infty,z]}\left(\frac{M_n}{\sqrt{n}}\right) = \Phi(z),$$

uniformly in $z \in \mathbf{R}_c$, where $\mathbf{R}_c$ is the set of all computable real numbers, and $\Phi(z)$ is the standard normal distribution function.

Proof. Let us fix a $z \in \mathbf{R}_c$. Since, from Section 6.3, for every $n \in \mathbf{N}$,

$$I_{(-\infty,z]}\left(\frac{M_n}{\sqrt{n}}\right) = \varphi_n + V_{n1}X_1 + \cdots + V_{nn}X_n,$$

where φ_n is a rational number in $[0, 1]$, such that $\varphi_n \to \Phi(z)$, as $n \to \infty$, and the V_{ni} are computable quantities which depend on the X_i only through $X_1, X_2, \ldots, X_{i-1}$, we can write

$$\frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n} I_{(-\infty,z]}\left(\frac{M_n}{\sqrt{n}}\right) = \frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n}\varphi_n + \frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n}(V_{n1}X_1 + \cdots + V_{nn}X_n).$$

Since the first term on the right-hand side tends to $\Phi(z)$, as $N \to \infty$, it remains to show that the second term tends to zero, as $N \to \infty$. Let us note that the sum in this second average, that is, in

$$\frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n}(V_{n1}X_1 + \cdots + V_{nn}X_n), \tag{6.6}$$

is not an M-martingale, and we cannot apply any strong law of large numbers directly to it. Nevertheless, the average (6.6) can be written as

$$\frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n}(V_{n1}X_1 + \cdots + V_{nn}X_n) = \frac{1}{\ln N} R_N - \frac{1}{\ln N} \sum_{n=N+1}^{\infty} \frac{1}{n}(V_{n1}X_1 + \cdots + V_{nN}X_N),$$

where

$$R_N = \Big(\sum_{r=1}^{\infty} \frac{1}{r} V_{r1}\Big) X_1 + \Big(\sum_{r=2}^{\infty} \frac{1}{r} V_{r2}\Big) X_2 + \cdots + \Big(\sum_{r=N}^{\infty} \frac{1}{r} V_{rN}\Big) X_N = \sum_{i=1}^{N} W_i X_i,$$

say, is a computable M-martingale, and W is a computable predictable sequence, negative, and, by using the asymptotic approximation (6.4), with

$$|W_i| = \Big|\sum_{r=i}^{\infty} \frac{1}{r} V_{ri}\Big| \sim \frac{1}{\sqrt{2\pi}} \sum_{r=1}^{\infty} \frac{1}{i+r} \frac{1}{\sqrt{r}} \leq \frac{c}{\sqrt{i}},$$

where c is a constant.

Then since, for $i_j = \lfloor e^j \rfloor$,

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{i=1}^{N} W_i X_i = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} (W_{i_{j-1}+1} X_{i_{j-1}+1} + \cdots + W_{i_j} X_{i_j}),$$

and since

$$\sum_{j=1}^{\infty} \frac{1}{j^2} (W_{i_{j-1}+1}^2 + \cdots + W_{i_j}^2) < \sum_{j=1}^{\infty} \frac{1}{j^2} \Big(\frac{c^2}{i_{j-1}+1} + \cdots + \frac{c^2}{i_j} \Big) < \sum_{j=1}^{\infty} \frac{c^2}{j^2} \frac{i_j - i_{j-1}}{i_{j-1}+1} < \infty,$$

by the strong law of large numbers of Theorem 5.3.3 we have that, for any M-typical sequence,

$$\frac{1}{\ln N} \sum_{i=1}^{N} W_i X_i \to 0,$$

as $N \to \infty$. Thus, since by the law of the iterated logarithm of Theorem 5.6.2, for any M-typical sequence,

$$\lim_{N\to\infty} \frac{1}{\ln N} |M_N| \Big| \sum_{n=N+1}^{\infty} \frac{1}{n} V_{nN} \Big| \leq \lim_{N\to\infty} \frac{1}{\ln N} \sqrt{2N \ln\ln N} \frac{c}{\sqrt{N}} = 0,$$

and since the V_{ni} are all non-positive, and for $i < N$,

$$\Big| \sum_{n=N+1}^{\infty} \frac{1}{n} V_{ni} \Big| < \Big| \sum_{n=N+1}^{\infty} \frac{1}{n} V_{nN} \Big|,$$

where the term on the right-hand side tends to zero, as $N \to \infty$, we have that

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=N+1}^{\infty} \frac{1}{n} (V_{n1} X_1 + \cdots + V_{nN} X_N) = \lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=N+1}^{\infty} \frac{1}{n} (V_{nN} X_1 + \cdots + V_{nN} X_N) = 0,$$

and the theorem is proved. **Q.E.D.**

Let us consider now the following result which is about the first moment, in logarithmic average, of the standardized martingale $M_n/\sqrt{n}$.

Lemma 6.4.1 *Consider the computable basic martingale $M_n = \sum_{i=1}^n X_i$, with $X_i \in \{-1,1\}$. Then, for any M-typical sequence ξ,*

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n} \frac{M_n}{\sqrt{n}} = 0.$$

Proof. Write

$$\frac{1}{\ln N} \sum_{n=1}^{N} \frac{1}{n} \frac{M_n}{\sqrt{n}} = \frac{1}{\ln N} \sum_{i=1}^{N} W_i X_i - \frac{1}{\ln N} \sum_{n=N+1}^{\infty} \frac{1}{n} \frac{M_n}{\sqrt{n}},$$

where

$$W_i = \sum_{r=i}^{\infty} \frac{1}{r\sqrt{r}} = o\Big(\frac{1}{\sqrt{i}}\Big),$$

is a computable predictable sequence.

Then, by considering that $\sum_{i=1}^N W_i X_i$ is a computable M-martingale, since, for $i_j = \lfloor e^j \rfloor$,

$$\lim_{N\to\infty} \frac{1}{\ln N} \sum_{i=1}^{N} W_i X_i = \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} (W_{i_{j-1}+1} X_{i_{j-1}+1} + \cdots + W_{i_j} X_{i_j}),$$

and since

$$\sum_{j=1}^{\infty} \frac{1}{j^2} (W^2_{i_{j-1}+1} + \cdots + W^2_{i_j}) < \sum_{j=1}^{\infty} \frac{1}{j^2} \Big(\frac{c^2}{i_{j-1}+1} + \cdots + \frac{c^2}{i_j} \Big) < \sum_{j=1}^{\infty} \frac{c^2}{j^2} \frac{i_j - i_{j-1}}{i_{j-1}+1} < \infty,$$

by the strong law of large numbers of Theorem 5.3.3 we have that, for any M-typical sequence,

$$\frac{1}{\ln N} \sum_{i=1}^{N} W_i X_i \to 0,$$

as $N \to \infty$.

