RANDOM SEQUENCES

ACADEMISCH PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR AAN DE UNIVERSITEIT VAN AMSTERDAM, OP GEZAG VAN DE RECTOR MAGNIFICUS, DR. S.K. THODEN VAN VELZEN, HOOGLERAAR IN DE FACULTEIT DER TANDHEELKUNDE, IN HET OPENBAAR TE VERDEDIGEN IN DE AULA DER UNIVERSITEIT (OUDE LUTHERSE KERK, SINGEL 411, HOEK SPUI) OP WOENSDAG 16 SEPTEMBER 1987 TE 13.30 UUR

DOOR

MICHIEL VAN LAMBALGEN

GEBOREN TE KRIMPEN AAN DE IJSSEL

Promotores: Prof. Dr. J.F.A.K. van Benthem
Prof. S.J. Doorman M.Sc.

Tout a dejà été éprouvé et expérimenté à mille réprises, mais souvent sans avoir été dit, ou sans que les paroles qui le disent subsistent, ou si elles le font, nous soient intelligibles et nous émeuvent encore. Comme les nuages dans le ciel vide, nous nous formons et nous dissipons sur ce fond d'oubli.

Marguerite Yourcenar
Archives du Nord

Preface

The research on the foundations of probability reported here was part of project 22 – 110 of the Dutch Foundation for Scientific Research ZWO, whose support is gratefully ackowledged.
Some of the central concerns of this essay took shape in a long afternoon stroll in Salzburg with Roger Cooke, who also intensely supervised their subsequent development. In spite of the heat generated by our debates, they have resulted in a state of low entropy in which few differences of opinion remain.
I owe a special debt to Johan van Benthem, who steered me toward mathematical logic long ago. The example of his technical acumen and systematic philosophizing has inspired me ever since. His already overburdened schedule notwithstanding, he took over part of my teaching duties during the last phase of the writing.
I am grateful to Joop Doorman, whose insistence on a philosophical justification of what was originally a collection of rather technical results, finally produced Chapter 2.
Lastly, I thank Mike Keane and Guus Balkema for mathematical assistance.
The cover was designed by Hans Sprangers.

Contents

1 Introduction

It may be taken for granted that any attempt at defining disorder in a formal way will lead to a contradiction. This does not mean that the notion of disorder is contradictory. It is so, however, as soon as I try to formalize it.

Hans Freudenthal

0111000101001000..... is an initial segment of a long sequence produced by tossing a coin. In its broadest outline, the subject of this thesis is the mathematical description of the sequences produced by random processes (such as coin tossing), which will be called *random sequences*. The tantalizing motto, taken from Freudenthal [29], expresses a negative verdict on this enterprise. But meanwhile it raises a no less interesting question: How can a non-contradictory concept *necessarily* defy formalisation?

Clearly, the two definite articles in the phrase "the mathematical description of the sequences produced by random processes" present a host of problems. There might not be such a description, as Freudenthal thinks; or it might be completely trivial, the reason being that all we can say a priori on the sequences produced by, say, coin tossing is, that these are sequences of zeros and ones. On the other hand, there *do* exist various definitions of random sequences; perhaps even too many.

The discussion in the pages that follow is therefore concentrated on two main questions:

1. Is a mathematical definition of random sequences possible and if so, why should one want to give such a definition?
2. Given the fact that various definitions have been proposed, does it make sense to ask for criteria which allow us to choose between them?

Even apart from its usefulness, the possibility of providing a definition of randomness has often been doubted. Here is a grab-bag of some of the a priori reasons which have been adduced for this conviction:

– As soon as you can define randomness, it ceases to be true randomness;
– Randomness is a property of processes, not of the sequences generated by such processes;
– It is characteristic of a random process that it may generate *any* sequence.

For the moment, we shall leave these a priori arguments unanalyzed.

The first person to discuss systematically the possibility, and indeed the necessity, of a definition of random sequences was Richard von Mises, who provided an axiomatisation of probability theory with "random sequence" as a primitive term. He argued, as did later

Kolmogorov, that, if probability is interpreted as relative frequency, then the applicability of probability theory to real phenomena (which is amply verified) entails that these phenomena must have certain properties of randomness, and he proposed to take these properties as basic for a definition of random sequences (other properties being optional).
A definition of randomness is therefore necessary to explain the applicability of probability theory and a minimal set of properties random sequences have to satisfy can be deduced from its rules, a priori reasons for the impossibility of such a definition notwithstanding. But in a philosophical analysis we must of course investigate the apparent conflict between compelling physical reasons *pro* and a priori reasons *contra* a mathematical definition of randomness.

Although in the thirties, but also in recent years, there has been a lively commerce in definitions of randomness, our second question, namely: Do there exist criteria to choose between these definitions?, has not been explicitly discussed in the literature. To be sure, there have been discussions among partisans of various schools, the Geneva conference on the foundations of probability (1937) being a notable example. But one is struck by the sheer monotony of these discussions, the same arguments pro and con being repeated over and over again, without noticible effects upon the opinions of the discussants.
We propose to break this stalemate by analyzing possible sources for the lack of mutual comprehension so clearly displayed. The conclusion of our analysis will be that von Mises, around whose axiomatisation of probability theory the discussion centered, had views on the foundations of probability and of mathematics in general, which were not shared, but also not fully understood, by his critics. His view on the foundations of mathematics, usually expressed only implicitly, was shaped by his work as an applied mathematician and is a mixture of constructivism and a tendency to introduce bold concepts whenever the description of real phenomena seems to necessitate it. From this mixture results what one might call an inhomogeneous mathematical universe, which is a far cry from the very homogeneous set theoretical universe that inspired some of the objections of his critics. Unfortunately, these assumptions were not made explicit in the debate. Von Mises' views on the foundations of probability did, of course, figure explicitly and prominently in the debate; but its critics did not show a reciprocal awareness of their own assumptions, thereby successfully creating the impression that, while von Mises' view was unnecessarily complicated, theirs was simplicity itself. We believe that this debate, if analyzed correctly, points to the conclusion that different (objective) interpretations of probability lead to different requirements for random sequences. Hopefully, this point of view is helpful in understanding the debate that raged between von Mises and his critics; and if it directs the reader away from bickering about definitions of randomness and toward the deeper questions of the foundations of probability, it has fulfilled its purpose.

These two main questions determine much of the technical work in Chapters 3 to 5. Given that there exist different types of definitions, each with its own minor variants, we must investigate how these definitions are related extensionally. Some of these relationships are well know, e.g. that between randomness in the sense of Martin-Löf and definitions involving (variants of) *Kolmogorov complexity*. In these cases, the novelty of our treatment consists solely in introducing new proof techniques. But other relations have been studied less thoroughly, notably that between von Mises' semi-formal definition, based on so called *admissible place selections* , and the other types. There are obvious reasons for this lack of attention. The fact that von Mises' definition is not quite formal renders a comparison with the other definitions, which *are* rigorous in the modern sense, difficult; and often the need for such a comparison is not felt acutely because, say, Martin-Löf's definition is considered to be an *improvement* on von Mises' proposal, rather than an *alternative*, based on radically different principles. We therefore have to study ways in which to make von Mises' definition precise; moreover, these attempts to instill precision should be based as much as possible upon his own philosophical premises. This problem has not been solved entirely, but the results that have been obtained, do enable us to effect a rigorous comparison between von Mises' definition and the other types.

So far, we have been concerned with *extensional* relationships between different *types* of definitions. But the exact meaning and justification of the definitions is no less important. The rationale behind von Mises' definition is studied extensively in Chapter 2. It turns out that it is completely justified on von Mises' own interpretation of probability, to be called *strict frequentism* . His opponents, on the other hand, usually start from a very different interpretation of probability, the *propensity interpretation*, and it is this interpretation which inspires those modern definitions of randomness which, following Martin-Löf, characterise randomness as the satisfaction of certain statistical tests. The choice of the particular type of test employed by Martin-Löf is, in our opinion, debatable, and the conclusion of the discussion is, in a nutshell, that whereas von Mises' definition is philosophically rigorous, but technically less so, with Martin-Löf's definition it is just the other way around.

The most promising approach to the characterisation of randomness appears to be the one inaugurated by Kolmogorov, using a notion of *complexity* for finite sequences. Philosophically, it stands midway between the definitions of von Mises and of Martin-Löf. Kolmogorov accepts von Mises' view that an analysis of the conditions of applicability of probability theory necessarily leads to the concept of a random sequence (although they differ in the place they allot to random sequences in the formal structure of probability theory). But while von Mises requires of random sequences only those properties which make the deductions of probability theory go through, Kolmogorov goes further and in a sense explains,

non-probabilistically, why random sequences must have those properties.
Technically, a major advantage of Kolmogorov complexity is, that it allows us to discuss degrees of randomness, both of sequences as a whole and within a single sequence. This feature leads to new problem which cannot be posed in the other frameworks, such as: What is the connection (if any) between the arithmetical complexity and the Kolmogorov complexity of a sequence? What is the connection between traditional measures of disorder such as entropy in its various forms, and Kolmogorov complexity?

The structure of this book is as follows. It consists of two parts, the first historical, the second technical, which can by and large be read independently. Each of the technical Chapters (3 to 5) has its own non-technical introduction, setting out the reasons for the constructions that follow. Of course, for the full motivation of the technical work, the reader is referred to the historical part.
Chapter 2 deals with von Mises' semi-formal definition of random sequences and its function in his axiomatisation of probability theory. We believe that insufficient attention to this context has lead to wholly unjustified criticisms of von Mises' theory. Accordingly, an overview of, and a critical commentary on, the debate to which von Mises' introduction of random sequences gave rise, will occupy half of the chapter. As intimated already, one of the main conclusions of our analysis will be that the evident lack of mutual understanding displayed in this debate is a consequence of widely divergent interpretations of probability.

The second, technical, part of the thesis opens with a chapter on Martin-Löf's definition of random sequences via statistical tests, and some of its variants. The emphasis of the discussion is, technically, on unified methods of proof, and, philosophically, on the virtues and vices of the particular type of test adopted by Martin-Löf. The fourth chapter is an updated version of [53] and contains results relating the definition of randomness of von Mises and that of Martin-Löf and its variants. In particular we study the behaviour of Martin-Löf random sequences under so called place selections (theorem 4.5.2) and we give a new proof of a famous theorem due to Ville, the philosophical implication of which is discussed in 2. We hope that this proof (4.6.1) has more explanatory power than Ville's original combinatorial argument.
Chapter 5 is concerned with what we consider to be the most promising development incited by von Mises' original proposal: Kolmogorov complexity. After studying several of its variants, we settle for the definition proposed by Chaitin, which allows an equivalent condition for randomness in the sense of Martin-Löf (5.4.3). As mentioned above, the decisive advantage of a complexity theoretic definition vis à vis other definitions of randomness is that it also enables us to measure the degree of randomness of a sequence, both locally, as a function of the initial segments of the sequence, and globally, as a number

attached to the sequence as a whole. The local behaviour of the degree of randomness is studied in a section on complexity oscillations (5.4, especially 5.4.2-3); the global behaviour is compared with measures of disorder defined in ergodic theory, such as metric entropy (5.5.2) and topological entropy (5.5.3).
The appendix (Chapter 6) contains notations and definitions not explained in the text.

Passages detracting from the main argument are labelled **Digression**. Those who are interested more in philosophical vistas than in technical details will find at the beginning of each chapter a list of sections which do not bear directly on foundational issues. But some mathematics is necessary; there is no royal road to the philosophy of science.

2 Roots of Randomness: Von Mises' Definition of Random Sequences

2.1 Introduction In 1919 Richard von Mises (1883–1957) published an (in fact the first) axiomatisation of probability theory which was based on a particular type of disorderly sequences, so called *Kollektivs.* The two features characterizing Kollektivs are, on the one hand, existence of limiting relative frequencies within the sequence (global regularity) and, on the other hand, invariance of these limiting relative frequencies under the operation of "admissible place selection" (local irregularity). An admissible place selection is a procedure for selecting a subsequence of a given sequence x in such a way that the decision to select a term x_n does not depend on the value of x_n.

After several years of vigorous debate, which concerned not only von Mises' attempted characterisation of a class of random phenomena, but also his views on the interpretation of probability, it became clear that most probabilists were critical of von Mises' axiomatisation and preferred the simple set of axioms given in Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung* of 1933. The defeat of von Mises' theory was sealed at a conference on probability theory in Geneva (1937), where Fréchet gave a detailed account of all the objections that had been brought to bear against von Mises' approach.

We believe that this debate, for all its vigor, has failed to produce a careful analysis of von Mises' views and the new concepts he introduced. In fact, when one reads the various contributions, one is immediately struck by its monotony: the same objections and refutations are repeated over and over again, with scarcely any new elements being brought in. (There is one major exception: the objections based on a construction due to Ville (1937).) When one takes into account the considerable scientific acumen of the participants in the debate, this monotony may be a cause for surprise.

In the following pages, we shall attempt both to analyse von Mises' theory in detail and to examine the reasons why the debate which ensued after its publication failed to lead to satisfaction. Our guiding principles in this analysis will be twofold.

First, we believe that von Mises' characterisation of random sequences has great intuitive appeal, for all its imprecision. We do not regard the lack of precision itself as objectionable. Instead, we subscribe to Kreisel's doctrine of *informal rigour*:

> The 'old fashioned' idea is that one obtains rules and definitions by analysing intuitive notions and putting down their properties. This is certainly what mathematicians thought they were doing when defining length or area, or, for that matter logicians when finding rules of inference or axioms (properties) of mathematical structures such as the continuum. [...] What the 'old fashioned' idea assumes is quite simply that the intuitive notions are *significant* , be it in the external world or in thought (and a *precise* formulation of what is significant in a subject is the result, not the starting point of research into that subject).
> Informal rigour wants (i) to make this analysis as precise as possible (with the means available), in particular to eliminate doubtful properties of the intuitive notions when drawing conclusions about them; in particular not to leave undecided questions which can be decided by full use of evident properties of these intuitive notions [52,138].

It will be seen, for instance, that the notion of Kollektiv is at least clear enough to refute the often repeated allegiation of inconsistency. We do not, however, claim to have reached the limits of analysis (but perhaps the idea of the ultimate analysis does not even make sense).

Second, we try to explain the sterility of the debate by assuming that the participants had widely diverging, but in part unarticulated, opinions on the foundations of mathematics and probability. We shall meet instances of this phenomenon when we discuss the alleged inconsistency of Kollektivs (in 2.3.3) and the force of Ville's objection (in 2.6.2).
The conclusion of our analysis will be that the criticisms directed against von Mises' theory are either misguided (such as the charge that von Mises was working with a wrong concept of what axiomatisation should be) or based on foundational views which are not his (the alleged inconsistency, or the objection that Kollektivs do not always satisfy the law of the iterated logarithm). One may then pursue the debate at the level of foundational issues, but here, it is much more difficult to decide who is right and who is wrong. And for our purpose, the conclusion that different views on the foundations of probability may lead to different requirements on definitions of random sequences, is sufficient to motivate the technical work of subsequent chapters.