Moreover, by using the law of the iterated logarithm of Theorem 5.6.2, we have that, for any M-typical sequence,

$$\lim_{N\to\infty} \Big| \frac{1}{\ln N} \sum_{n=N+1}^{\infty} \frac{1}{n} \frac{M_N}{\sqrt{n}} \Big| \leq \lim_{N\to\infty} \frac{1}{\ln N} \sqrt{2N \ln\ln N} \frac{c}{\sqrt{N+1}} = 0.$$

Q.E.D.

Note that, in the same way, we could consider all the moments, in logarithmic average, of the standardized martingale $M_n/\sqrt{n}$.

6.4.2 Subsequences

The following theorem is a version for M-typical sequences, always in the case of a basic martingale M representing a coin-tossing process, of Schatte's strong central limit theorem for subsequences. It is worth noting that its proof follows closely the proof of Theorem 6.4.1.

Theorem 6.4.2 *Consider the computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, with $X_i \in \{-1, 1\}$, and, for a fixed rational number $\alpha \in (0,1)$, the computable subsequence $\{n_k\}$ given by (6.5). Then, for any M-typical sequence ξ,*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K} I_{(-\infty,z]}\Big(\frac{M_{n_j}}{\sqrt{n_j}}\Big) = \Phi(z),$$

uniformly in $z \in \mathbf{R}_c$.

Proof. Let us fix a $z \in \mathbf{R}_c$. From Section 6.3, for every n_j, $j = 1, 2, \ldots$,

$$I_{(-\infty,z]}\Big(\frac{M_{n_j}}{\sqrt{n_j}}\Big) = \varphi_{n_j} + V_{n_j,1}X_1 + \cdots + V_{n_j,n_j}X_{n_j},$$

where φ_{n_j} is a rational number in $[0,1]$, such that $\varphi_{n_j} \to \Phi(z)$, as $n_j \to \infty$, and the $V_{n_j,i}$ are computable quantities which depend on the X_i only through $X_1, X_2, \ldots, X_{i-1}$. Then we have

$$\frac{1}{K}\sum_{j=1}^{K} I_{(-\infty,z]}\Big(\frac{M_{n_j}}{\sqrt{n_j}}\Big) = \frac{1}{K}\sum_{j=1}^{K} \varphi_{n_j} + \frac{1}{K}\sum_{j=1}^{K}(V_{n_j,1}X_1 + \cdots + V_{n_j,n_j}X_{n_j}),$$

and since the first term on the right-hand side tends to $\Phi(z)$, as $K \to \infty$, we have to show that the second term tends to zero, as $K \to \infty$. To show this, we can write

$$\frac{1}{K}\sum_{j=1}^{K}(V_{n_j,1}X_1 + \cdots + V_{n_j,n_j}X_{n_j}) = \frac{1}{K}R_K - \frac{1}{K}\sum_{j=K+1}^{\infty}(V_{n_j,1}X_1 + \cdots + V_{n_j,n_K}X_{n_K}),$$

where

$$R_K = \sum_{j=1}^{K}(W_{n_{j-1}+1}X_{n_{j-1}+1} + \cdots + W_{n_j}X_{n_j}),$$

and

$$W_i = \sum_{r=j}^{\infty} V_{n_r,i}, \qquad n_{j-1} < i \le n_j,$$

is a negative computable predictable sequence, such that, by using the asymptotic approximation (6.4),

$$|W_{n_j}| = \Big|\sum_{r=j}^{\infty} V_{n_r,n_j}\Big| \sim \frac{1}{\sqrt{2\pi}} \sum_{r=j+1}^{\infty} \frac{1}{\sqrt{n_r - n_j}} \leq \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n_j}} \sum_{r=1}^{\infty} \frac{\alpha^r}{\sqrt{1-\alpha^{2r}}} \leq c\, \alpha^{j-1},$$

where c is a constant.

Then, since $|W_i| \leq |W_{n_j}|$, when $i \leq n_j$,

$$\sum_{j=1}^{\infty} \frac{1}{j^2}(W^2_{n_{j-1}+1} + \cdots + W^2_{n_j}) \leq \sum_{j=1}^{\infty} \frac{1}{j^2} W^2_{n_j}(n_j - n_{j-1}) \leq c^2(1-\alpha^2) \sum_{j=1}^{\infty} \frac{1}{j^2} < \infty,$$

and by the strong law of large numbers of Theorem 5.3.3 we have that, for any M-typical sequence,

$$\lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K} (W_{n_{j-1}+1} X_{n_{j-1}+1} + \cdots + W_{n_j} X_{n_j}) = 0.$$

Also, by the law of the iterated logarithm of Theorem 5.6.2, for any M-typical sequence,

$$\lim_{K\to\infty} \frac{1}{K} |M_{n_K}| \Big| \sum_{j=K+1}^{\infty} V_{n_j,n_K} \Big| \leq \lim_{K\to\infty} \frac{1}{K} \sqrt{2 n_K \ln\ln n_K} \frac{c}{\sqrt{n_K}} = 0,$$

and since the $V_{n_j,i}$ are all non-positive, and for $i < n_K$,

$$\Big| \sum_{j=K+1}^{\infty} V_{n_j,i} \Big| < \Big| \sum_{j=K+1}^{\infty} V_{n_j,n_K} \Big|,$$

where the term on the right-hand side tends to zero, as $K \to \infty$, we have that

$$\lim_{K\to\infty} \frac{1}{K} \sum_{j=K+1}^{\infty} (V_{n_j,1} X_1 + \cdots + V_{n_j,n_K} X_{n_K}) = \lim_{K\to\infty} \frac{1}{K} \sum_{j=K+1}^{\infty} (V_{n_j,n_K} X_1 + \cdots + V_{n_j,n_K} X_{n_K}) = 0,$$

and the theorem is proved. **Q.E.D.**

Let us consider now the analogue for subsequences of Lemma 6.4.1.

Lemma 6.4.2 *Consider the computable basic martingale $M_n = \sum_{i=1}^{n} X_i$, with $X_i \in \{-1, 1\}$, and, for a fixed rational number $\alpha \in (0,1)$, the computable subsequence $\{n_k\}$ given by (6.5). Then, for any M-typical sequence ξ,*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{K} \frac{M_{n_j}}{\sqrt{n_j}} = 0.$$

Proof. We have

$$\frac{1}{K}\sum_{j=1}^{K}\frac{M_{n_j}}{\sqrt{n_j}} = \frac{1}{K}\sum_{j=1}^{K} W_j(X_{n_{j-1}+1}+\cdots+X_{n_j}) - \frac{1}{K}\sum_{j=K+1}^{\infty}\frac{M_{n_K}}{\sqrt{n_j}},$$

where

$$W_j = \sum_{r=j}^{\infty}\frac{1}{\sqrt{n_r}} \leq \sum_{r=j}^{\infty}\alpha^{r-1} = \frac{\alpha^{j-1}}{1-\alpha},$$

since $\sqrt{n_{j-1}}/\sqrt{n_j} \leq \alpha$. Then

$$\sum_{j=1}^{\infty}\frac{1}{j^2}W_j^2(n_j - n_{j-1}) \simeq \frac{1-\alpha^2}{(1-\alpha)^2}\sum_{j=1}^{\infty}\frac{1}{j^2} < \infty,$$

and by the strong law of large numbers of Theorem 5.3.3 we have that, for any M-typical sequence,

$$\lim_{K\to\infty}\frac{1}{K}\sum_{j=1}^{K} W_j(X_{n_{j-1}+1}+\cdots+X_{n_j}) = 0.$$

Also, by the law of the iterated logarithm of Theorem 5.6.2, for any M-typical sequence,

$$\lim_{K\to\infty}\frac{1}{K}|M_{n_K}|\sum_{j=K+1}^{\infty}\frac{1}{\sqrt{n_j}} \leq \lim_{K\to\infty}\frac{1}{K}\sqrt{2n_K\ln\ln n_K}\,\frac{\alpha^K}{1-\alpha} = 0,$$

and the result is proved. **Q.E.D.**

6.5 Discussion

In real applications, the distributions of values considered in Section 6.1 would represent the statistical distributions of large sets of outcomes observed sequentially, under some, fairly elementary, stochastic mechanisms. These results can be seen as the analogue of the strong consistency, in Kolmogorov's probability framework, of the usual 'empirical' distribution functions. However, these limiting distributions are obtained without assuming any Kolmogorovian probability distribution P. As far as we know, Theorem 6.4.1 and 6.4.2 represent the first algorithmic versions, in any probability framework, of the strong central limit theorem. Indeed, it seems that in the traditional algorithmic approaches, where a Kolmogorovian probability

distribution P is employed, no version of this result has yet been proved (Vovk, 1995b).

In Chapter 2, in Kolmogorov's probability axiomatics, we noted that, by restricting the class of allowable probabilistic models, a simple application of the standard martingale central limit theorem leads to assessments, in this case significance levels, which respect the prequential principle. That is, which depend only on the actual realized sequence of outcomes and on the actual realized sequence of forecasts. These significance levels, however, even if respecting the prequential principle (in the restricted class of models), lack any 'within-sequence' interpretation. The limiting standard normal distribution, valid under the martingale central limit theorem, has a repeated-sampling interpretation, which refers to all possible realizations. Now, as far as we are concerned, the strong central limit theorem does not lead as well to any asymptotic significance level with a 'within-sequence' interpretation. Even if, appealing to the strong central limit theorem, referring the realized value of the test statistic $S_n/\sqrt{n}$ to the standard normal distribution does not involve any realization that did not materialize, but only an infinite continuation of the finite sequence actually at hand, the strong central limit theorem, just for this reason, does not lead to any asymptotic significance level, with whichever interpretation. Nevertheless, the strong central limit theorem seemed to be the result Seillier-Moiseiwitsch (1986, Chapter 8) was looking for as to provide a probabilistic background for asymptotic significance levels with a 'purely prequential within-sequence' interpretation.

Chapter 7

Conclusions

We conclude with a few words about the interpretation of some of the probabilistic concepts involved in the previous chapters. Like Vovk (1993a), we believe that the only source of probability (using this term very broadly) is probabilistic theories, where for probabilistic theories we mean empirical theories about the world whose interpretation come from the definition of their empirical content. To make a parallel, for us probabilistic theories are theories about some aspect of the world, based on some concept of probability, much as Newton's theory of gravity is a theory about the falling of bodies and the movement of stars and planets, based on the concept of gravity.

PROBABILITY. Traditionally, the concept of Probability (now, using this term in its usual strict sense), for instance, in Kolmogorov's probability axiomatics, is usually provided with either an interpretation admitting repetitions, or an interpretation for which it makes sense only to speak of a single realization of an event. Whereas for the former Kolmogorov's (1933) *propensity* interpretation can, to me, still provide a valid elucidation, for the latter there are, I believe, at least two different understandings. I distinguish between these two different understandings, which very often are not distinguished at all, not on the basis of the different information available (see Vovk, 1993a), but on the basis of the different paradigms which are being used. For me, whereas one interpretation is based on the idea of

proportion of an ideal population, the other (the real epistemic one) is based on the idea of *pure degrees of belief*. When asked about the probability of having a head in a tossing of a coin, or of having a black mouse in a genetic experiment, we almost always look out in our mind for the right proportion in an ideal population of similar events. It does not matter if this population is completely arbitrary or if we are thinking of a real one. The point is that we actually use a population, ideal or real as we please, to give our probabilities. On the other hand, if we were asked on a Friday evening to give the probability of a football team winning the day after or the probability that the distance to the sun is greater than 10^8 km, most of us would not use any ideal population of similar events to state our probabilities. We would hardly think, for instance, of a population of similar football matches considering all the peculiarities that led to that particular one. And we will not think at all of any population when asked about the probability that the 10^{10}th digit of $\pi = 3.1415\ldots$ is one.

Nevertheless, however common it might be to employ interpretations of Probability for a single realization of an event, a serious point could be raised against them, particularly those based on pure degrees of belief, by results like the strong law of large numbers or the calibration theorem. Under a pure degrees of belief interpretation of the concept of Probability, it seems it does not have much meaning to group together the 'outcomes' of uncertain events or quantities like, for example, the marriage of my sister, the height of the Eiffel Tower, the freezing point of mercury, and so on. But if we do it, by considering, for instance, a strong law of large numbers or a calibration theorem, for an infinite sequence of such events, then we have to expect (with a probabilistic degree of belief one) some 'frequency property' even for the outcomes of such events. That is, to an event (the frequency property) which does not have any meaning under a degrees of belief interpretation, we have to give nonetheless probability one. To me, this would seem to suggest that the concept of Probability has in itself an unavoidable interpretation in terms of repetitions. Dawid (1982) argued about the destructive implications for the theory of coherence of the calibration theorem. Here, my suggestion is that there seems

to be some incompatibility in considering limit results like those above, when we adopt an interpretation of the concept of Probability for a single realization of an event. In such a case, a possible solution seems to be, particularly in the case of a pure degrees of belief interpretation, to admit that, in our mind (or in the logical 'mind' of a computer program making probabilistic predictions), due to the impossibility of reducing the complexity of the situations in question, we actually use the concept of Probability with some sort of repetitive interpretation, when comparing two completely unrelated events by giving them similar probabilities.