The plan of this chapter is as follows. In 2.2 we examine von Mises' version of the frequency interpretation, its surprising consequences and its possible rival, the propensity interpretation. In 2.3 we introduce Kollektivs and discuss their metamathematical status. 2.4 is centered around the demonstration that any form of the frequency interpretation assumes that the phenomena to which it is applicable are Kollektivs. In 2.5 we study some of the attempts to achieve precision in the definition of Kollektivs. 2.6 is devoted to a discussion of the objections brought forth by Fréchet. Our conclusions will be summed up in 2.7.
It will be clear from this outline that we shall mostly be concerned with two problems and their relation: the interpretation of probability and the definition of random sequences.

2.2 The frequency interpretation of probability

2.2.1 Methodological considerations. In the early thirties, two books were published on the foundations of probability theory, which express widely divergent attitudes: the *Wahrscheinlichkeitsrechnung*, von Mises' definitive treatise (1931) and the *Grundbegriffe der Wahrscheinlichkeitsrechnung* by Kolmogorov (1933). A convenient starting point for a discussion of von Mises' views is given by the following juxtaposition of quotations:

> Die Wahrscheinlichkeitstheorie als mathematische Disziplin soll und kann genau in demselben Sinne axiomatisiert werden wie die Geometrie oder die Algebra. Das bedeutet, daß, nachdem die Namen der zu untersuchenden Gegenstände und ihrer Grundbeziehungen sowie die Axiome, denen diese Grundbeziehungen zu gehorchen haben, angegeben sind, die ganze weitere Darstellung sich ausschließlich auf diese Axiome gründen soll und keine Rücksicht auf die jeweilige konkrete Bedeutung dieser Gegenstände und Beziehungen nehmen darf.
>
> Dementsprechend wird im §1 der Begriff eines *Wahrscheinlichkeitsfeldes* als eines gewissen Bedingungen genügenden Mengensystems definiert. Was die Elemente dieser Mengen sind, ist dabei für die mathematische Entwicklung der Wahrscheinlichkeitsrechnung völlig gleichgültig (man vergleiche die Einführung der geometrische Grundbegriffe in HILBERTs "Grundlagen der Geometrie" oder die Definitionen von Gruppen, Ringen und Körpern in der abstrakten Algebra).
>
> Jede axiomatische (abstrakte) Theorie läßt bekanntlich unbegrenzt viele konkrete Interpretationen zu. In dieser Weise hat auch die mathematische Wahrscheinlichkeitstheorie neben derjenigen ihrer Interpretationen, aus der sie aufgewachsen ist, auch zahlreiche andere. Wir kommen so zu Anwendungen der mathematische Wahrscheinlichkeitstheorie auf Untersuchungsgebiete, die mit den Begriffen des Zufalls und der Wahrscheinlichkeit im konkreten Sinne dieser Begriffe nichts zu tun haben (Kolmogorov [44,1]).

> Die Wahrscheinlichkeitstheorie wird in dieser Vorlesungen aufgefaßt als eine *mathematische Naturwissenschaft* von der Art etwa wie die Geometrie oder die Mechanik. Ihr Ziel ist es, für eine bestimmte Gruppe beobachtbarer Erscheinungen, die Massenerscheinungen und Wiederholungsvorgänge, eine übersichtliche Beschreibung zu geben, wie sie die Geometrie für die räumlichen, die Mechanik für die Bewegungserscheinungen liefert. An der Spitze einer derartigen Theorie stehen Aussagen, durch die die Grundbegriffe definiert werden und die man oft Axiome nennt; in ihnen kommen allgemeine Erfahrungseinhalte zur Verwertung, ohne daß sie unmittelbar als Erfahrungssätze angesprochen werden dürften. Aus den Axiomen werden dann auf deduktivem Wege, oder, wie man jetzt besser sagt, durch "tautologische Umformungen" mannigfache Sätze gewonnen, die vermöge des Zusammenhanges, der zwischen den Grundbegriffen und der Erfahrungswelt besteht, bestimmten, durch Beobachtung nachprüfbaren Tatbeständen entsprechen. So weist die Theorie am Anfang und am Ende jeder Gedankenreihe Berührung mit der Welt der Beobachtungen auf; ihren eigentlichen Inhalt aber, der uns vorzugsweise beschäftigen wird, bilden die rein *mathematischen Überlegungen, die zwischen dem Anfang und dem Ende stehen* (von Mises [68,1]).

These quotations emphasize two different aspects of the mathematical method. The quotation from Kolmogorov is concerned mainly with faultless derivations from axioms, which should proceed regardless of the actual meanings of the primitive concepts involved.

Von Mises, of course, does not deny the importance of this procedure, but he stresses the role of mathematics in describing real structures, in as much detail as is necessary, a feature less prominent in Kolmogorov's book.
The following quotation from von Mises' *Wahrscheinlichkeit, Statistik und Wahrheit* [70] further clarifies the sense in which probability theory is *mathematische Naturwissenschaft* :

> Die Wahrscheinlichkeitsrechnung (oder die Theorie der zahlenmäßig erfaßbaren Wahrscheinlichkeiten) ist die Theorie bestimmter, der Beobachtung zugänglicher Erscheinungen, der Wiederholungs– und Massenvorgänge etwa vom Typus der Glücksspiele, der Bevölkerungsbewegung, der Bewegung Brownscher Partikel usf. Das Wort "Theorie" ist hier in demselben Sinn gemeint wie die Hydromechanik Theorie der Flüssigkeitsströmungen, die Thermodynamik Theorie der Warmevorgänge, die Geometrie Theorie der räumlichen Erscheinungen heißt [70,128].

Statements such as these have led critics (e.g. Feller in his talk at the Geneva conference on probability theory [23,9]; see also Fréchet [28]) to object that von Mises' conception of a scientific theory was not true to the example set by Hilbert's *Grundlagen* and confused mathematical and empirical considerations; and since Kolmogorov's theory did not fall prey to this alleged confusion, it had to be preferred.

This objection is untenable. The axiomatisations of Kolmogorov and von Mises both attempt to provide a *rigorous* mathematical foundation for probability theory, but they choose, as we shall see, different sets of primitive terms. In particular, perhaps somewhat surprisingly, the term "probability" does not occur in von Mises' axioms, but is a defined notion, whereas it *is* a primitive term in the Kolmogorov axioms. These different languages reflect different motives, as Kolmogorov was well aware. Von Mises believed that only the frequency interpretation of probability makes sense and attempts to say in mathematical terms what this interpretation amounts to. Kolmogorov's preferences are expressed in the continuation of the passage cited above:

> Die Axiomatisierung der Wahrscheinlichkeitsrechnung kann auf verschiedene Weisen geschehen, und zwar beziehen sich diese verschiedenen Möglichkeiten sowohl auf die Wahl der Axiome als auch auf die der Grundbegriffen und Grundrelationen. Wenn man allerdings das Ziel der möglichen Einfachheit des Axiomensystems und des weiteren Aufbaus der darauf folgenden Theorie im Auge hat, so scheint es am zweckmäßigsten, die Begriffe eines zufälligen Ereignisses und seiner Wahrscheinlichkeit zu axiomatisieren. Es gibt auch andere Begründungssysteme der Wahrscheinlichkeitsrechnung, nämlich solche, bei denen der Wahrscheinlichkeitsbegriff nicht zu den Grundbegriffe zählt, sondern durch andere Begriffe ausgedrückt wird [a footnote refers to von Mises]. Dabei wird jedoch ein anderes Ziel angestrebt, nämlich der größtmögliche Anschluß der mathematischen Theorie an die empirische Entstehung des Wahrscheinlichkeitsbegriffes [44,2].

Although Kolmogorov is clearly aware of the possibility of different axiomatisations of probability theory, this passage has generally been overlooked by von Mises' critics (a phenomenon which will recur again). Accordingly, Kolmogorov was used unwillingly as support for a cause that was not his.

Our attitude toward the problem of axiomatising probability theory is as follows. There is no need to deviate from the Kolmogorov axioms in purely mathematical investigations. But von Mises' theory is a useful (indeed necessary) counterpart to that of Kolmogorov, since it attempts to provide a frequentistic interpretation for the theorems of probability theory which are, strictly speaking, statements about *measure* only. Interestingly, this attempt does not always succeed, as with the law of the iterated logarithm when formulated as a theorem about infinite sequences. Such cases lead one to question the empirical content of some of the results of measure theoretic probability theory. (Ideally, a derivation from the Kolmogorov axioms should be followed by a derivation from von Mises' axioms, to see what the result really means.)
Furthermore, the frequency interpretation is not so crystal clear as to render superfluous attempts at a precise formulation; even a rough formalisation shows that there exist essentially different versions (see 2.2.3). Not least among the merits of von Mises' theory is that it pursues one such interpretation, called *strict frequentism* in 2.2.3, to the bitter end.

2.2.2 Kollektivs (informal exposition). When we look at the list of examples of phenomena to which probability theory should be applicable (see the quotation from von Mises' [70] on p.9): coin tossing, demographic events, Brownian motion, it is clear that these examples exhibit a common trait: either an unlimited repetition of an experiment or a great number of events is involved. But the examples also differ in some probabilistic properties; in modern parlance, we would say that coin tossing is a Bernoulli process, whereas Brownian motion is a Markov process. The essence of von Mises' theory is, that it uses properties of games of chance such as coin tossing as a tool to deduce properties of other processes *and* as an instrument to define probability. In order to have at our disposal a technical term for this privileged case, we introduce the word *Kollektiv*.
Informally, a Kollektiv is a sequence of elements of a sample space (which are also called *attributes*), which is akin to a typical sequence of events produced by coin tossing. To say precisely what "akin to" means, we have to list some of the properties of coin tossing which we regard as essential. Two of these properties, amply verified by experience, are:

(i) Approximate stability of the relative frequency of an attribute if the number of observations (or experiments) is increased;

(ii) The impossibility of a successful gambling strategy, that is, the impossibility of making unlimited amounts of money in a game of chance, using some kind of system. A gambling strategy may roughly be thought of as a rule for betting on some trials and skipping others.

The informal statement of these properties of Kollektivs is sufficient to explain von Mises' version of the frequency interpretation. A more formal statement will be given in 2.3; but, in a sense, all the subsequent chapters are devoted to a formalisation of properties 2.2.2(i) and (ii).

2.2.3 Strict Frequentism: "Erst das Kollektiv, dann die Wahrscheinlichkeit"

We may now give an explicit, albeit informal, definition of probability.

2.2.3.1 Definition The *probability* of an attribute in a Kollektiv equals the relative frequency of that attribute within the Kollektiv.

In 2.3 a more formal definition (involving infinite Kollektivs and limiting relative frequencies) will be given, but the salient points can be illustrated as well using the finite version. Von Mises summarizes his attitude in the slogan: "Erst das Kollektiv, dann die Wahrscheinlichkeit", an innocuous–sounding formula with far reaching implications.

1. There is no probability of an individual event, e.g. that of Rachel dying at age 40, as such. One may, however, *metaphorically* assign various probabilities to this event, corresponding to each Kollektiv to which Rachel belongs: that of female heavy smokers, that of sports car drivers and so forth. So far, only the first property of Kollektivs, 2.2.(i), is used.
2. The second property of Kollektivs gives a special twist to the definition of probability and severely restricts its applicability. Basically, defining the probability of an attribute with respect to a Kollektiv only, means that probability enjoys a multiplicative property. Details will be given in 2.4, but an example will make clear what we mean. The paradigmatic example of a Kollektiv is a sequence of tosses with a fair coin. The probability of heads in such a Kollektiv will (approximately) be $\frac{1}{2}$; and the probability of heads on two consecutive tosses will be $\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$. Now the relative frequency $\frac{1}{2}$ may be called a probability *only* if this multiplicative property holds. In this respect, von Mises' nomenclature differs from that of Kolmogorov, who requires of a probability only that it be a positive measure with norm one. The multiplicative property creeps in only afterwards, when he defines the notion of independence, two events being independent if the probability of their joint occurrence equals the product of the probabilities of the events themselves. He then duly remarks that it is this notion of independence which distinguishes probability theory from measure theory [44,8]. Of course, mass phenomena which do not satisfy the second property of Kollektivs can be

handled as well in the theory, but von Mises' convention is such that in this case, the relative frequency is not a probability.

3. Von Mises' definition is not the only one which establishes some connection between probability and relative frequency. In 2.4, and again in 2.6, we shall meet the *propensity interpretation*, which proceeds along rather different lines. We shall use the term *strict frequentism* for any interpretation of probability which *explicitly* defines probability in terms of relative frequency. Von Mises also thinks that there is more to probability than the definition (2.2.3.1):

> Die Wahrscheinlichkeit, Sechs zu zeigen, ist *eine physikalische Eigenschaft* eines Würfels, von derselben Art, wie sein Gewicht, seine Wärmedurchlässigkeit, seine elektrische Leitfähigkeit usw [70,16].

but this aspect of probability does not figure in the definition.

To appreciate the strictness with which von Mises himself applied his doctrine, it is instructive to consider the case of attributes of probability zero. If computed in a finite Kollektiv, probability zero is of course equivalent to the non-occurrence of that attribute. But when our Kollektiv is infinite, as the precise version of the explicit definition of probability (2.2.3.1) requires, then probability zero of an attribute is compatible with the attribute occurring infinitely often. Although this idea is formulated in terms of infinite Kollektivs, it has consequences for observable events. If $x \in 2^{\omega}$ is a Kollektiv with probability distribution (1,0) and if we derive from x a Kollektiv $y \in (2^n)^{\omega}$ by selection and combination as is done in 2.4, then some of the y_j (which represent finite, observable populations) may contain 1's, although the probability of 1 is zero.

In the case of a continuous sample space, the idea that probability zero does not imply impossibility is universally accepted. But the application of this idea to a discrete sample space seemed too much to swallow, witness the following remark by Martin–Löf, when he contrasts his own approach to the definition of random sequences with that of von Mises:

> [...] an event with vanishing limit frequency is actually impossible. This contrasts sharply with the conception of von Mises, who explicitly stated that the opposite might occur. It seems as if he strained his seldom failing intuition on this point in order not to conflict with his somewhat arbitrary definition of randomness [62,619].

We shall see in 2.5–6 that this divergence of opinions, small as it may seem, actually points to irreconcilable intuitions as regards the principles which should govern the definition of Kollektivs. (And our conclusion will be that Martin-Löf's definition and its relatives are rather more arbitrary than that of von Mises.)

4. Another way to illustrate the strictness of strict frequentism, is to consider the role played by the laws of large numbers (and in fact all weak and strong limit laws of probability theory)

in von Mises' set–up; or rather, the role they do *not* play. We introduce some notation first.