In Chapter 2, we started our investigation by considering Dawid's prequential principle, as proposed by Dawid (1984, 1991), in the classical probability framework of Kolmogorov. In an attempt to find significance levels (which have a repeated-sampling interpretation under an assumed probability distribution P) satisfying the prequential principle, in Section 2.5.1 and 2.5.2 a simulation study was carried out of a conjecture put forward by Dawid (1991) on the supposed prequential asymptotic behaviour of significance levels based on test statistics Y_n having the form in (2.1). This simulation work did not corroborate the general conjecture, although standardized test statistics, in which Z_i is such that $\mathrm{E}(Z_i|\mathbf{X}^{i-1}) = 0$ and $\mathrm{Var}(Z_i|\mathbf{X}^{i-1}) = 1$, lead to significance levels which are much more 'robust' from a prequential point of view.

The study of the asymptotic behaviour of the above test statistics was also motivated by a more philosophical question. The fact that the test statistic Y_n has an asymptotic standard normal distribution, under some mild conditions, not involving independence, on the form of the underlying probability distribution P, suggested (see, Seillier-Moiseiwitsch (1986), Dawid (1992), and Seillier-Moiseiwitsch and Dawid (1993)), together with other investigations (e. g., Dawid, 1985), that it could have been possible to define, in contrast with the more traditional 'sample-space' interpretations, some sort of Probability with a 'within-sequence' interpretation, for events defined only in terms of the sequence of the actual realized outcomes. However, about such a suggestion, we did not find any reasonable solution, neither in Kolmogorov's probability axiomatics, nor in Vovk's prequential proba-

bility framework, nor in Vovk's purely martingale probability framework. And we now doubt that such a solution could exist.

FAIRNESS. The probabilistic foundations, alternative to Kolmogorov's axiomatics, considered in Chapter 3, had all been inspired by the idea of building a probability framework on a tree-like structure with one-step-ahead probabilities, given the past. In the prequential probability framework proposed by Vovk (1993a), instead of considering a usual filtered probability space $(\Omega^\infty, (\mathcal{F}_n)_{n\geq 0}, \mathcal{F}, P)$, we consider a partially specified probability forecasting system π giving one-step-ahead probabilities, given the past, not necessarily for every finite sequence of past outcomes. This framework was shown to be essentially equivalent, at least in the discrete case, and when π is total, to the event tree framework of Shafer (1985, 1993). In the purely martingale probability framework of Vovk (1993c), neither a probability distribution P nor a probability forecasting system π were introduced. Mathematically, we needed just to introduce a sequence of measurable functions, which we called a basic martingale, and to use the principle of the excluded gambling strategy. Even if from the point of view of the applications, either Vovk's prequential probability framework or Vovk's purely martingale probability framework are essentially equivalent to the classical Kolmogorov axiomatics, these newer probability frameworks are interesting from the point of view of their interpretation. Both these frameworks come with a genuinely sequential interpretation which seems to be more complete than the usual sample-space interpretations of the concept of Probability. Indeed, these traditional interpretations do not give any relevance to the sequential aspect of the observation process. For them, a finite or infinite sequence is always considered as a 'point' in some space, and only the quantity $P(E)$, with $E \in \mathcal{F}$, is provided with an interpretation. However, this sequential interpretation is not based on the idea of Probability, but rather on the idea of fairness, as embodied by the concept of a *martingale*. And it takes account of all possible realizations that did not obtain, not only of the sequence of the actual realized outcomes.

As for the interpretation, in the prequential probability framework, of the one-step-ahead probabilities provided by π, I believe, unlike Vovk (1993a), that these are completely compatible with more that one interpretation of the concept of Probability, and, in particular, with both of the above interpretations for a single realization of an event, namely that based on the idea of proportion of an ideal population and that based on pure degrees of belief. Moreover, I believe that Kolmogorov's propensity interpretation is perfectly compatible with forecasting systems π admitting dependence: we have just to start again from the beginning.

On the problem of the relation of disagreement between theory and observations, in the case of interpretations for a single realization of an event, I think, like Vovk (1993a), that this can only be solved with the introduction of an appropriate principle. With respect to sample-space interpretations of the concept of Probability this would be based on events with very small probabilities; whereas with respect to Vovk's martingale interpretation the measure of disagreement would be based on martingales taking very large values.

RANDOMNESS. In Chapter 4, by introducing some algorithmic constraints, we gave a definition of *random sequences*, which we called M-typical, in a purely martingale framework, on the lines of the algorithmic approach based on the property of typicalness proposed by Martin-Löf (1966). This definition, instead of being given, as in this classical approach, with respect to a probability distribution P, is given only with respect to a basic martingale, by using the principle of the excluded gambling strategy. The idea underlying this approach, and giving an interpretation to it, is that, if we are to play an infinite sequence of fair games against an infinitely rich bookmaker, then, whatever computable strategy we choose, we will never become richer and richer as the game goes on. This martingale interpretation of the concept of randomness provides the most sequential interpretation of the concept of probability (in a broad sense) among those we have considered so far. Nevertheless, even the theory of M-typical sequences, as presented in Chapters 4, 5 and 6, is not based only on the single realized sequence of outcomes, since it cannot

completely avoid considering all the possible realizations that did not materialize. It has already been noted in different forms (von Mises (1951), Dawid (1985), etc.), that the principle of the excluded gambling strategy as a basis for probability has much in common with the principle of the impossibility of perpetual motion as a basis for physics.

In Chapter 5 and 6, these M-typical sequences were shown to satisfy analogues of the classical Kolmogorov strong law of large numbers; of Kolmogorov's upper half of the law of the iterated logarithm; and of Schatte's strong central limit theorem. In Chapter 6 we also investigated the distribution of the values corresponding to a given M-typical sequence, for some basic stochastic sequences, in the case of a basic martingale which was essentially equivalent to a Kolmogorovian probability distribution P. These results, together with the strong central limit theorem, represent an instance in which distributional properties are obtained without using any probability distribution P, or forecasting system π whatsoever. Similarly to the results obtained with the theory of typical sequences of Martin-Löf, these results would reassure us that frequency properties, considered by von Mises as a basis for probability, and, in particular, the distributional properties considered in Chapter 6, are a consequence of the more powerful concepts of fairness and randomness, which in our case are embodied by the concept of an M-typical sequence. From a conceptual point of view, these results would also seem of help in the distinction between the nature of the statistical distributional properties embodied by a probability distribution P, and those embodied by a distribution of the values, or, using a traditional term, by an 'empirical' distribution function. In a way, it would seem that the concept of a distribution of the values, or, equivalently, the concept of an empirical distribution function, still provide the best understanding of a concept of Probability with a sequential within-sequence interpretation.