2.2.3.2 Definition Let $p \in [0,1]$. The measure μ_p on 2^ω is defined to be the product measure $(1–p,p)^\omega$. We put

$$LLN(p) := \{x \in 2^\omega \mid \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = p\}.$$

2.2.3.3 Theorem *Strong law of large numbers* : $\mu_p LLN(p) = 1$.

An influential interpretation of probability (influential because apparently unconsciously adopted by most mathematicians), the *propensity interpretation*, holds that probability should primarily be thought of as a physical characteristic. Now von Mises could concede this much (cf. the passage quoted on p.12) but, contra von Mises, the propensity interpretation claims to be able to *derive* the frequency interpretation from the strong law of large numbers together with an auxiliary hypothesis. (Some use the weak law for this purpose; see e.g. the passage from Fréchet [28] cited in 2.6.) In other words, propensity theorists claim that it is possible to derive statements on relative frequencies from premisses which are (almost) probability–free.
We present the alleged derivation of the frequency interpretation in the form given in Popper's *Realism and the Aim of Science* [83]. This presentation might seem anachronistic. But expositions of the propensity interpretation which do show some awareness of its assumptions are rare (the reader may wish to compare Popper's version with that of Fréchet, quoted in 2.6.1). Since the propensity interpretation has inspired some of the work on random sequences in the literature, we have chosen to present it in its (for all its naiveté) most articulated form[1].
The derivation goes as follows.
Suppose we have a coin; after a thorough examination of its physical characteristics (weight, center of mass etc.) we conclude that the probability, *as a physical characteristic or propensity*, of coming up heads will be p. The strong law of large numbers is then invoked to conclude that the set of outcome sequences which show limiting relative frequency of heads equal to p has μ_p–measure one. Now the auxiliary hypothesis comes in. After explaining why the weak law cannot be used in this context, Popper goes on to say:

> The case is different if we obtain a probability that is *exactly* equal to 1 (or 0, as in the case of measure zero). Admittedly, even in this case, "probability" has to mean something connected with frequency if we are to obtain the required result. But no precise connection need be assured – no limit axiom and no randomness axiom [the two conditions formally defining Kollektivs; see 2.3]; for these have been shown to be valid except for cases which have a probability (a measure) zero, and which therefore may be neglected. Thus all we need to assume is that zero probability (or zero measure) means, in the case of random events, *a probability which may be neglected as if it were an impossibility* [84,380].

Stated like this, the argument is quite like the type of reasoning employed in the ergodic foundation of statistical mechanics. Here, one tries to justfy the auxiliary hypothesis on physical grounds:

> [...] one could have an invariant ensemble where every particle moves on the same straight line reflected at each end from a perfectly smooth parallel wall. The obviously exceptional character of this motion is reflected mathematically in the fact that this ensemble, though invariant, is confined to a region of zero "area" on *S* [a surface of constant energy] and therefore has no ensemble density. To set up such a motion would presumably be physically impossible because the slightest inaccuracy would rapidly destroy the perfect alignment (Lebowitz and Penrose [81,24]; for a variation on this argument, see Malament and Zabell [60]).

Von Mises declines any use of the laws of large numbers in the way indicated above. He rightly remarks that this use amounts to an adoption of the frequency interpretation for certain special values of the probabilities, namely those near to 0 and 1 (or equal to 0 or 1 if you use the strong law), and asks: Why not adopt the frequency interpretation from the start, for *all* values of the probability distribution? The obvious answer is that the above procedure explains (or at least pretends to) the frequency interpretation:

> Thus, there is no question of the frequency interpretation being *inadequate.* It has merely become *unnecessary*: we can now derive consequences concerning frequency limits even if we do not assume that probability means a frequency limit; and we thus make it possible to attach to "probability" a wider and vaguer meaning, without threatening the bridge on which we can move from probability statements on the one side to frequency statements which can be subjected to statistical tests on the other (Popper [84, 381]).

In the same way, the ergodic theorem plus the auxiliary hypothesis are taken to explain the statistical behaviour of gases; and we may remark in passing that von Mises also declines such uses of ergodic theory (see the last chapter of [68]).

It is not our purpose here to judge between these two interpretations of probability, strict frequentism and propensity interpretation. We only note that the assumptions underlying the interpretations are of a rather different character:
– The auxiliary hypothesis of the propensity interpretation is of highly *theoretical* nature and badly in need of justification; indeed it is not clear what form a justification should take. In any case it seems more profitable to study concrete examples of its use, for instance in statistical mechanics.
– Von Mises starts from two brute *facts*, amply corroborated by experience, and makes no attempt to explain these facts.
Obviously, in order to turn Popper's *deduction* of the frequency interpretation into a true *explanation,* his premisses have to be analysed further. But since we shall show in the sequel that adherence to the propensity interpretation justifies requirements on the definition of

Kollektivs which are quite unjustified from a strict frequentist point of view, we ask the reader to be alive to both possibilities of interpretation.

2.2.4 Structure and task of probability theory. After all that has been said, it will come as no surprise that the outward appearance of von Mises' theory is rather different from that of Kolmogorov's. We now proceed to give a concise description of its structure; the mathematical details, in so far as they are relevant, will be given in 2.3 and 2.4.

Von Mises emphatically presents probability theory as an empirical theory, designed to transform data, in the form of probabilities, into predictions or explanations, again in the form of probabilities (we omit complications due to the fact that some relative frequencies, e.g. those in Markov processes, are not probabilities. See for these [68] and [70]). The theory should be judged solely on its empirical merits, its adequacy in predicting or explaining observable phenomena.

Since the data are probabilities, they are supplied in the form of relative frequencies in Kollektivs. It follows that probability theory must consist of rules transforming given Kollektivs into other Kollektivs. Accordingly, the axioms of the theory posit the validity of one type of transformations (so called *place selections*); the validity of the other necessary rules of transformation is derivable from these axioms. Consequences of the axioms include the Kolmogorov axioms (albeit with finite additivity only), the multiplicative property alluded to above (in 2.2.3.2) and the formula for conditional probability.

The axioms themselves are a translation into mathematical terms of the facts of experience mentioned in 2.2.2: approximate stability of relative frequencies in long series of trials and the impossibility of a successful gambling strategy. As such, these axioms exhibit a certain amount of idealisation; in particular, the Kollektivs, which in practice are finite, are represented by infinite sequences. This procedure is equally justified as concept formation in geometry: the ideal entities are introduced for their technical advantages, but their properties are studied only on so far as they are relevant to the prediction of observable, hence finite, phenomena. If the infinities can be eliminated, then so much the better.

It would, therefore, be a grave mistake to suppose that von Mises' theory is a *mathematical* theory of *infinite* Kollektivs, as is, for instance, the definition of random sequences proposed by Martin–Löf (for which see Chapter 3). Von Mises introduced infinite Kollektivs only for their technical advantages, not as autonomous objects of study [70,103-4]. We shall discuss Kolmogorov's attempt to define *finite* Kollektivs in 5.2.

2.3 Axiomatising Kollektivs. We now introduce a mathematical description of Kollektivs, essentially by expressing properties 2.2.2(i),(ii) in mathematical terms.

The formal set–up is as follows. Let M (for "Merkmalraum") be a sample space, i.e. the set of possible outcomes of some experiment. The doctrine of strict frequentism says that

probabilities P(A) for $A \subseteq M$ must be interpreted as the relative frequency of A in some Kollektiv. In our mathematical description the probability P(A) will be identified with the limiting relative frequency of the occurrence of A in some infinite Kollektiv $x \in M^{\omega}$.

2.3.1 The axioms (as given by von Mises [67,57]).

2.3.1.1. Axiom A sequence $x \in M^{\omega}$ is called a *Kollektiv* if

(i) for all $A \subseteq M$, $P(A) := \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_A(x_k)$ exists

(ii) Let $A,B \subseteq M$ be non–empty and disjoint; and suppose that $A \cup B$ occurs infinitely often in x. Derive from x a new sequence x', also in M^{ω}, by deleting all terms x_n which do not belong to either A or B. Now let Φ be an *admissible place selection*, i.e. a selection of a subsequence $\Phi x'$ from x' which proceeds as follows:
"Aus der unendliche Folge [x' wird] eine unendliche Teilfolge dadurch ausgewählt, daß über die Indizes der auszuwählenden Elemente ohne Benützung der Merkmalunterschiede verfügt wird."

Then $P'(A) := \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_A(\Phi x')_k$ and $P'(B) := \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_B(\Phi x')_k$ exist and

$\frac{P'(A)}{P'(B)} = \frac{P(A)}{P(B)}$ when $P(B) \neq 0$.

A few remarks on the above definition are in order.

1. The set of axioms 2.3.1.1 will alternatively be called a *definition* of Kollektivs. In Hilbertian jargon, 2.3.1.1 provides an implicit definition of Kollektivs (rather than of probability, as in the Kolmogorov axioms).
2.The quantifier "for all $A \subseteq M$" should not be taken too seriously. In the *Wahrscheinlichkeitsrechnung* [68,17] von Mises remarks that all one needs to assume is that (i) and (ii) hold for "simply definable" sets. For definiteness, we may substitute "Peano–Jordan measurable" for "simply definable".
3. The function P defined in (i) is called *the probability distribution determined by the Kollektiv x*, in conformity with the slogan of 2.2.3. We shall occasionally use the phrase "$x \in M^{\omega}$ is a Kollektiv with respect to distribution P"; this phrase might suggest that the distribution is primary, but should be taken to mean only that P satisfies (i). In the same vein, the phrase "a fair coin" is used to designate a coin whose relative frequencies are approximately equal to ½. It will be clear from the discussion in 2.2.3 that no reference to the physical properties of the coin is intended.

4. We shall use the phrase "$x \in M^\omega$ is invariant under an admissible place selection Φ" to mean that the limiting relative frequency in the subsequence selected by Φ are the same as those in x. The notation " $x \in M^\omega$ " should be read to mean only that each term of x is an element from M; we do not imply that Kollektivs are elements of a universe described by Zermelo–Fraenkel set theory. Similarly, the notation "Φx" for the subsequence selected from x by the admissible place selection Φ should, until further notice (in 2.5) *not* be read as the application of a function $\Phi: M^\omega \to M^\omega$ to x, since at this stage it is not clear that an admissible place selection is indeed a function. The reasons for this caution will gradually become clear in the sequel. (Note also that the notation "Φx" is ambiguous: do we keep track of where the terms of the subsequence originate in x?)

5. Of course the enigmatic condition (ii) will take pride of place among our considerations. In the relevant literature the first part (replacing x by x', obtained from x by deleting terms not in $A \cup B$) is usually omitted. For the paradigmatic case of coin tossing, the sample space M equals $2 = \{0,1\}$ and condition (ii) reduces to:

If Φ is an admissible place selection, $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(\Phi x)_k = P(\{1\})$.

As will be made clear in 2.4, the more elaborate condition is necessary in order to ensure the validity of the rule for conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

It is interesting that the validity of this rule has to be built in blatantly into the axioms (thus emphasizing its empirical origin), especially in view of attempts such as Accardi's [1] to put the blame for the failure of classical probability theory in quantum mechanics upon this rule. (Wald [100, 41-2] claims that, also in the general case, condition (ii) can be reduced as for Kollektivs in 2^ω; but his proof uses evidently non–admissible place selections.)

6. The condition "$P(B) > 0$" is necessary for the ratios in (ii) to be well-defined. On the other hand, it is clear that in von Mises' set-up conditionalisation on a set B is possible if B occurs infinitely often in the Kollektiv, a strictly weaker requirement (cf. the discussion in 2.2.3). One could extend condition (ii) to incorporate B which occur infinitely often, but for which $P(B) = 0$ by means of non-standard analysis: if B occurs infinitely often, then, in any non-standard universe, P(B) is a *positive* infinitesimal, so the ratios in (ii) are well defined.

It will be noted that 2.3.1.1, and especially condition (ii) does not fully conform to present standards of mathematical rigour. In the sequel we shall review a number of attempts to make this condition precise; but let us first try to give an idea of what is meant by means of some examples.

2.3.1.2 Example Admissible place selections may be viewed as gambling strategies: if n is chosen, that means that a bet is placed on the outcome of the n^{th} trial; otherwise, the n^{th} trial is skipped. In the examples we consider the simplest case, cointossing; in other words, Kollektivs x in 2^{ω}.
(a) Choose n if n is prime. (This strategy caused Doob to remark that its only advantage consists in having increasing leisure to think about probability theory in between bets.)
(b) Choose n if the $n–9^{th}$,....., $n–1^{st}$ terms of x are all equal to 1. (The strategy of a gambler who believes in "maturity of chances".)
(c) Now take a second coin, supposed to be independent of the first is so far as that is possible (no strings connecting the two coins, no magnetisation etc.). Choose n if the outcome of the n^{th} toss with the second coin is 1.
Condition (ii) is intuitively satisfied in all three cases, although in (c) a heavy burden is put upon the word "independent". We shall call selections of type (a) and (b) *lawlike* (since they are given by some prescription) and those of type (c) *random*.

Condition (ii) will usually be called the *axiom of randomness* (from *Regellosigkeitsaxiom*). Von Mises alternatively uses the designation *principle of the excluded gambling strategy* (from *Prinzip vom ausgeschlossenen Spielsystem*). Unfortunately, he uses the term "gambling strategy" in two different senses (which he evidently considers to be the same):

> Diese Unmöglichkeit, die Gewinstaussichte beim Spiel durch ein Auswahlsystem zu beeinflussen, die *Unmöglichkeit des Spielsystems*....[70,29]
>
> Daß sie nicht zum gewünschten Ziele führten, nämlich zu einer Verbesserung der Spielchancen, also zu einer Veränderung der relativen Häufigkeiten...[70,30].

Apparently, von Mises thinks that a gambling strategy making unlimited amounts of money can operate only by selecting a subsequence of trials in which relative frequencies are different. It was shown by Ville that this idea is mistaken: there exist gambling strategies (called *Martingales*)which cannot be represented as place selections. We shall come back to this point in 2.6.2 and in 3.4.

Put concisely, the definition of Kollektivs consists of two parts: global regularity (existence of limiting relative frequencies) and local irregularity (invariance under admissible place selections implies that a Kollektiv is unpredictable). Both separately and in conjunction, these parts have come under fire. The most pertinent objections will be reviewed in 2.6, except one, the charge that the theory is outright inconsistent. In view of its urgent character, this charge will be taken up in 2.3.3, after a brief review of some of the consequences of the axioms in 2.3.2.

2.3.2 Some consequences of the axioms The following propositions are literal translations of some of von Mises' *Sätze* in [67].

2.3.2.1 Proposition [67,57] Let $x \in M^{\omega}$ be a Kollektiv, P the probability distribution induced by x. Then P(M) = 1 and P is finitely additive.