Appendix A

Computable Functions

In this Appendix we present in a concise and intuitive form some simple properties about operations with real-valued computable functions essential to our definition of M-typical sequences, and to the proof of their properties, which do not seem easy to find in the literature in the form we need, and which can be derived directly from some basic definitions. It is probably worth noting that whereas the classical theory of computability over natural numbers is dealt extensively in the literature and presented in many books, the standard reference for the advanced study of the theory of recursive functions being Rogers (1967), there seems to be a shortage of references dealing with computations over real numbers. Two books which deal explicitly with computable real numbers and computable real functions are Bridges (1994) and Kushner (1984), the former being an introductory book containing just a short account on the topic and the latter a fairly thorough monograph on recursive mathematical analysis. Books which include also an account of algorithmic probability theory are rare; two exceptions are Uspensky and Semenov (1993), and Li and Vitányi (1993), this last being particularly complete.

A.1 Basic Definitions

The basic facts and definitions from the theory of algorithms grouped in this section are taken mainly from the account on computability given in Vovk and V'yugin

(1993), and in some other papers from the same authors, which are already in a form convenient to us.

Intuitively, an *algorithm* is a precise prescription defining a discrete, deterministic process of transforming finite objects. *Finite* objects are objects that can be given by a word in some fixed alphabet, that is, by a finite sequence of symbols from a finite fixed primitive list. Examples of finite objects are given by integers, rational numbers, intervals of the real line with rational end-points, finite sequences of finite objects, but not by real numbers. When an algorithm $\mathcal{U}$ is fed with a finite object it yields as output, in those cases where the process terminates, a finite object. An infinite sequence of finite objects is computable if some algorithm transforms any positive integer i into the ith term of the sequence.

A real number z for which there is an algorithm $\mathcal{U}$ transforming every natural number n into a rational approximation to it to within 2^{-n} is said to be a *computable real number*. For any computable real number there is in general more than one algorithm giving rational approximations to it, possibly differing, to any desired precision, and a computable real number can always be given by any one of these algorithms. We denote the set of all computable real numbers by $\mathbf{R}_c$.

Definition A.1.1 *A real-valued function $f: A \to \mathbf{R}$, where A is a set of finite objects, is computable if there is an algorithm $\mathcal{U}$ which transforms any input $a \in A$ and positive integer n into a rational number r satisfying $|f(a) - r| \leq 2^{-n}$.*

In simple words, f is computable if its values can be computed arbitrarily accurately by some fixed algorithm. It follows from the definition that the values $f(a)$ of a real-valued computable function are computable real numbers.

Definition A.1.2 *A real-valued function $f: A \to \overline{\mathbf{R}}$ is said to be lower or upper semicomputable if there is an algorithm $\mathcal{U}$ which, when fed with a rational number r and an input $a \in A$, eventually stops if $f(a) > r$ or $f(a) < r$ respectively and never stops otherwise.*

This means, for example, that, for a lower semicomputable function f, if $f(a) > r$ this fact will sooner or later be learned, whereas if $f(a) \leq r$ we may be for ever

uncertain. A real-valued function $f: A \to \mathbf{R}$ is computable if and only if it is both lower and upper semicomputable. Sometimes we shall use the expressions computable or lower (upper) semicomputable by $\mathcal{U}$ to indicate the particular algorithm giving to a function the corresponding algorithmic property. For our purposes, we will need to consider only real-valued computable functions defined over sets of finite objects and we will not consider computable functions defined over sets of real numbers or computable real numbers.

The following concepts are sometimes useful in the derivation of the subsequent properties. We term *vicinities* the following finite objects: open intervals of the real line with rational end-points; sets consisting of a single finite object; products of vicinities. We say that a set U is *effectively open* if it is the union of a computable sequence of vicinities.

Then we have that a function $f: A \to \overline{\mathbf{R}}$ is lower (resp. upper) semicomputable if and only if its subgraph $G_l = \{(a,r) : a \in A, r < f(a)\}$ (resp. supergraph $G_u = \{(a,r) : a \in A, r > f(a)\}$) is effectively open. For example, f is lower semicomputable if $G_l = \bigcup_{i=1}^{\infty} V_i$, where $V_i = \{(a,s) : s \in (q_1, q_2)\}$, some $a \in A$, $q_1, q_2 \in \mathbf{Q}$, and there exists an algorithm $\mathcal{U}$ which transforms any positive integer i into the ith term V_i. Also, since a real-valued function is computable if and only if it is both lower and upper semicomputable, a real number f is computable if and only if the rays $(-\infty, f)$ and (f, ∞) are effectively open.

An alternative definition which is sometimes useful is the following. A function $f: A \to \overline{\mathbf{R}}$ is lower semicomputable if and only if the set $D_l = \{(a,q) : a \in A, q \in \mathbf{Q}, q < f(a)\}$ is semicomputable, that is, there exists an algorithm $\mathcal{U}$ which stops when fed with $(a,q) \in D_l$ and never stops otherwise (alternatively we could say that D_l is the domain of a 2-ary partial recursive function).

A.2 Arithmetic Operations

Let us consider now the problem of determining the computability of the result of some arithmetic operations upon computable and lower semicomputable functions.

The computability of the result of transformations like e^t or $\log t$ could then be determined from the computability of these simpler operations by using Taylor series expansions. A treatment of other more elaborate constructions, including Fourier series and Fourier transforms, can be found in Pour-El and Richards (1983). Computability of the result of the following operations is determined directly from the above definitions without introducing any other concept or result.

A.2.1 Computable Functions

Let $f_i: A \to \mathbf{R}$, $i = 1, 2, \ldots$, be computable functions defined over a set A of finite objects. The algorithms computing $f_1, f_2, \ldots$, that is, giving rational approximations to them to any desired precision, are denoted by $\mathcal{U}_1, \mathcal{U}_2, \ldots$ respectively.

Negative. The negative $-f_1$ of a computable function is a computable function. It is trivial to see that, given $\mathcal{U}_1$, an algorithm computing $-f_1$ exists.

Sum. The sum $f_1 + f_2$ of two computable functions f_1 and f_2 is a computable function. Since there exist two algorithms $\mathcal{U}_1$ and $\mathcal{U}_2$ giving rational approximations q_1 and q_2 to f_1 and f_2 to any desired precision, and the addition of two rational numbers is a computable operation, there exists also an algorithm giving rational approximations q to $f_1 + f_2$ to any desired precision. Indeed, for every $a \in A$, to approximate $f_1(a) + f_2(a)$ to 2^{-n}, we just need to consider $q = q_1 + q_2$, where q_1 and q_2 are approximations of $f_1(a)$ and $f_2(a)$ to $2^{-(n+1)}$, since

$$|q - (f_1(a) + f_2(a))|$$
$$= |q_1 + q_2 - f_1(a) - f_2(a)| \leq |q_1 - f_1(a)| + |q_2 - f_2(a)| \leq \frac{1}{2^{n+1}} + \frac{1}{2^{n+1}} = \frac{1}{2^n}.$$

Of course, since we can always add two functions at a time, any finite sum of computable functions is again a computable function. Besides, since $-f_2$ is computable, also the function $f_1 - f_2$ is computable.

Infinite sum. If $f_1, f_2, \ldots$ is a computable sequence of computable functions (that is, there exists an algorithm $\mathcal{U}$ which for any given (i, a, n) gives as output a rational

number r such that $|f_i(a) - r| \leq 2^{-n}$) and $f_i \geq 0$, for all i, then the function $\sum_{i=1}^{\infty} f_i$ is lower semicomputable. Without going into the details, this fact can be proved by considering an algorithm which, for any given (a, q), $q > 0$, tries to 'allocate' q among all the values $f_1(a), f_2(a), \ldots$ and eventually stops if it will succeed and never stops otherwise.

Reciprocal. If $f_1 \neq 0$, then the function $1/f_1$ is computable. For every $a \in A$, since $f_1 \neq 0$, by asking better and better approximations to the algorithm computing f_1, we would know, sooner or later, if $f_1(a)$ is positive or negative, and if 2^{-m}, $m \geq n$, is the minimum precision required for this information, and q is the corresponding rational approximation provided by $\mathcal{U}_1$, we would have that

$$|f_1(a)| \geq |q| - \frac{1}{2^m} > 0.$$

To obtain an approximation of $1/f_1(a)$ to 2^{-n}, we would have then to consider $1/r$, where r is an approximation of $f_1(a)$ to $2^{-\eta}$, and η is the minimum natural number such that $2^\eta \geq 2^m$ and

$$2^\eta \geq \frac{2^{2m+n} + 2^m(2^m|q| - 1)}{(2^m|q| - 1)^2}.$$

In fact, we can see that

$$\left|\frac{1}{r} - \frac{1}{f_1(a)}\right| = \frac{|f_1(a) - r|}{|r||f_1(a)|} \leq \frac{1}{|r|}\frac{1}{|f_1(a)|}\frac{1}{2^\eta} \leq \frac{2^m 2^\eta}{2^m 2^\eta |q| - 2^\eta - 2^m}\ \frac{2^m}{2^m|q| - 1}\ \frac{1}{2^\eta} \leq \frac{1}{2^n}.$$

Product. The product $f_1 f_2$ of two computable functions is computable. To show this, consider the algorithm which, for any given couple (a, n), obtains from $\mathcal{U}_1$ and $\mathcal{U}_2$ the preliminary approximations q_1 and q_2 of $f_1(a)$ and $f_2(a)$ to 2^{-n}, then the final approximations r_1 and r_2 to 2^{-m}, where

$$m = \max\left\{n, 1 + [\log_2(2^n|q_1| + 2^n|q_2| + 5)]\right\},$$

and gives as output the rational $r_1 r_2$. Then

$$|r_1 r_2 - f_1(a) f_2(a)|$$

$$= |r_1(r_2 - f_2(a)) + f_2(a)(r_1 - f_1(a))| \leq |r_1(r_2 - f_2(a))| + |f_2(a)(r_1 - f_1(a))|$$
$$\leq |r_1|\frac{1}{2^m} + |f_2(a)|\frac{1}{2^m} \leq \left(|q_1| + |q_2| + \frac{5}{2^n}\right)\frac{1}{2^m} \leq \frac{1}{2^n}.$$

As a special case of this, note that, if $c \in \mathbf{R}_c$, then the function cf_1 is computable.

Minimum and maximum. The functions $\min(f_1, f_2)$ and $\max(f_1, f_2)$ are both computable. To see this, consider the algorithms which, for any given couple (a, n), feed $\mathcal{U}_1$ and $\mathcal{U}_2$ with (a, n), to obtain the rational approximations q_1 and q_2, and give as output $\min\{q_1, q_2\}$ and $\max\{q_1, q_2\}$ respectively.

Absolute value. The function $|f_1|$ is computable. It is easy to see this by considering an algorithm which, for any given couple (a, n), feeds $\mathcal{U}_1$ with (a, n) to obtain the rational approximation q, and gives as output q, if $q \geq 0$, and $-q$, if $q < 0$.

A.2.2 Lower Semicomputable Functions

Here, $f_i: A \to \overline{\mathbf{R}}$, $i = 1, 2, \ldots$, where A is still a set of finite objects, will be lower semicomputable functions. Similarly to before, we will denote by $\mathcal{U}_1, \mathcal{U}_2, \ldots$ the algorithms lower semicomputing, in terms of Definition A.1.2, the functions $f_1, f_2, \ldots$ respectively.

Negative. The negative $-f_1$ is upper semicomputable. To see this just consider the algorithm which, for any given couple (a, q), feeds $\mathcal{U}_1$ with $(a, -q)$, and gives as output the output of $\mathcal{U}_1$. In this way, the algorithm will eventually stop when $-q < f_1(a)$, that is when $q > -f_1(a)$, and never stop otherwise, as required by the definition of upper semicomputable function.

Sum. The sum $f_1 + f_2$ of two lower semicomputable functions is again a lower semicomputable function. This can be shown by using the alternative definition of lower semicomputable functions in terms of their effectively open subgraphs. An algorithm lower semicomputing $f_1 + f_2$, that is generating a sequence of vicinities covering its subgraph, can be built by running simultaneously the algorithms, which

can be determined from $\mathcal{U}_1$ and $\mathcal{U}_2$, generating the coverings of the subgraphs of f_1 and f_2 respectively.

Infinite sum. If $f_1, f_2, \ldots$ is a lower semicomputable sequence of lower semicomputable functions (in the sense that, in Church's λ-notation, the function $\lambda ia.f_i(a)$ which transforms i, a into $f_i(a)$ is lower semicomputable) and $f_i \geq 0$, for all i, then $\sum_{i=1}^{\infty} f_i$ is a lower semicomputable function. Similarly to the case of an infinite sum of computable functions, it is possible to find an algorithm lower semicomputing this infinite sum by trying, for any given (a, q), $q > 0$, to 'allocate' q among all the values $f_1(a), f_2(a), \ldots$.

Reciprocal. If $f_1 \geq 0$, then the function $1/f_1$ is upper semicomputable. To see this, we have just to consider an algorithm which, for any given couple (a, q), never stops, if $q \leq 0$, and gives as output the output of the algorithm $\mathcal{U}_1$, when fed with $(a, 1/q)$, if $q > 0$. In fact, $\mathcal{U}_1$ would eventually stop for $f(a) > 1/q$, that is, for $1/f(a) < q$, and never stop otherwise.