This proposition might seem to be trivially true, but in fact its truth value is undetermined until it has been specified on which subsets of M P is defined. Wald [100,46] has shown that for continuous sample spaces M, there exists no non–atomic P defined on all subsets of M; but a finitely additive probability can be defined for all Peano–Jordan measurable subsets of M.

2.3.2.2 Proposition [67,58] An admissibly chosen subsequence of a Kollektiv x is again a Kollektiv, with the same distribution.

Proof Composition of two admissible place selections yields a new selection which still proceeds *ohne Benützung der Merkmalunterschiede*.
This proposition will be the starting point for some of our investigations in Chapters 4 and 5.

2.3.2.3 Proposition [67,59] A Kollektiv x is determined completely by its distribution; it is not possible to specify a function $n \rightarrow x_n$.

Proof Choose $A \subseteq M$ such that $0 < P(A) < 1$. If there were such a function we could use it to define an admissibly selected subsequence of x which consists of elements of A only. □

This consequence contains the essence of the new concept: a Kollektiv has no other regularities than frequency regularities. Von Mises adds the comment that 2.3.2.3 implies

> das man die "Existenz" von Kollektivs nicht durch eine analytische Konstruktion nachweisen kann, so wie man etwa die Existenz stetiger, nirgends differentierbarer Funktionen nachweist. Wir müssen uns mit der abstrakten logischen Existenz begnügen, die allein darin liegt, daß sich mit den definierten Begriffe widerspruchsfrei operieren läßt
> [67,60].

In other words, Kollektivs are *new* mathematical objects, not constructible from previously defined objects. Hence in one place [68,15; see also 70,112] von Mises compares Kollektivs to Brouwer's free choice sequences, one extreme example of which is the sequence of outcomes produced by successive casts of a die[2]. In another place he contrasts his approach with that of Borel [8], in a way which makes clear that Kollektivs are not to be thought of as numbers, i.e. *known* objects:

> ...den von Borel u.a. untersuchten Fragen (z.B. über das Auftreten einzelner Ziffern in den unendlichen Dezimalbrüchen der irrationale Zahlen), wo das Erfülltsein oder Nicht–Erfülltsein der Forderung II [i.e. 2.3.2.1.(ii)] ohne Bedeutung ist [67,65].

The reference is to Borel's Strong Law of Normal Numbers, i.e. Theorem 2.3.2.3 for $p = \frac{1}{2}$ (or rather its analogue for sequences in 10^{ω})! To modern eyes, accustomed to set theory, von Mises' statement may look surprising: (dyadic) numbers and Kollektivs (as they arise in a coin tossing game) can both be thought of as elements of Cantor space. But it will be seen time and again that the set theoretic perspective is not very helpful in understanding von Mises' ideas and the debates to which they gave birth; because at that time, these set theoretic notions were still fresh and not part of the thinking habits of mathematicians. (Neither is set theory very helpful in understanding Borel's ideas on probability; see Novikoff and Barone [79] for a particularly disastrous example of prejudiced historiography. We shall come back to this point in 2.6.)

Digression Perhaps Borel wouldn't have disagreed with von Mises' comment. When he introduces the considerations which lead up to the strong law of normal numbers, he states [8,194–5]

> Nous nous proposons d'étudier la probabilité pour qu'une fraction décimale appartienne à un ensemble donné, en supposant que
> 1 Les chiffres décimaux sont indépendants;
> 2 Chacun d'eux a une probabilité égale a 1/q (dans le cas de la base q) de prendre chacun de ces valeurs possibles: 0,1,2,3,...., q–1.
> Il n'est pas besoin d'insister sur le charactère partiellement arbitraire de ces deux hypothèses; la première, en particulier, est nécessairement inexacte, si l'on considère, *comme on est toujours forcé de le faire dans la pratique*, un nombre décimal défini par *une loi*, quelle que soit d'ailleurs la nature de cette loi. Il peut néanmoins être intéressant d'étudier les conséquences de cette hypothèse, afin que précisement de se rendre compte de la mesure dans laquelle les choses se passent *comme si* cette hypothèse est vérifiée.

In this context it may be interesting to remark that, at the time when Borel proved his strong law (1909), it was by no means considered to be self–evident; in fact one expected the opposite result. Here is Hausdorff's comment [37,420]

> Dieser Satz ist merkwürdig. Auf der einen Seite erscheint er als plausibele Übertragung des "Gesetzes der großen Zahlen" ins Unendliche; andererseits ist doch die Existenz eines Limes für eine Zahlenfolge, noch dazu eine vorgeschriebene Limes, ein sehr spezieller Fall, den man a priori für sehr unwahrscheinlich halten sollte.

And in 1923 Steinhaus still called the strong law of normal numbers *le paradoxe de Borel* [94,286]. Evidently, the strong law was considered to be paradoxical because a regularity such

as the existence of limiting relative frequencies was felt to be incompatible with chance. It is perhaps useful to keep in mind that such was the intellectual climate in which von Mises first published his ideas.

2.3.3 Do Kollektivs exist? Objections to von Mises theory were not long in coming. Although his efforts met with sympathy, doubts were raised concerning the soundness of the foundation. In this respect the following comment is typical:

> Ich glaube nicht, daß Versuche, die von Misessche Theorie rein mathematisch zu fassen, zum Erfolg führen können, und glaube auch nicht daß solche Versuche dieser Theorie zum Nutzen gereichen. Es liegt hier offentsichtlich der sehr interessante Fall vor, daß ein praktisch durchaus sinnvoller Begriff – Auswahl ohne Berücksichtigung der Merkmalunterschiede – prinzipiell jede rein mathematische, auch axiomatische Festlegung ausschließt. Wohl aber wäre es wünschenswert, das sich diesem Sachverhalt, der vielleicht von grundlegender Bedeutung ist, das Interesse weiter mathematischen Kreise zuwendet (Tornier [96,320]).

A catalogue of objections (with their rebuttals) will be given in 2.6, but one simple objection, reiterated ad nauseam, will be dealt with rightaway. The objection states that the appeal to the "abstrakten logischen Existenz" in 2.3.2.3 is illusory, since it is easily shown that Kollektivs with respect to non–trivial distributions do not exist.

For suppose that $x \in 2^{\omega}$ is a Kollektiv which induces a distribution P with $0 < P(\{1\}) < 1$. Consider the set of strictly increasing sequences of (positive) integers. This set can be formed independently of x; but among its elements we find the strictly increasing infinite sequence $\{n \mid x_n = 1\}$, and this sequence defines an admissible place selection which selects the subsequence 11111..... from x. Hence x is not a Kollektiv after all. The above argument, purporting to show the inconsistency of 2.3.1.1 is translated almost literally from Kamke's report to the Deutsche Mathematiker Verein [41,23]. (It may not be entirely out of place to mention that Kamke is the author of a textbook on set theory.) The argument calls for several remarks.

1. It is obviously very insensitive to von Mises' intentions; in fact, it is almost verbally the same as the proof of 2.3.2.3, the proposition which states that a Kollektiv cannot be given by a function! Von Mises had no trouble in dismissing the argument: the set $\{n \mid x_n = 1\}$ does not define an admissible place selection since it uses *Merkmalunterschiede* in a most extreme way. The real problem is rather, to understand why the argument was considered to be convincing at all. It seems that this is one of those cases in which there was no common ground for discussion between von Mises and his adversaries. Kamke speaks as a set theorist: the set of all infinite binary sequences exists "out there", together with all its elements, some of which are Kollektivs. Hence the set $\{n \mid x_n = 1\}$ is available for admissible place selection

in much the same sense as is the set of primes (our example 2.3.1.2(a)).
Von Mises, on the other hand, considers Kollektivs to be *new* objects which, like choice sequences, are not pre–existent; hence $\{n \mid x_n = 1\}$ is *not* available. For him, $n \rightarrow x_n$ is not a legitimate mathematical function; functions are objects which have been constructed. (For evidence of von Mises' constructivist tendencies see, e.g., [71].)

2. Kamke's argument is somewhat beside the mark in that it fails to appreciate the purpose of von Mises' axiomatisation; namely, to provide a mathematical description for certain physical phenomena. The argument refers to what *could* happen, whereas von Mises' axioms are rooted in experience and refer to what *does* happen.
The empirical roots are twofold: in some cases (e.g. in example 2.3.1.2(c), where we use random selection) it is an empirical matter to decide whether a proposed place selection is admissible; and even if we have established to our satisfaction that a place selection is admissible (e.g. on a priori grounds, as for lawlike selections (examples 2.3.1.2(a,b)), the truth of the axiom is by no means self–evident, but at most a fact of experience.
An analogy may be helpful here. In various places (see for instance [70,30]) von Mises likens condition (ii) to the first law of thermodynamics. Both are statements of impossibility: condition (ii) is the principle of the excluded gambling strategy, while the first law (conservation of energy) is equivalent to the impossibility of a perpetuum mobile of the first kind.
It may be even more appropriate to compare condition (ii) to the second law of thermodynamics, the law of increase of entropy or the impossibility of a perpetuum mobile of the second kind, especially in view of Kamke's criticism. Indeed, Kamke's objection is reminiscent of Maxwell's celebrated demon, that "very observant and neat–fingered being", invented to show that entropy decreasing evolutions may occur. Maxwell's argument of course in no way detracts from the validity of the second law, but serves to highlight the fact that statistical mechanics cannot provide an absolute foundation for entropy increase, since it does not talk about what happens *actually*.

3. Another point completely overlooked by Kamke's argument is the *intensional* character of admissible selection, where we use "intensional" in Troelstra's sense:

> Whenever we are led to consider information on sets or sequences beyond their extensions or graphs, we shall speak loosely of "intensional aspects" [98,203].

Clearly, admissibility is not a property of the place selection itself; but, as can be seen from the definition ("Auswahl ohne Benützung der Merkmalunterschiede"), it also involves the consideration of the Kollektiv from which the choice is to be made, or perhaps the process generating that Kollektiv.

Only in the degenerate case where one is tempted to infer the admissibility of a place selection on a priori grounds (e.g. when the selection is lawlike) admissibility may be predicated of the place selection itself, but it must be kept in mind that this is an elliptical way of speaking only. It is not unusual for physical quantities to have an intensional character in the above sense. The notion of a disturbing measurement in quantum mechanics is intensional and likewise admits a degenerate case, namely the measurements which are disturbing because they destroy the system. In this example it is clear that the intensional element, the fact that "disturbing" is not a property of the observable representing the measurement, can be completely explained, using only extensional notions, in a more elaborate theory (via non–commuting operators etc.).

We do not, of course, mean to suggest that these considerations themselves suffice to instill precision in the phrase "Auswahl ohne Benützung der Merkmalunterschiede". But they do serve to show that Kamke has not grasped von Mises' point *and* to direct one's attention to possible formalisations of the enigmatic phrase.

Concluding this part of the discussion and having cleared the theory of the charge of outright inconsistency, we now take a closer look at its metamathematical status. Admittedly, the theory is not formalised, but then, formalisation is not an end in itself. One may expect to derive two benefits from formalisation: the possibility of mechanical checking of proofs, and a proof of consistency.
As can be guessed from the presence of *two* new primitive terms in 2.3.1.1, von Mises' theory is really two in one: *probability theory* , in which it is assumed that some Kollektiv is invariant under certain place selection; and an *explanation of invariance via admissibility*.
The structure of the first part is crystal clear: all notions can be defined in ordinary mathematical terms (even Kollektivs, as follows from results of Wald presented in 2.5) and proofs are just computations (as will be clear from the sample proofs given in 2.4).
Von Mises later came to regard this part as the essential mathematical part of the theory (see, for instance, [68] and [70]; we return to this point in 2.5); the verification that a certain Kollektiv is indeed invariant under a given set of place selections then had to proceed empirically. He considered admissibility to be the *intuitive* explanation of invariance under place selections, but admissibility as such dropped out of the theory [70,29].

The second part of the older theory (explanation of invariance via admissibility) is indeed less clear than the first part; but this does not mean that the notion of admissibility is completely unclear or even inconsistent. In particular, the notion is clear enough to show the validity of arguments like the proof of 2.3.2.2, which is of the form:

If Φ is an admissible place selection on x, then Ψ is an admissible selection

on y.

The same type of argument occurs in 2.4, when it is shown that the Kollektivs are closed under certain operations (admissible place selections being a special case).

An axiomatisation of admissibility could proceed by *postulating* the validity of 2.3.2.2 and related propositions in 2.4, with an additional postulate which says that lawlike place selections are admissible. This is more or less the approach chosen by Dörge [22] and amounts to an implicit definition of admissibility.

We believe that the second part of the theory has enough physical plausibility to make further attempts at formalisation worthwhile. In 5.6 we present two different explicit definitions of admissibility, involving Kamae entropy and Kolmogorov complexity; we do not claim that these definitions exhaust the possible meanings of admissibility. Rather, these definitions should be viewed as different projections of the universe where Kollektivs "live", the formalisation of which still has to be found.

2.4 The use of Kollektivs In the previous section we examined the meaning of proposition 2.3.2.3 from the point of view of the foundations of mathematics.We saw that it laid the theory open to the (albeit unjustified) charge of inconsistency. Now we investigate its probabilistic meaning. On the face of it, proposition 2.3.2.3 seems to make von Mises' theory pointless: on the one hand a Kollektiv is completely determined by its distribution (in the sense that nothing more can be said about it), on the other hand, Kollektivs are deemed to be necessary for the interpretation of probability. Then a natural question arises: Why do we need Kollektivs at all? Why isn't it sufficient to use the distribution (as in effect happens in Kolmogorov's theory) instead of the unwieldy formalism of Kollektivs?

In what sense, then, do Kollektivs occur in computations, over and above their distribution? The answer, as we shall see, is that anybody who believes in the frequency interpretation and in the validity of the usual rules for probability, is bound to believe in Kollektivs. That is, not necessarily in the idealized, infinite Kollektivs as they occur in von Mises' axioms, but rather as finite approximations to these. In other words, *Kollektivs are a necessary consequence of the frequency interpretation.* This point is made by von Mises, when he states that

> Die Autoren, die die allgemeine Regellosigkeit "ablehnen" und durch eine beschränkte ersetzen, schließen entweder alle Fragen der Beantwortung aus, die nicht der von ihnen willkürlich gesetzten Beschränkung entsprechen; oder sie nehmen in jedem konkreten Fall die Regellosigkeit, die gerade gebraucht wird, als ein Datum der betreffenden Aufgabe an, was nur auf eine Änderung der Darstellungsform hinauslauft [70,128-9].