Note, that, in the same way, if $f_1 \geq 0$, then the lower semicomputability of f_1 can be derived from the upper semicomputability of $1/f_1$. Besides, it is also easy to see that if $f_1 \leq 0$ and $f_1 > -\infty$, then $1/f_1$ is upper semicomputable.

Product. If $f_1, f_2 \geq 0$, then the product $f_1 f_2$ is a lower semicomputable function. An algorithm generating a sequence of vicinities covering the subgraph of $f_1 f_2$ can be obtained by running simultaneously the algorithms generating the coverings of the subgraphs of f_1 and f_2.

Moreover, for $c \in \mathbf{R}_c$, the function cf_1 is: computable, if $c = 0$; lower semicomputable, if $c > 0$; upper semicomputable, if $c < 0$.

Minimum and maximum. The functions $\min(f_1, f_2)$ and $\max(f_1, f_2)$ are both lower semicomputable. To show that $\min(f_1, f_2)$ is lower semicomputable, we need just to consider an algorithm which, for any given couple (a, q), feeds $\mathcal{U}_1$ and $\mathcal{U}_2$ with (a, q), and stops when both algorithms stop, and never stops otherwise. On the

other hand, to show that $\max(f_1, f_2)$ is lower semicomputable, we need to consider an algorithm which, for any (a, q), feeds $\mathcal{U}_1$ and $\mathcal{U}_2$ with (a, q), and stops when either one or the other of the two algorithms stops and never stops otherwise, running $\mathcal{U}_1$ and $\mathcal{U}_2$ simultaneously.

Absolute value Here, unlike for the computability case, from the lower semicomputability of f_1, we cannot deduce the lower, or upper, semicomputability of $|f_1|$. All that we can show is the lower semicomputability of $|f_1|$, when f_1 is considered only on the subset of A on which it is negative, or the upper semicomputability of $|f_1|$, when f_1 is considered only on the subset of A on which it is positive.

A.3 A Note

Let us make a technical note about the introduction of algorithmic notions in probability theory. Broadly speaking, following the classification given by Uspensky and Semenov (1993), algorithmic studies in mathematical analysis could be classified either as constructive analysis, computable analysis, or partly computable analysis.

In *constructive analysis*, at least in some of its strongest ramifications, we are not allowed to use either the law of the excluded middle, which permits the abstraction to actual infinity and the use of indirect proofs, or the axiom of choice, which is, for instance, necessary in the proof of the countable additivity of Lebesgue measure. As a consequence of this requirement, in this approach we are only allowed to deal with constructive, or computable, objects, and ideally everything should have to be given or to be proved in a constructive way.

In *computable analysis* and *partly computable analysis*, on the other hand, the problem of computability is addressed without any logical restriction and in particular the law of the excluded middle and the axiom of choice can still be used. The difference between these two less extreme approaches lies in the fact that while in the former we only deal with computable objects, in the latter (also called approximational approach) we consider the computable objects as a subset of all

objects. For instance, while in the latter we can still deal with functions defined over real numbers, as in classical mathematical analysis, in the former we can deal just with functions defined over computable real numbers, which take values in the computable real numbers. A classical result which can help to highlight the differences between the two approaches is probably given by the existence of monotonic bounded sequences of computable real numbers which do not converge to a computable real number. In this case, while in partly computable analysis such sequences are still considered to converge, in computable analysis they are not, and in this last approach a monotonic bounded sequence does not always converge.

Now, as far as we are concerned, most of the studies in algorithmic probability theory, either in the traditional approach using probability distributions or in the newer approach using martingales, could be classified either as partly computable analysis or as computable analysis. In our previous definitions, for instance, the algorithm computing a real-valued computable function did not have to be actually specified, as would have been required by a constructive logic, but needed just to exist. It is also common practice, in most of algorithmic probability theory, to actually consider functions defined over sets of real numbers and to ask for their computability only as a later requirement, and so to follow an approximational approach.

Bibliography

BASU, D. (1975). Statistical information and likelihood. *Sankhyā* A **37**, 1–71.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57**, 269–326.

BRIDGES, D. S. (1994). *Computability. A Mathematical Sketchbook.* Springer-Verlag, New York.

BROSAMLER, G. (1988). An almost everywhere central limit theorem. *Math. Proc. Cambridge Philos. Soc.* **104**, 561–574.

BUEHLER, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.* **30**, 845–863.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.

DAWID, A. P. (1982). The well-calibrated Bayesian (with discussion). *J. Amer. Statist. Assoc.* **77**, 605–613.

DAWID, A. P. (1983). Inference, statistical: I. In *Encyclopedia of Statistical Sciences* (eds S. Kotz, N. L. Johnson and C. B. Read), vol. 4, pp. 89–105. Wiley-Interscience, New York.

DAWID, A. P. (1984). Statistical theory. The prequential approach (with discussion). *J. R. Statist. Soc.* A **147**, 278–292.

DAWID, A. P. (1985). Calibration-based empirical probability (with discussion). *Ann. Statist.* **13**, 1251–1285.

DAWID, A. P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences* (eds S. Kotz, N. L. Johnson and C. B. Read), vol. 7, pp. 210–218. Wiley-Interscience, New York.

DAWID, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *J. R. Statist. Soc.* B **53**, 79–109.

DAWID, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference (with discussion). *Bayesian Statistics 4* (eds J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith), pp. 109–125. Oxford University Press, Oxford.

DAWID, A. P. (1993). Discussion on "A logic of probability, with application to the foundations of statistics" by V. G. Vovk. *J. R. Statist. Soc.* B **55**, 341–343.

DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.

ELLIOTT, P.D.T.A. (1979). *Probabilistic Number Theory I. Mean-Value Theorems.* A series of comprehensive studies in mathematics, vol. 239. Springer-Verlag, New York.

ELLIOTT, P.D.T.A. (1980). *Probabilistic Number Theory II. Central Limit Theorems.* A series of comprehensive studies in mathematics, vol. 240. Springer-Verlag, New York.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume I, Third edition, Revised.* Wiley, New York.

DE FINETTI, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68. (Engl. transl. (1964): Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability* (eds H. E. Kyburg and H. E. Smokler), pp. 93–158. Wiley, New York.)

FISHER, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700–725.

FISHER, R. A. (1956a). On a test of significance in Pearson's *Biometrika Tables* (No. 11). *J. R. Statist. Soc.* B **18**, 56–60.

FISHER, R. A. (1956b). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

HARTMAN, P. and WINTNER, A. (1941). On the law of the iterated logarithm. *Amer. J. Math.* **63**, 169–176.

HOWARD, J. V. (1993). Discussion on "A logic of probability, with application to the foundations of statistics" by V. G. Vovk. *J. R. Statist. Soc.* B **55**, 343–344.