One of the main goals of this section is to establish the claim that Kollektivs are necessary for the frequency interpretation of probability (otherwise the reader might think that, von Mises'

theory being superseded by Kolmogorov's, there is no use anymore in investigating Kollektivs). This will be done in 2.4.2. To do so, we need some facts concerning operations on Kollektivs, which will be presented in 2.4.1. There, we also have the opportunity to stress the differences in the treatment of *independence* in the theories of von Mises and Kolmogorov. In 2.4.3, we consider the role of the laws of large numbers in von Mises' theory, a subject already touched upon in 2.2.3.

2.4.1 The fundamental operations: definition and application.

2.4.1.1 Definition of the operations We indicate briefly how the usual rules of probability theory can be derived using 4 operations, which transform Kollektivs into Kollektivs. That is, we shall prove, using our intuitive understanding of admissibility, that these operations preserve *Kollektiv*hood. These proofs can be made fully rigorous if we start with a given set of place selections, in the spirit of von Mises' later ideas; alternatively, we may use the four operations to axiomatise admissibility.

1. *Place selection* This operation transforms a Kollektiv into a Kollektiv with respect to the same distribution; indeed, this is the content of proposition 2.3.2.2.
2. *Mixture* Let $x \in M^\omega$ be a Kollektiv with respect to a distribution P on M. Let N be a sample space and $f: M \to N$ a function (which, of course, must in some sense be constructive). Consider the sequence $y = (f(x_n))_n$ in N^ω. Obviously y induces the distribution Pf^{-1}. Moreover, y is a Kollektiv with respect to this distribution: since f is defined by a mathematical law, an admissible place selection operating on y can be transformed, using f, to an admissible place selection on x.
3. *Division* Let A be proper subset of M, $x \in M^\omega$ a Kollektiv with respect to P, and suppose that A occurs infintely often in x. Division allows one to define the conditional probability P(B|A) for $B \subseteq M$: we transform x into a sequence $x' \in A^\omega$ by retaining only those terms of x which belong to A. If we also suppose that $P(A) > 0$, then we may define (for $B \subseteq A$) $P(B|A) := P(B)/P(A)$ and x' is a Kollektiv with respect to $P(\bullet|A)$. In fact, the whole point of the elaborate condition of randomness 2.3.1.1 (ii) is just to ensure that x' *is* Kollektiv (a point missed by Schnorr [88,18]). If A occurs infinitely often in x, but nonetheless $P(A) = 0$, we may use non-standard analysis as indicated in 2.3.1. If *P denotes the extension of P to the non-standard universe and st(•) the standard part map, the distribution in x' is given by $P(B|A) = st(*P(B)/*P(A))$. (Related ideas can be found in [104].)
4. *Combination* Let M,N be sample spaces, $x \in M^\omega$ a Kollektiv with respect to P, $y \in M^\omega$ a Kollektiv with respect to Q. Combining Kollektivs is the operation of forming the sequence $(<x_n,y_n>)_n$ in $(M\times N)^\omega$. We then need to know conditions under which this sequence is again a Kollektiv and if so, with respect to which distribution. If we analyse the meaning of applying

an admissible place selection to a sequence $(<x_n,y_n>)_n$, we arrive at the following necessary and sufficient condition for this sequence to be a Kollektiv:

Independence Let x, y be as above. $(<x_n,y_n>)_n$ is a Kollektiv with respect to the distribution P×Q on M×N if x and y are *independent*[3] *Kollektivs*, i.e. if the following operation leads to a Kollektiv x" in M^ω with distribution P:
Fix arbitrary $A \subseteq N$. Apply an admissible place selection to y, giving a subsequence $(n_k)_k$ of natural numbers and a sequence y' such that y'_k equals the n_k^{th} term of y. Then select a subsequence x' from x as follows: the n_k^{th} term of x is retained if $y'_k \in A$; and, lastly, apply an admissible place selection to x', giving x". (It is not difficult to check that the relation of independence is symmetric. The last condition is necessary in order to ensure that x and y are themselves Kollektivs.)

Similarly, one may define independence of three Kollektivs: we say that x,y and z are independent Kollektivs if they are pairwise independent (in the above sense) and if each of them is independent (again in the sense introduced above) of the combination of the other two. The extension to n independent Kollektivs is routine.
In [70,58], von Mises calls the operation of selecting a subsequence x' from x as follows: the n_k^{th} term of x is retained if $y'_k \in A$, *sampling*. We have met sampling already in example 2.3.1.2(c), as a special case of admissible place selection: if $x,y \in 2^\omega$ are Kollektivs supposedly generated by independent coins, choose those x_n for which $y_n = 1$. But note that in the above condition, sampling is used to *define* what it means for two Kollektivs to be independent.
The particular type of sampling displayed in example 2.3.1.2(c) will occur so often, that is denoted by a special symbol: for $x \in 2^\omega$, $y \in 2^\omega$, where y contains infinitely many ones, $^x/_y$ is defined as:

$$(^x/_y)_m = x_n \text{ if n is the index of the } m^{th} \text{ 1 in y.}$$

(This notation is slightly ambiguous; do we keep track of which n were chosen or not? We shall never need to.)

We now illustrate the condition of independence with two examples, one pertaining to two tosses with a single coin, the other to two coins, supposed to be physically independent. Whereas in Kolmogorov's theory these two cases are treated alike by *postulating* that probabilities multiply, in von Mises' theory the two cases are distinguishable in that in the first case independence, hence the product rule, is provable, while in the second case independence has to be assumed.

2.4.1.2 Examples

1. We are interested in the probability of obtaining two times heads with two tosses in succession of a fair coin. Let x be a Kollektiv with respect to distribution $(\frac{1}{2},\frac{1}{2})$. A new Kollektiv, representing the situation in which we are interested, is obtained as follows:
choose first those x_n for which n is odd, then those x_n for which n is even; then combine the two Kollektivs thus obtained, which gives $\xi = (<x_{2n-1},x_{2n}>)_{n\geq 1}$.
In this case it is *provable* that ξ is a Kollektiv with respect to the product distribution on $\{<0,0>,<0,1>,<1,0>,<1,1>\}$; in other words, it is provable that (x_{2n-1}), (x_{2n}) are independent Kollektivs. To calculate the distribution in ξ (e.g. the probability of $<1,1>$), we may proceed as follows: single out those odd n for which $x_n = 1$; this operation gives us a sequence $f: \omega \to \omega$ such that $x_{2f(k)-1} = 1$. For this particular f, consider $(x_{2f(k)})_k$. This sequence can be thought of as being chosen from x by the following admissible place selection: x_n is chosen if n is even and $x_{n-1} = 1$. Hence this sequence is a Kollektiv with distribution $(\frac{1}{2},\frac{1}{2})$. The computation is now a matter of bookkeeping:

if we put $y = (x_{2n-1})$, z is (x_{2n}), $Y(m) = \sum_{n=1}^{m} y_n$, then $(x_{2f(k)}) = z/y$ and we may write

$$\frac{1}{m}\sum_{n=1}^{m} 1_{<1,1>}(<x_n,y_n>) = \frac{1}{m}Y(m)\cdot\frac{1}{Y(m)}\sum_{k=1}^{Y(m)}(z/y)_k;$$

and the desired value $\frac{1}{4}$ is obtained by taking limits.
In the same way one proves that ξ is a Kollektiv. Let Φ be an admissible place selection operating on ξ. Φ determines an admissibly chosen subsequence $(x_{2g(i)-1})_i$, for some sequence $g: \omega \to \omega$; and also an admissibly chosen subsequence of (x_{2n}), the latter determined by the procedure: choose those n such that $n = 2g(i)$ and $x_{2g(i)-1} = 1$. The computation now proceeds as above, with the sequences just defined replacing (x_{2n-1}) and $(x_{2f(k)})$.

We thus see that in von Mises' theory the product rule is part of the *meaning* of probability; it is provable from the properties of Kollektivs that the probability of the outcome of two tosses in succession is obtained by multiplying the probabilities of the single outcomes. The same holds for the probabilities for the outcomes of n tosses in succession.
A single Kollektiv x in 2^ω thus induces a product probability distribution on the binary words of length n, for each n. This example therefore illustrates the claim made earlier, that place selections are intended to capture the independence of successive tosses.

2. Now consider two tosses with two fair coins, supposed to be independent (in some physical sense). In this case we also expect the product rule to hold. But now its validity must be assumed; there is no way to deduce it from the theory. To be specific, if x and y are Kollektivs representing the two coins, we must assume that x and y are independent in the sense of the

condition given in 2.4.1.1.4. Once this assumption is made, a simple computation, exactly as in the previous example, shows that the probabilities of the outcomes <0,0>, <0,1>, <1,0> and <1,1> are given by the productrule:

if we put $Y(m) = \sum_{n=1}^{m} y_n$, then $\frac{1}{m}\sum_{n=1}^{m} 1_{<1,1>}(<x_n, y_n>) = \frac{1}{m}Y(m)\cdot\frac{1}{Y(m)}\sum_{k=1}^{Y(m)}(x/y)_k$;
since y is a Kollektiv with respect to $(\frac{1}{2},\frac{1}{2})$, $\lim_{n\to\infty}\frac{1}{m}Y(m) = \frac{1}{2}$, so after taking limits
the left hand side equals $\frac{1}{4}$.

The case of n independent coins (possibly with different distributions) is treated similarly; but note that we now need n independent Kollektivs to induce a product probability distribution on the binary words of length n, whereas in the previous example one Kollektiv sufficed for all binary words.

2.4.1.3 Comparison At this point, having seen some of the differences between von Mises' theory and that of Kolmogorov, the reader may well wonder how the results of the two theories are related. The answer is somewhat intricate. Recall the different definitions of probability in the two theories: Kolmogorov provides an implicit definition of probability as a positive measure with norm one, while in von Mises' theory, probability is basically a measure *together with* a Kollektiv which induces that measure.

It is clear from this description that not necessarily every theorem of the form "the probability of such-and-such is so-and-so" derived from the Kolmogorov axioms is derivable in von Mises' theory, *for the latter's interpretation of probability*.
In fact, we shall see in 2.6 that the law of the iterated logarithm, *when stated in this form*, is a counterexample. Roughly speaking, we may say that von Mises' theory can reproduce that part of Kolmogorov's theory (with the probability distribution interpreted in a Kollektiv), which makes no essential use of the σ-additivity of the measure. The first volume of Feller' treatise [25] gives a fair sample of problems which fall in this category, as do, of course, von Mises' own technical works on probability theory, *Wahrscheinlichkeitsrechnung* [68] and *Mathematical theory of probability and statistics* [74] (this is not to say that the books mentioned contain all that can be derived in von Mises' theory).
That part of Kolmogorov's theory which *does* use σ-additivity essentially, can be derived in von Mises' theory purely conventionally, as a statement concerning *measure*, which in some cases, but not in all, can also be interpreted as a statement concerning *probability*. The strong limit laws belong to this category, when stated in their usual form:

"The measure of the following set of infinite sequences {..|.....} is 1".

Nevertheless, as readers of Feller's [25] well know, the strong limit laws can also be stated in terms of finite sequences and in that form they *are* derivable in von Mises' theory. This holds

for the finite version of the strong law of large numbers, briefly considered in 2.4.3, as well as for the finite version of law of the iterated logarithm (for which see Kolmogorov [45] and 2.4.3).

2.4.2 Necessity of Kollektivs A natural question suggested by the existence of these two different formalisms for probability theory is: Which formalism is to be preferred? The course of history has already provided some sort of an answer: no one uses von Mises' formalism anymore. Apparently, we must conclude from this fact that Kollektivs have no relevance for probability theory as such. They may perhaps be studied for their own sake, in some far-out corner of mathematics; but, to use Poincaré's famous distinction, as a problem that one poses, not as a problem that poses itself. This conclusion, however, is mistaken.
Indeed, it is quite trivial to show that anyone who interprets probability as relative frequency and accepts the Kolmogorov axioms plus the product rule for (physically) independent events, also has to believe in Kollektivs. (If the sample space has cardinality greater than 2, the rule for conditional probabilities must be added to this list.)
In practice, we have to operate with relative frequencies in finite sequences, so strictly speaking one can't deduce the existence of infinite Kollektivs. However, for simplicity we shall assume that probability is interpreted as limiting relative frequency, in which case the existence of infinite Kollektivs *can* be deduced. With suitable approximations the argument works as well for finite sequences. (In fact, Kolmogorov's later conviction that his axioms needed to be supplemented by a precise form of the frequency interpretation, led him to the first satisfactory definition of randomness for *finite* sequences; see 5.2)

We shall now give the argument, which consists essentially only in inverting the examples in 2.4.1.2.
Referring to the first example, we claim the following. Consider an infinite sequence of tosses with a fair coin; if the probability of heads is identified with its limiting relative frequency in the sequence (in this case $\frac{1}{2}$), and if this probability satisfies the usual rules plus the product rule for two consecutive tosses, then the sequence must be invariant under the place selections which occur in the proof of the product rule.

To prove the claim, recall that three place selections occuured in example 1: if x denotes a sequences of tosses with a fair coin, we select from x
(i) x_m with m odd
(ii) x_m with m even
(iii) x_m with m even and $x_{m-1} = 1$.
We show that in each of the selected subsequences, the limiting relative frequency of 1 is $\frac{1}{2}$. We assume the frequency interpretation and the product rule: the probability of each of the

outcomes <0,0>, <0,1>, <1,0>, <1,1> in $(<x_n,y_n>)_n$ is ¼. The computation goes as follows.

(i) $$\lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}x_{2n-1} = \lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}1_{<1,0>}(<x_{2n-1},x_{2n}>) + \lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}1_{<1,1>}(<x_{2n-1},x_{2n}>)$$

$= ¼ + ¼ = ½$.

(ii) is treated analogously.

(iii) If we put $y = (x_{2n-1})$, $z = (x_{2n})$, $Y(m) = \sum_{n=1}^{m} y_n$, then the selected subsequence can be written z/y and we have

$$\lim_{n\to\infty}\frac{1}{Y(m)}\sum_{k=1}^{m}(z/y)_k = \frac{\lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}1_{<1,1>}(<z_n,y_n>)}{\lim_{m\to\infty}\frac{1}{m}Y(m)};$$

by (i), $\lim_{n\to\infty}\frac{1}{m}Y(m) = ½$, so the right hand side equals $\frac{¼}{½} = ½$.

The same trivial argument can be applied to the second example, to show that the sequence x of outcomes of the tosses with the first coin must be invariant under a place selection defined by the second coin: choose those x_n for which $y_n = 1$ (for example).

Summarizing: interpreting probability as limiting relative frequency and applying the deductions of probability theory to a sequence x entails assuming that x is a Kollektiv, or at least that it has the Kollektiv-properties required for the particular deduction at hand (and one is tempted to argue: since we could have chosen to perform a different calculation, e.g. that of the probability of n times heads on n consecutive tosses, x must in fact be a Kollektiv, invariant under all admissible place selections).
Part of probability theory is adequately represented by Kolmogorov's axioms, but as soon as it comes to interpreting the results (as results on relative frequency), one necessarily has to consider Kollektivs. And to say precsiely what the frequency interpretation is, one has to give a precise definition of Kollektivs.

2.4.3 Strong limit laws Twice already, strong limit laws were mentioned in connection with von Mises' theory and both times we stressed a negative aspect. In 2.2.3 it was said that the existence of limiting relative frequencies in a Kollektiv cannot be inferred from the strong law of large numbers (which states that these limiting relative frequencies exist in "almost all" sequences). Rather, they were assumed to exist because that is a reasonable idealisation of experience. In 2.4.1.3 we remarked that the law of the iterated logarithm, when stated in its

usual form (that is, for infinite sequences), is not derivable in von Mises' system. Given the central role of the strong limit laws in probability theory, it is natural to inquire into their status in von Mises' theory.

Von Mises devoted a chapter of *Wahrscheinlichkeit, Statistik und Wahrheit* [70,129-163] to this problem; and elsewhere in this book, in a description of the contents of "das schöne und sehr lesenswehrte Büchlein von A. Kolmogoroff" [70,124], the *Grundbegriffe*, he indicated in what sense the law of the iterated logarithm is derivable in his system ([70,125]; a passage which has apparently gone unnoticed).

We shall follow von Mises' description of the strong law of large numbers; after that, little need be added to clarify the status of the law of the iterated logarithm.

Let x be a Kollektiv in 2^ω with respect to distribution (1–p,p). Fix $n,m \in \omega$ with $m < n$ and let $\varepsilon \in (0,1)$. From x a Kollektiv y in $(2^n)^\omega$ is derived as in example 1 of 2.4.1.2: y is a *combination* (in the sense of the fourth operation discussed in 2.4.1.1) of the n Kollektivs $(x_{kn+i})_k$ for $1 \le i \le n$. As in the example, one shows that y is a Kollektiv with respect to the product distribution on the binary words of length n. From y we derive by *mixing* a Kollektiv z in 2^ω as follows (recall that each y_j is an n–tuple):

$$z_j = \begin{cases} 1 & \text{if for some } k,\ m \le k \le n,\ \left|\frac{1}{k}\sum_{i=1}^{k}(y_j)_i - p\right| > \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

As von Mises presents it, the strong law of large numbers then says that the limiting frequency of 1 in z, i.e. of the event:

$$\exists k(m \le k \le n \ \&\ \left|\frac{1}{k}\sum_{i=1}^{k}(y_j)_i - p\right| > \varepsilon) \text{ in } y,$$

is less than $\varepsilon^{-2}\cdot m^{-1}$, independent of the values of n and p. What is the relation of this form of the strong law of large numbers to the form stated as Theorem 2.3.3?

Put

$$A_{mn}(\varepsilon) := \{w \in 2^n \mid \forall k\,(m \le k \le n \ \rightarrow \left|\frac{1}{k}\sum_{i=1}^{k} w_i - p\right| \le \varepsilon)\}.$$

Let P_n be the probability distribution on 2^n induced by x (via y). Von Mises' version of the strong law then implies:

$$\forall \varepsilon > 0\ \forall \delta > 0\ \exists m\ \forall n \ge m\ P_n(A_{mn}(\varepsilon)) > 1-\delta.$$

Now P_n may *formally* be regarded as the restriction of the measure $\mu_p = (1–p,p)^\omega$ on 2^ω to 2^n. We may then write equivalently (*we just use a different notation*):

$$\forall\varepsilon > 0\forall\delta > 0\exists m\forall n \geq m\ \mu_p\{x \in 2^\omega | \forall k\ (m \leq k \leq n \rightarrow \left|\frac{1}{k}\sum_{i=1}^{k} x_i - p\right| \leq \varepsilon)\} > 1-\delta.$$

This statement is, *using the* σ*-additivity of* μ_p, equivalent to

$$\mu_p\{x \in 2^\omega | \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = p\} = 1.$$

In other words, the usual version of the strong law can be derived from the version acceptable to von Mises if we take the collection of *probability distributions* (P_n), induced by the Kollektiv $x \in 2^\omega$, to define a single σ-additive *measure* μ_p on 2^ω. From the standpoint of von Mises, however, the extension of the collection (P_n) to μ_p is a purely conventional matter, bereft of probabilistic significance.

It is perhaps not superfluous to recall that Kolmogorov was of the same opinion; in fact, von Mises credits his presentation of the strong law to Kolmogorov [45], the paper which contains the general form of the law of the iterated logarithm for independent random variables. In this article, Kolmogorov emphatically states that the only meaningful form of the law pertains to *finite* sequences. Since we do not at present need the result in full generality, we state it for i.i.d. two-valued random variables. Modulo this simplification, Kolmogorov's version reads as follows:

(a) $\forall\varepsilon > 0\forall\delta > 0\exists m\forall n \geq m\ \mu_p\{x \in 2^\omega | \exists k\colon m \leq k \leq n\ \&\ \sum_{j=1}^{k} x_j > (1+\delta)\sqrt{2k \cdot p(1-p)\log\log n}\ \} < \varepsilon$

(b) $\forall\varepsilon > 0\forall\delta > 0\exists m\forall n \geq m\ \mu_p\{x \in 2^\omega | \forall k\colon m \leq k \leq n\ \&\ \sum_{j=1}^{k} x_j < (1-\delta)\sqrt{2k \cdot p(1-p)\log\log n}\ \} < \varepsilon;$

with analogous conditions for the lower bound on the relative frequency. (For notational convenience we have used μ_p instead of the P_n; but it will be clear that we refer in fact to finite sequences only.)

As with the strong law of large numbers, *this* form is derivable from von Mises' axioms, the extension to the version for infinite sequences then being purely conventional. Ville's theorem, discussed in 2.6 and improved upon in 4.6, will in fact show that there is no straightforward frequency interpretation for the infinite version.

In conclusion we emphasize again that, for von Mises, the limiting relative frequencies in a Kollektiv do not owe their existence to the strong law of large numbers. Rather, it is the other way around, as the above derivation should have made clear: only because our x satisfies the two conditions on Kollektivs, it allows us to deduce the strong law, as a statement on the

relative frequency of a particular event.

2.5 Making Kollektivs respectable: 1919 – 1940 For a while, from 1919 to 1933, the only explicit, more or less rigorous, axiomatisation of probability theory (von Mises') made use of Kollektivs, hence the imperative need to make this objects mathematically acceptable. Two principal lines of attack can be distinguished.

1. Restricting *a priori* the class of admissible place selections and trying to construct explicitly a Kollektiv with respect to the class so obtained (Reichenbach, Popper, Copeland; 2.5.1);
2. Showing that von Mises' theory is consistent *in context*, that is, showing that in each specific application we may assume the existence of a Kollektiv with respect to the place selections required for the application (von Mises, Wald; 2.5.2).

After the appearance of Kolmogorov's *Grundbegriffe* in 1933, and especially after the Geneva conference in 1937, at which strict frequentists and the proponents of an implicit definition of probability came into head-on collision (see 2.6), attempts to define Kollektivs petered out, with Church's [16] (1940) as a notable exception. Only in 1963, with the publication of Kolmogorov's [47], hostilities were resumed. We now discuss attempts 1. and 2.; for simplicity, we consider Kollektivs in 2^{ω} only.

2.5.1 Lawlike selections Common to all attempts which fall under the heading 1. above, is the conviction that "admissible place selection" should mean "place selection given by a mathematical law", as in the first two examples illustrating the definition of admissible place selection (choose the n^{th} term if n is prime; choose the n^{th} term if it is preceded by 10 1's). We comment on this interpretation later, but let us first consider some representative examples of this approach.
Various authors (e.g. Popper [83], Reichenbach [85], Copeland [17]) independently arrived at a class of place selections which is a generalisation of the second example (2.3.1.2(b)): the so-called *Bernoulli selections*. They can be described as follows: let x be a Kollektiv; fix a binary word w and choose all x_n such that w is a final segment (or *suffix*) of x(n–1).
Note that this selection chooses an infinite subsequence of x if x contains infinitely many occurrences of w (which is for instance the case if x is a Kollektiv with respect to (1–p,p), for $0 < p < 1$).
We henceforth treat place selections as partial functions $\Phi: 2^{\omega} \to 2^{\omega}$, where Φx is the infinite subsequence selected from x by Φ. This identification is not unproblematic. It has the technical disadvantage that it does not keep track of where the n^{th} selected term occurred in the original sequence. Its main philosophical disadvantage is, that it is most appropriate for place selections which are judged admissible on a priori grounds. It is considerably less so for

place selections which are admissible for a given Kollektiv, the general case of admissibility (cf. the discussion in 2.3.3). Since we are concerned in this section with place selections which are, for various reasons, judged admissible on a priori grounds, the identification is harmless here.

The *domain* of a place selection Φ will be the set of those x such that Φ operating on x produces an infinite subsequence of x. Intuitively, a place selection Φ is completely determined by a function $\phi: 2^{<\omega} \to \{0,1\}$, when we interpret the statement "$\phi(w) = 1$" as: choose the $|w|+1^{th}$ term, and "$\phi(w) = 0$" as: skip the $|w|+1^{th}$ term. To bridge the gap between ϕ and Φ it is convenient to use a place selection Φ' which operates on finite sequences. We formalize these remarks in the following definition; we first introduce a general definition of place selection, and then specialize to Bernoulli selections, as introduced informally above.

2.5.1.1 Definition Let $\phi: 2^{<\omega} \to \{0,1\}$ be any function. ϕ determines a place selection Φ in two steps:

(i) $\Phi': 2^{<\omega} \to 2^{<\omega}$ is given by $\Phi'(uj) = \begin{cases} \Phi'(u)j & \text{if } \phi(u) = 1 \\ \Phi'(u) & \text{if } \phi(u) = 0 \end{cases}$ where $j \in \{0,1\}$

(ii) a partial function $\Phi: 2^{\omega} \to 2^{\omega}$ is defined by

(a) $\text{dom}\,\Phi = \{x \in 2^{\omega} \mid \forall n \exists k \geq n\ \phi(x(k)) = 1\}$

(b) $x \in \text{dom}\,\Phi$ implies $\Phi(x) = \bigcap_n [\Phi'(x(n))]$

2.5.1.2 Definition Let $w \in 2^{<\omega}$ and $\phi_w: 2^{<\omega} \to \{0,1\}$ defined by

$$\phi_w(u) = \begin{cases} 1 & \text{if w is a final segment of u} \\ 0 & \text{otherwise} \end{cases}$$

$\Phi_w: 2^{\omega} \to 2^{\omega}$ is a *Bernoulli selection* if it results from ϕ_w by application of (i) and (ii) of 2.5.1.1.

Recall that for $p \in [0,1]$, the set LLN(p) was defined as (2.2.3.2):

$$\text{LLN}(p) = \{x \in 2^{\omega} \mid \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k = p\}.$$

2.5.1.3 Definition Let $p \in [0,1]$. $x \in 2^{\omega}$ is called a *Bernoulli sequence* with parameter p (notation: $x \in B(p)$) if for all w: $x \in \text{dom}\,\Phi_w$ implies $\Phi_w(x) \in \text{LLN}(p)$.

It is not difficult to show that, if x is a Bernoulli sequence with parameter p, for each word w the limiting relative frequency of w in x equals $\mu_p[w]$.

2.5.1.4 Lemma Let $p \in [0,1]$. Then

$$x \in B(p) \text{ iff } \forall w \in 2^{<\omega}: \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{[w]}(T^k x) = \mu_p[w],$$

where $T: 2^\omega \to 2^\omega$ is the left shift and $\mu_p = (1-p,p)^\omega$.

Proof See, e.g., Schnorr [88,22]. □

Remark The preceding lemma has as a consequence that, at least for $1 > p > 0$, $x \in B(p)$ implies for all words w: $x \in \text{dom } \Phi_w$. The "implies" in definition 2.5.1.3 might therefore have been replaced by "and".

In the special case $p = \frac{1}{2}$, Bernoulli sequences are commonly called, *normal numbers*. Now, although Kollektivs were not supposed to be constructible (cf. proposition 2.3.2.3), Bernoulli sequences can be constructed explicitly. E.g.

2.5.1.5 Lemma There exists a recursive normal number.

Proof (Champernowne [15]) Let x = 0100011011000....., i.e. the set of all finite binary words written in lexicographic order. For the construction of Bernoulli sequences for arbitrary p, see von Mises [69]. □

In one sense, normal numbers and, more generally, Bernoulli sequences, are clearly not satisfactory models of Kollektivs, if only because problems involving two coins (say), cannot be treated in the way von Mises intended[4]. On the other hand, the beautiful work of Kamae [40], which is described in 5.6, shows that there are really many more place selections Φ such that $x \in B(p)$ implies $\Phi x \in B(p)$; in fact an uncountable set and what's more, with an appealing physical description.

Bernoulli selections are examples of lawlike selections, but by no means the only ones; e.g. our second example: choose x_n if n is prime, is not of this form. The apparently most general characterisation of lawlike place selections is due to Church [16] (the article is from 1940, a time when von Mises' theory was no longer a hot issue).

2.5.1.6 Definition A function Φ: $2^{\omega} \to 2^{\omega}$ is called a *recursive place selection* if it is generated by a total recursive ϕ: $2^{<\omega} \to \{0,1\}$ according to (i) and (ii) of 2.5.1.1.

2.5.1.7 Definition Let $p \in [0,1]$. x is *Church random* with parameter p (notation: $x \in C(p)$) if for all recursive place selections Φ: $x \in \text{dom } \Phi$ implies $\Phi x \in LLN(p)$.

Remark Unlike the situation for Bernoulli sequences, in this case the "implies" cannot be replaced by "and". In other words, while

$$B(p) \subseteq \bigcap_w \text{dom } \Phi_w$$

we do not have

$$C(p) \subseteq \bigcap \{\text{dom } \Phi | \ \Phi \text{ recursive}\}.$$

In 2.6.2 we shall meet an example of a place selection Φ such that $C(p) \not\subset \text{dom } \Phi$. This observation implies that, with the above definition of Church randomness, some Bernoulli sequences are Church random for fairly trivial reasons. Note that, from the point of view of von Mises' theory, it would be natural to require that a Kollektiv belongs to the domain of the place selections needed to solve a particular problem, since the theory consists essentially of transformations of (infinite) Kollektivs into (infinite) Kollektivs. Also, the wording of the definition of Kollektivs (2.3.1.1; originally [67,57]) suggests that it is *assumed* that admissible place selections select infinite subsequences. However, it is customary in the literature to use the implication in 2.5.1.7 (see e.g. Schnorr [88,22]) and for good reason, since there exist (recursive!) place selections with disjoint domains.

We now discuss the merits of the identification of "admissible place selection" with "lawlike place selection".