KOLMOGOROFF, A. (1929). Ueber das Gesetz des iterierten Logarithmus. *Math. Ann.* **101**, 126–135.

KOLMOGOROV, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin. (Engl. transl. (1950): *Foundations of the Theory of Probability.* Chelsea, New York.)

KOLMOGOROV, A. N. and USPENSKII, V. A. (1987). Algorithms and randomness. *Theory Probab. Appl.* **32**, 389–412.

KUSHNER, B. A. (1984). *Lectures on Constructive Mathematical Analysis.* Translations of mathematical monographs, vol. 60. American Mathematical Society, Providence, Rhode Island.

LACEY, M. T. and PHILIPP, W. (1990). A note on the almost everywhere central limit theorem. *Statist. Probab. Lett.* **9**, 201–205.

LI, M. and VITÁNYI, P. (1993). *An Introduction to Kolmogorov Complexity and Its Applications.* Springer-Verlag, New York.

MARCINKIEWICZ, J. and ZYGMUND, A. (1937a). Sur les fonctions indépendantes. *Fund. Math.* **29**, 60–90.

MARCINKIEWICZ, J. and ZYGMUND, A. (1937b). Remarque sur la loi du logarithme itéré. *Fund. Math.* **29**, 215–222.

MARTIN-LÖF, P. (1966). The definition of random sequences. *Informn. Control* **9**, 602–619.

VON MISES, R. (1951). *Wahrscheinlichkeit, Statistik und Wahrheit.* Springer, Vienna. (Engl. transl. (1957): *Probability, Statistics and Truth.* Hodge, London.)

ODIFREDDI, P. (1992). *Classical Recursion Theory. The Theory of Functions and Sets of Natural Numbers.* Studies in logic and the foundations of mathematics, vol. 125. North-Holland, Amsterdam.

POUR-EL, M. B. and RICHARDS, I. (1983). Computability and noncomputability in classical analysis. *Transactions of the American Mathematical Society* **275**, 539–560.

RÉVÉSZ, P. (1968). *The Laws of Large Numbers.* Academic Press, New York.

RISSANEN, J. (1989). *Stochastic Complexity in Statistical Enquiry.* World Scientific, Singapore.

ROGERS, H. (1967). *Theory of Recursive Functions and Effective Computability.* McGraw-Hill, New York.

ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23**, 470–472.

SCHATTE, P. (1988). On strong versions of the central limit theorem. *Math. Nachr.* **137**, 249–256.

SCHNORR, C. P. (1971). A unified approach to the definition of random sequences. *Math. Syst. Theory* **5**, 246–258.

SCHNORR, C. P. (1977). A survey of the theory of random sequences. In *Basic Problems in Methodology and Linguistics* (eds R. E. Butts and J. Hintikka), pp. 193–211. Reidel, Dordrecht.

SEILLIER-MOISEIWITSCH, F. (1986). Assessment of sequential probabilistic forecasting procedures. PhD Thesis, Department of Statistical Science, University College London, London.

SEILLIER-MOISEIWITSCH, F. and DAWID, A. P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.* **88**, 355–359.

SEILLIER-MOISEIWITSCH, F., SWEETING, T. J. and DAWID, A. P. (1992). Prequential tests of model fit. *Scand. J. Statist.* **19**, 45–60.

SHAFER, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton.

SHAFER, G. (1985). Conditional probability (with discussion). *Int. Statist. Rev.* **53**, 261–277.

SHAFER, G. (1990). The unity and diversity of probability (with discussion). *Statist. Sci.* **5**, 435–462.

SHAFER, G. (1991). What is probability? In *Perspectives on Contemporary Statistics* (eds D. C. Hoaglin and D. S. Moore), to appear. Mathematical Association of America.

SHAFER, G. (1993). Can the various meanings of probability be reconciled? In *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (eds G.Keren and C. Lewis), pp. 165–196. Erlbaum, Hillsdale, New Jersey.

SHAFER, G. (1995). *The Art of Causal Conjecture.* MIT Press.

SHIRYAYEV, A. N. (1984). *Probability.* Springer, Berlin.

SPIEGELHALTER, D. J., DAWID, A. P., LAURITZEN, S. L. and COWELL, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statist. Sci.* **8**, 219–283.

STOUT, W. F. (1970a). A martingale analogue of Kolmogorov's law of the iterated logarithm. *Z. Wahrscheinlichkeitstheorie verw. Geb.* **15**, 279–290.

STOUT, W. F. (1970b). The Hartman–Wintner law of the iterated logarithm for martingales. *Ann. Math. Statist.* **41**, 2158–2160.

STOUT, W. F. (1974). *Almost Sure Convergence.* Academic Press, New York.

USPENSKII, V. A., SEMENOV, A. L. and SHEN', A. KH. (1990). Can an individual sequence of zeros and ones be random? *Russian Math. Surveys* **45**, 121–189.

USPENSKY, V. and SEMENOV, A. (1993). *Algorithms: Main Ideas and Applications.* Kluwer Academic Publishers, Dordrecht.

VILLE, J. (1939). *Etude Critique de la Notion de Collectif.* Gauthier-Villars, Paris.

VOVK, V. G. (1988a). Kolmogorov–Stout law of the iterated logarithm. *Math. Notes* **44**, 502–507.

VOVK, V. G. (1988b). The law of the iterated logarithm for random Kolmogorov, or chaotic, sequences. *Theory Probab. Appl.* **32**, 413–425.

VOVK, V. G. (1990a). Prequential probability theory. Submitted to *Probability Theory and Related Fields.*

VOVK, V. G. (1990b). Prequential variants of the central limit theorem and the law of the iterated logarithm. Submitted to *Probability Theory and Related Fields.*

VOVK, V. G. (1991). Finitary prequential probability: asymptotic results. Unpublished manuscript.

VOVK, V. G. (1993a). A logic of probability, with application to the foundations of statistics (with discussion). *J. R. Statist. Soc.* B **55**, 317–351.

VOVK, V. G. (1993b). Forecasting point and continuous processes: prequential analysis. *Test* **2**, 189–217.

VOVK, V. G. (1993c). A purely martingale version of Kolmogorov's strong law of large numbers. *Theory Probab. Appl.*, to appear.

VOVK, V. G. (1995a). A purely martingale version of Lindeberg's central limit theorem. Unpublished manuscript.

VOVK, V. G. (1995b). Personal communication.

VOVK, V. G. and V'YUGIN, V. V. (1993). On the empirical validity of the Bayesian method. *J. R. Statist. Soc.* B **55**, 253–266.

VOVK, V. G. and V'YUGIN, V. V. (1994). Prequential level of impossibility with some applications. *J. R. Statist. Soc.* B **56**, 115–123.

WILLIAMS, D. (1991). *Probability with Martingales.* Cambridge University Press, Cambridge.