1. It is an illusion to suppose that one can restrict oneself to the existence of lawlike place selections only. As the paragraph on combination in 2.4.1.1 shows, a lawlike selection on $(<x_n,y_n>)$ factors as a lawlike selection on y and a *random* selection on x. Hence, by the argument given in 2.4.2, it follows that an application of the theory, even to such a simple problem as that of the probability of two coins coming up heads, assumes that x and y satisfy stronger properties of randomness than just being Church random. And if it is maintained that the admissibility of lawlike place selections can be recognized a priori, this has a consequence that the admissibility of the above random selection on y is also an priori; a consequence which should perhaps instill some caution in the use of the a priori in this context.

2. The recursive analogue of proposition 2.3.2.2:

An admissibly chosen subsequence of a Kollektiv is again a Kollektiv, with the same distribution,

is

If x ∈ C(p), *then for every recursive place selection* Φ, x ∈ dom Φ *implies* Φx ∈ C(p).

If the admissible place selections were identified with the recursive place selections, C(p) would be the set of Kollektivs with distribution C(p); so if x ∈ C(p), we have by the above analogue of 2.3.2.2 at least countably many subsequences of x which are also Kollektivs with respect to (1–p,p). Now it seems very implausible that, for a satisfactory definition of Kollektivs, *only* countably many subsequences of a Kollektiv are themselves a Kollektiv (with the same distribution).

On the contrary, we shall prove the following *principle of homogeneity*, which can be read as a quantitative version of proposition 2.3.2.2:

If x *is a Kollektiv with respect to* (1–p,p), *so is almost every subsequence of* x.

To turn this rather vague principle into a precise mathematical statement requires some effort; this will be done in Chapters 3 and 4 and involves, perhaps somewhat surprisingly, a study of modern definitions of randomness. But to give the reader already at this stage an impression of the formal version of the principle, we state it in semi-formal terms (where / denotes the operation of *sampling* intoduced in 2.4.1.1):

If x *is a Kollektiv with respect to distribution* (1–p,p), *then* $\mu_p\{y|\ x/y$ Kollektiv with respect to (1–p.p)$\} = 1$.

Already from this form of the principle, which is considerably weaker that the version that will be proved in 4.5, it is clear that the content of proposition 2.3.2.2 is not likely to be exhausted by its recursive analogue stated above. In other words, the principle of homogeneity, which is in itself a purely quantitative statement not mentioning admissibility, suggests that there are many more admissible place selections than just those which are recursive.

3. The recursive place selections owe their appeal to the circumstance that they are a priori admissible. But there might be many more such selections, even disregarding possible wider interpretations of the term "lawlike". We shall not consider these wider interpretations (such as hyperarithmetical, constructible), since, although the admissibility of selections thus defined is a priori, the *truth* of the axiom that Kollektivs are invariant under these admissible place selections is by no means a priori; and our experience with constructible, non-recursive, place selections is restricted, to say the least. In fact, one might also argue that the class of recursive place selections is already much too large.

Physical processes are a possible source of *a priori* selections, that is to say, if these processes are in some sense physically independent of the process which generates the Kollektiv from which is to be selected. Another source is the human mind (but perhaps this example can be subsumed under the previous one): a choice sequence seems no less an admissible place selection than e.g., the sequence of primes (at least if the mind generating the sequence has no prognostic or telepathic abilities). The trouble with these examples is, that they do not lead to

a well defined class of admissible place selections, considered as functions on the infinite binary sequences. If we select from the Kollektiv produced by a coin using the outcomes of the tosses of a second coin, all we can say a priori is that the second coin will produce a sequence in 2^{ω}.

Of course we trust that it will produce a sequence which is independent of the first sequence and hence an admissible selection for that sequence. But to describe this situation, we must widen our framework and consider, not only a priori admissibility, but also admissibility with respect to a given Kollektiv, in conformity with the intensional character of admissibility mentioned in 2.3.3.

However, there exist situations in which the old framework (i.e. admissibility as a priori property) suffices and which nevertheless give rise to continuously many admissible place selections: the special case of independence discussed under the name *disjointness* by Furstenberg [30], is a case in point. The place selections obtained in this way are defined in 5.6.

4. The remarks in 3. point toward a general conclusion: lawlikeness is not as fundamental as may seem at first sight. What *is* fundamental is a relation of physical independence between the process generating the Kollektiv and the process determining the selection. A lawlike selection rule is (as far as we know!) indeed independent of coin tossing in this sense; but there are many other such selection procedures. The physical roots of probability theory, emphasized by von Mises, are obscured rather than illuminated by Church' definition.

2.5.2 The contextual solution We have noted already that von Mises' later presentations of the theory differs slightly from the version given in 2.3 (which dates from 1919). The new version is best described as being contextual: in each specific application of the theory it is assumed that the Kollektiv under consideration is invariant under the place selections needed for that application. This assumption of course has to be justified, and in the process of justification notions such as admissibility or independence may come into play; but they do not form part of the theory.

> Die Festsetzung daß in einem Kollektiv jede Stellenauswahl die Grenzhäufigkeit unverändert läßt, besagt nichts anderes als dieses: Wir verabreden daß, wenn in einer konkreten Aufgabe ein Kollektiv einer bestimmten Stellenauswahl unterworfen wird, wir annehmen wollen, diese Stellenauswahl ändere nichts an den Grenzwerten der relativen Häufigkeiten. Nichts darüber hinaus enthält mein Regellosigkeitsaxiom [i.e. 2.3.1.1(ii)].
> Da nun in einer bestimmten Aufgabe niemals "alle" Auswahlen in Frage kommen, sondern deren nur wenige, so das man jedesmal mit einer eingeschränkten, ad hoc zugeschnittenen Regellosigkeit das Auslangen finden könnte, so kann tatsächlich nichts von dem eintreten, was ängstliche Gemüter befürchten [namely, inconsistency] [70,119].

As an instrumentalist position, von Mises' position is no more absurd than, say, the complementarity interpretation of quantum mechanics. But, if taken to be the whole truth, it

leads to the same type of objection, known as "counterfactual definiteness": the real, physical Kollektiv does not know which computation we are going to perform; we could have chosen to perform a computation different from the one we in fact performed; hence the real Kollektiv must be invariant under "all" place selections. In other words, although for computational purposes an instrumentalist reading of the randomness axiom, *with its abandonment of a definition of Kollektivs*, suffices, explaining the applicability of probability seems to require more (recall that the older theory had both these aims).

The consistency of the contextual version of the theory was settled by Wald [100]. (Note that von Mises wrote the passage quoted just now before Wald's results became known.)

2.5.2.1 Theorem Let $p \in [0,1]$ and let $\mathfrak{H}$ be a countable set of place selections. Put $C(\mathfrak{H},p) := \{x \mid \forall \Phi \in \mathfrak{H}\ (x \in \text{dom}\ \Phi \rightarrow x \in LLN(p))\}$. Then $C(\mathfrak{H},p)$ has the cardinality of the continuum.

This theorem provides for the existence of many Bernoulli sequences or Church random sequences; but its applicability is of course not so restricted. Von Mises was perfectly satisfied with this result [75,92], since any specific application of the theory never involves more than countably many place selections.

We now give a proofsketch of a measure theoretic version of the above theorem, a proofsketch which will at the same time illustrate von Mises' stand on the laws of large numbers.

2.5.2.2 Lemma (Doob [20], Feller [24]) Let $p \in (0,1)$ and let $\Phi: 2^\omega \rightarrow 2^\omega$ be a place selection. Then for all Borel sets $A \subseteq 2^\omega$: $\mu_p\Phi^{-1}A \leq \mu_pA$. If $\mu_p\text{dom}\ \Phi = 1$, we have equality for all A.

Proof See Schnorr [88,23]. □

As a consequence, we have

2.5.2.3 Theorem Let $p \in (0,1)$ and let $\mathfrak{H}$ be a countable set of place selections. Then $\mu_pC(\mathfrak{H},p) = 1$.

Proof Let $\Phi \in \mathfrak{H}$. Since $\mu_pLLN(p)^c = 0$ (theorem 2.2.3.3) we get $\mu_p\Phi^{-1}LLN(p)^c = 0$, by the preceding lemma. □

The theorem is of course most interesting for those $\mathfrak{H}$ which contain only place selections

whose domain has full measure (an assumption which is usually made). Note that we have surreptitiously changed the condition "$p \in [0,1]$" in theorem 2.5.2.1 to "$p \in (0,1)$" in 2.5.2.3, for the simple reason that for $p = 0,1$, the measure μ_p is concentrated at one point. It is possible to give a measure theoretic proof of theorem 2.5.2.1 for the extremal values of p, but in that case one has to use the techniques of 4.6.

The correct interpretation of theorem 2.5.2.3 (from von Mises' point of view) is *not* given by the following quotation from Feller [25,204]:

> Taken in conjunction with our theorem on the impossibility of gambling systems, the law of large numbers implies the existence of the limit [relative frequency] not only for the original sequence of trials but also for all subsequences obtained in accordance with the rules of selection [i.e. admissible place selections]. Thus the two theorems together describe the fundamental properties of randomness which are inherent in the intuitive notion of probability and whose importance was stressed with special emphasis by von Mises.

Feller's remark fits in with the propensity interpretation, which allows one to say that theorem 2.5.2.3 *explains* the impossibility of gambling strategies; but, as we know by now, this is not von Mises' interpretation of probability.
For him, theorem 2.5.2.3 has significance as an existence result only, since μ_p is a measure, not a probability distribution (cf. the careful discussion in [74,41-2]). The theorem shows that the concept of Kollektiv is free of contradiction (in context), but does not thereby render superfluous the empirically motivated axioms for Kollektivs.

2.6. The Geneva conference: Fréchet's objections In 1937, the Université de Genève organized a conference on the theory of probability theory, part of which was devoted to foundational problems (the proceedings of this part have been published as [35]). The focal point of the discussion was von Mises' theory, and especially its relation to the newly published axiomatisation of probability theory by Kolmogorov. The prevailing attitude towards von Mises' ideas was critical. A fairly complete list of objections was drawn up in Fréchet's survey lecture on the foundations of probability [35,23-55]. Von Mises himself was absent, but his rebuttals of the objections were published in the proceedings [35,57-66]. To no avail: the same objections were reiterated in Fréchet's [103]; and, for that matter, ever since. Fréchet's criticism has more or less become the standard wisdom on the subject and for this reason we shall present it in some detail. Our conclusion will be that most of the objections, those based on Ville's famous construction included, are unfounded.

2.6.1 Fréchet's philosophical position In view of the persistent controversy between von Mises and his critics, with arguments seemingly having little or no effect, it seems worthwhile

to investigate why the participants in the debate had so little common ground for discussion. As stated in 2.1, we shall adopt as working hypothesis that the lack of mutual comprehension is due to widely differing views on the foundations of mathematics as well as on the foundations of probability.

The first difference comes out clearly when Fréchet advances the usual "proof of inconsistency" against von Mises. Although the argument itself is identical to that of Kamke reported in 2.3.3, it is worth quoting since it shows the extent of the mutual incomprehension.

> Or la deuxième condition [i.e. 2.3.1.1(ii)] n'imposait aucune limitation au choix de la selection des épreuves après laquelle la fréquence totale devait garder la même valeur. On pouvait donc conclure: ou bien qu'en faisant intervenir la totalité des selections imaginables, elle faisait intervenir un ensemble sans signification concrète precise, ou bien que si l'on considère cet ensemble de selections comme bien défini, il contient la selection S_1 qui retient seulement la suite des épreuves ou l'évenement considéré E s'est produit – ou aura lieu – et la selection S_2 qui ne retient que les autres. L'une au moins de ces suites partielles est infini; si c'est S_1, la fréquence totale de E y est égale à 1; si c'est S_2, elle est égale à zéro. Il n'existe donc pas de collectif où la probabilité d'un évenement soit supérieure à zéro et inférieur à l'unité. Cette observation évidente ayant été faite depuis longtemps de diverses côtés, il nous est difficile de comprendre ce qu'entend M. de Misès, en écrivant que jamais on n'a pu signaler un cas concret de contradictions qui pourraient se produire dans l'application de la notion de collectif [28,29-30].

Cette observation évidente..... it is astonishing to see that Fréchet has not grasped any of the subtle properties of Kollektivs: the intensional character of admissible place selections and the fact that Kollektivs have to be considered as new mathematical objects, so that the above selections S_1 and S_2 cannot be elements of a collection of place selections "bien défini".

Like Kamke, Fréchet reveals himself in this passage as one who believes that all mathematical objects are equally accessible; a view clearly not shared by von Mises (cf. his comparison of Kollektivs with choice sequences)[5].

So far, we have been concerned with different viewpoints on the foundations of mathematics. We now turn to the foundations of probability. We shall assume as working hypothesis that Fréchet is an adherent of the propensity interpretation. This hypothesis will explain at least in part why Fréchet thought that Ville' theorem dealt such a devastating blow to von Mises program. But part of Fréchet's conviction also results from plain confusion.

We shall now compile some passages from Fréchet [28,45-7] to show that he indeed subscribes to the propensity interpretation.

> [...] "la probabilité d'un phénomène est une propriété de ce phénomène qui se manifeste à travers sa fréquence et que nous mésurons au moyen de cette fréquence".
>
> Voici donc comment nous voyons répartis les différents rôles dans la théorie des probabilités. Après avoir constaté comme un fait pratique, que la fréquence d'un évenement fortuit dans un

grand nombre d'épreuves se comporte comme la mesure d'une constante physique attachée à cette évenement dans une certaine catégorie d'épreuves, constante qu'on peut appeler probabilité, on en deduit, par des raisonnements dont la rigueur n'est pas absolue, les lois des probabilités totales et composées et on verifie pratiquement ces lois. La possibilité de cette vérification enlève toute importance au peu de rigueur des raisonnements qui ont permis d'induire ces lois. Ici s'arrête la synthèse inductive.

On fait correspondre maintenant à ces réalités (toutes entachées d'erreurs expérimentales), un modèle abstrait, celui qui est décrit dans l'ensemble des axiomes, lesquelles ne donnent pas – contrairement à ceux de M. de Misès –, une définition constructive de la probabilité, mais une définition descriptive. [...]

Sur l'ensemble d'axiomes est bâtie la théorie déductive ou mathématique des probabilités. Enfin la valeur du choix de cet ensemble est soumise au contrôle des faits, non par la vérification directe, mais par celle des conséquences qui en ont été déduites dans la théorie déductive. La vérification la plus immédiate se présentera en géneral de la façon suivante: on adopte comme mesures expérimentales de certaines probabilités p, p',.... les fréquences f, f',..., correspondantes dans les groupes dépreuves nombreuses. Certains théorèmes de la théorie déductive établissent les expressions de certaines autres probabilités, P,P',..., en fonction de p, p',... . Ayant calculé P,P',... au moyen de ces expressions où l'on a remplacé approximativement p,p',.. par f, f',.., la vérification consistéra à s'assurer que les valeurs approchées ainsi obtenus pour P, P',... sont aussi approchées des fréquences F, F',... qui sont les mesures expérimentales directes de P, P',...

On peut d'ailleurs réduire beaucoup les difficultés pratiques de ces vérifications. Si l'on appelle P_n la probabilité pour que la fréquence dans n épreuves d'un évenement de probabilité p, diffère de p de plus de ε, alors d'après le théorème de Bernoulli, P_n converge vers zéro avec 1/n. Si donc on se content de vérifier expérimentalement qu'un évenement de probabilité assez petite est pratiquement très rare et même qu'une évenement de probabilité extrêmement petite est pratiquement impossible, le théorème de Bernoulli se traduit pratiquement ainsi: quel que soit le nombre ε>0, la fréquence dans n épreuves pourra pratiquement être considérée comme différant de la probabilité correspondante, de moins de ε, si le nombre des expériences est assez grand. Autrement dit, il est inutile d'opérer, pour toutes les valeurs de la probabilité p, la vérification qu'on se proposait. On peut se contenter de la faire quand p est petit. Or cela est beaucoup plus facile; il n'est pas nécessaire de faire de long relevés.

Except for the use of the weak law of large numbers where Popper uses the strong law, Fréchet's version of the propensity interpretation follows the lines laid out in 2.2.3 (although Fréchet seems to be much less aware of his assumptions than e.g., Popper!). It is evident from [28] and [103] that Fréchet considers the propensity interpretation to be much simpler than the strict frequency interpretation. Superficially, this is indeed so: much of that which von Mises struggled to formulate precisely is relegated here to the "synthèse inductive", where "c'est l'intuition qui domine et cherche à dégager, comme elle peut, l'essentiel de la complexité des choses" [28,45]. In particular, as we have seen, the rules of probability do not have to be rigorously derived from the interpretation, in contrast with von Mises' approach. Similarly, Fréchet can do without *limiting* relative frequencies and Kollektivs.

But, although the outward appearance of the propensity interpretation is indeed simple, it is so only because it takes so much for granted. The rules of probability theory are valid for certain phenomena because these phenomena are Kollektivs (2.4.2) and Fréchet's use of the weak law supposes either a large amount of randomness (2.4.3) or some highly theoretical assumption 2.2.3; but even then...). Pragmatic solutions indeed look simple, but a pragmatic attitude does

not contribute much toward an understanding of foundations.

2.6.2 Formal objections Above we considered Fréchet's methodological objections. We now discuss the objections which concern the formal structure of von Mises' theory.

2.6.2.1 Inconsistency Since Fréchet, as we have seen, advances the same "proof of inconsistency" as the one discussed at length in 2.3.3, we need not dwell upon it here. Let us recall only that this objection eventually led Wald to prove the consistency of von Mises' theory in context, on the assumption that each specific computation employs at most countably many place selections.
Fréchet objects that the revision by Wald causes the theory to lose much of its primordial simplicity and elegance. It is hard to make sense of this objection, since Wald's theorem is *metamathematical* in character and shows only that the ordinary deductions can be performed without fear of contradiction. The deductions themselves are in no way affected by the consistency proof.

A really forceful objection, which brings out clearly the underlying difference in the interpretation of probability, is provided by:

2.6.2.2 Ville's construction To understand this objection, we have to go back to the law of the iterated logarithm. In 2.4.3 we stated this law for finite sequences. This time, we state it for infinite sequences, since this is the form used in Fréchet's objection.

Law of the iterated logarithm Let $p \in (0,1)$.

(a) For $\alpha>1$, $\mu_p \left\{x \in 2^\omega \mid \exists k \forall n \geq k \left|\sum_{j=1}^{n} x_j - np\right| < \alpha\sqrt{2p\cdot(1-p)n\log\log n}\right\} = 1$

(b) For $\alpha<1$, $\mu_p \left\{x \in 2^\omega \mid \forall k \exists n \geq k \left(\sum_{j=1}^{n} x_j - np\right) > \alpha\sqrt{2p\cdot(1-p)n\log\log n}\right\} = 1$ and

for $\alpha<1$, $\mu_p \left\{x \in 2^\omega \mid \forall k \exists n \geq k \left(np - \sum_{j=1}^{n} x_j\right) > \alpha\sqrt{2p\cdot(1-p)n\log\log n}\right\} = 1$.

Part (b) in particular shows that the quantities

$$\sum_{j=1}^{n} x_j - np\,,\ np - \sum_{j=1}^{n} x_j$$

exhibit fairly large oscillations. This observation provides the starting point for Ville's construction [99,55-69], which proceeds in two stages (actually, our presentation is slightly

anachronistic, since Ville uses *Lévy's Law*, a precursor of the law of the iterated logarithm, instead of the latter).

1. Given any countable set $\mathcal{H}$ of place selections $\Phi: 2^\omega \to 2^\omega$, Ville is able to construct a sequence $x \in 2^\omega$ with the following properties (we assume the identity is in $\mathcal{H}$):

(i) $x \in C(\mathcal{H}, \frac{1}{2})$ (for $C(\mathcal{H},p)$, see definition 2.5.2.1)

(ii) $\forall n \ \frac{1}{n}\sum_{k=1}^{n} x_k \geq \frac{1}{2}$.

Part (ii) means that the relative frequency of 1 approaches its limit from above, a property which is atypical in view of the law of the iterated logarithm. A very much stronger form of (i) and (ii) will be proven in 4.6.

2. In the second stage of the construction, Ville temporarily adopts von Mises' viewpoint and interprets probability measures on 2^ω as in effect being induced by Kollektivs $\xi \in (2^\omega)^\omega$; so that $\mu_{\frac{1}{2}}A = 1$ must mean:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_A(\xi_\kappa) = 1$$

So far we have considered only Kollektivs in 2^ω; in particular, we have not defined what place selections $\Psi: (2^\omega)^\omega \to (2^\omega)^\omega$ are. Fortunately, we need not do so here, since we may, for the sake of argument, assume that Ville has done so in a satisfactory manner (for those interested in the details, see [99,63-67]). Now put

$$A := \left\{ x \in 2^\omega \mid \forall n \exists k \geq n \left(np - \sum_{j=1}^{n} x_j \right) > \tfrac{1}{2}\sqrt{\tfrac{1}{2} n \log\log n} \right\}.$$

Then Ville shows the following, using 1. :

For any countable set $\mathcal{H}$ of place selections $\Psi: (2^\omega)^\omega \to (2^\omega)^\omega$, there exists $\xi \in (2^\omega)^\omega$ such that

(iii) ξ induces $\mu_{\frac{1}{2}}$ and is a Kollektiv with respect to $\mathcal{H}$

(iv) for A as defined above, $\lim_{n\to\infty} \frac{1}{m}\sum_{j=1}^{m} 1_A(\xi_\kappa) = 0$.

Remark The reader may well wonder what "induces" in (iii) means in view of (iv), since we defined "ξ induces P" to mean:

for all $B \subseteq 2^\omega$, $P(B) := \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_B(\xi_k)$;

but since $P(A) = 0$ (by (iv)), the induced measure P cannot be equal to $\mu_{\frac{1}{2}}$ as claimed by (iii). Therefore (iii) should be understood as follows. A σ-additive measure on 2^ω is determined

completely by its values on the cylinders [w], for finite binary words w; and we do have for the ξ constructed by Ville:

for all w, $2^{-|w|} = \mu_{\frac{1}{2}}[w] = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{[w]}(\xi_k)$.

Ville's construction is thus a very interesting case of the phenomenon that relative frequency is not a σ-additive measure; since if P were σ-additive, it would coincide with $\mu_{\frac{1}{2}}$.

From 1. and 2., Fréchet and Ville derived the following three objections to von Mises' theory.
(a) (From 2) The theory of von Mises is weaker than that of Kolmogorov, since it does not allow the derivation of the law of the iterated logarithm.
(b) (From 1) Kollektivs do not necessarily satisfy all asymptotic properties proved by measure theoretic methods and since the type of behaviour exemplified by (ii) will not occur in practice (when tossing a fair coin), Kollektivs are not satisfactory models of random phenomena.
(c) (From 1) Von Mises' formalisation of gambling strategies as place selections is defective, since one may devise a strategy (a so called *Martingale*) which makes unlimited amounts of money of a sequence of the type constructed in 1., whereas ipso facto (by (i)), there is no place selection which does this.
For those who are accustomed to see Ville's construction as the deathblow to the theory of Kollektivs, its cavalier dismissal by von Mises may come as a surprise: "J'accepte ce théorème, mais je n'y vois pas une objection" [72,66]. In fact, von Mises to some extent anticipated Ville's construction in his discussion of the meaning of probability zero [70,38]. As we have seen (in 2.2.3), von Mises thought that an event having zero probability might occur infinitely often in a Kollektiv. But in this case, the limiting relative frequency is necessarily approached unilaterally, as for the sequence constructed by Ville.
We must now try to understand why von Mises could remain unmoved, when apparently the foundations of his work lay shattered. We believe that objections (a) and (b) are either untenable or based on an interpretation of probability which was not his. Objection (c) is justified, but of no consequence. Before we go deeper into the objections, however, we discuss in more detail the formal structure of Ville's argument.

We simplify a suggestion of Wald[6] to show that Ville's theorem is appreciably less general than may seem at first sight. Consider a countable set $\mathfrak{K}$ of place selections, as in 1. Obviously (i) would be *trivially* true if the x constructed did not belong to the domains of the place selections contained in $\mathfrak{K}$; and the construction would seem to be less interesting in that case. Unfortunately, such cases do occur. For we may define a countable set $\mathfrak{K}$ of recursive place selections as follows:

$\mathcal{H} := \{ \Phi^-_\alpha \mid \alpha \in (0,1)\cap\mathbb{Q} \} \cup \{ \Phi^+_\alpha \mid \alpha \in (0,1)\cap\mathbb{Q} \}$; Φ^-_α (Φ^+_α) is generated by ϕ^-_α (ϕ^+_α)

as in definition 2.5.1.1; ϕ^-_α is determined by $\phi^-_\alpha(x(n)) = 1$ iff $\left(\frac{n}{2} - \sum_{j=1}^{n} x_j\right) > \alpha\sqrt{\frac{1}{2}n\log\log n}$

and similarly $\phi^+_\alpha(x(n)) = 1$ iff $\left(\sum_{j=1}^{n} x_j - \frac{n}{2}\right) > \alpha\sqrt{\frac{1}{2}n\log\log n}$.

Obviously

$$x \in \text{dom } \Phi^-_\alpha \text{ iff } \forall k \exists n \geq k \left(\frac{n}{2} - \sum_{j=1}^{n} x_j\right) > \alpha\sqrt{\frac{1}{2}n\log\log n}$$

and similarly for the Φ^+_α.

Hence, if a sequence x belongs to the domains of the place selections in $\mathcal{H}$, it must exhibit the oscillations prescribed by the law of the iterated logarithm. This means that, when Ville's construction is applied to the set of recursive place selections (say), the constructed sequence x is Church–random partly for trivial reasons. An analogous statement holds for the strengthened form of Ville's theorem proved in Chapter 4. It is then of interest to ask to which countable sets of place selections Ville's construction can be applied non-trivially. The advantage of the measure theoretic proof given in Chapter 4 is, that a furnishes a characterisation of sets of place selections to which the construction is non-trivially applicable:

Ville's theorem applies non-trivially to a collection of place selections if for each Φ *in the collection and for each product measure* $\mu = \Pi_n(1-p_n,p_n)$ *such that* p_n *converges to* $\frac{1}{2}$, $\mu(\text{dom } \Phi) = 1$.

The Φ_w satisfy this condition, but the Φ_α don't. Roughly speaking, the theorem applies to place selections which do not have too much "memory".

These considerations show that Ville's theorem is somewhat restricted in scope. One might even go further and argue that sequences such as constructed by Ville are not Kollektivs at all, even on von Mises' definition; for this it suffices to replace the "implies" in definition 2.5.2.1 by "and". When we discussed this question in 2.5, we remarked that von Mises' *use* of Kollektivs seemed to make such a convention natural: Kollektivs are useful in a particular calculation only if the place selections needed for that application select an infinite subsequence from the Kollektiv. On the other hand, in a Church–style definition of randomness it is clearly impossible to demand that a random sequence belong to the domain of *all* recursive place selections: just consider place selections based on the law of the iterated logarithm for $p \neq \frac{1}{2}$. Fortunately we need not consider the merits of such a modification of the definition of

randomness in detail, since there are weightier arguments which show that the above objections are unjustified. So let us state the import of Ville's theorem in the following way: place selections with "limited memory" do not enforce satisfaction of the law of the iterated logarithm. We now investigate the consequences of this result upon von Mises' theory.

Objection (a) is easiest to dispose of; in fact we have done so already in 2.4.3, when we discussed the meaning of the strong limit laws in von Mises' theory. Stage 2 of Ville's construction shows that, although the version of the law of the iterated logarithm for finite sequences is derivable in von Mises' theory (which implies that it can be interpreted via relative frequency), the version for infinite sequences is not so derivable.
But the latter statement does not mean that von Mises is not able to derive the law as stated in 2.6.2.2, only that this theorem does not have a frequency interpretation (in the space of *infinite* binary sequences).

Far from being a drawback of the theory, this seems to be a very interesting subtlety, which illuminates the status of the law of the iterated logarithm and which nicely illustrates Kolmogorov's note of caution when introducing σ-additivity:

> Wenn man die Mengen (Ereignisse) A aus $\mathbb{E}$ [which in this case is the algebra generated by the cylinders [w]] als reelle und (vielleicht nur annäherungsweise) beobachtbare Ereignisse deuten kann, so folgt daraus natürlich nicht, daß die Mengen des erweiterten Körpers $B(\mathbb{E})$ [the σ-algebra generated by $\mathbb{E}$] eine solche Deutung als reelle beobachtbare Erscheinungen vernünftiger Weise gestatten. Es kann also vorkommen, daß das Wahrscheinlichkeitsfeld $(\mathbb{E},P)$ als ein (vielleicht idealisiertes) Bild reeller zufälliger Erscheinungen betrachtet werden kann, während das erweiterte Wahrscheinlichkeitsfeld $(B(\mathbb{E}),P)$ eine reine mathematische Konstruktion ist [44,16].

Objection (b) raises questions which go to the heart of the foundations of probability. It consists of two parts:
(b_1) Kollektivs are not satisfactory models of random phenomena, since a unilateral approach of the limit will not occur in practice;
(b_2) Kollektivs apparently do not necessarily satisfy all asymptotic laws derived by measure theoretic methods; it is an arbitrary decision to demand the satisfaction of one asymptotic law, viz. the strong law of large numbers at the expense of another, the law of the iterated logarithm.

Ad (b_1). "In practice" we see only finite sequences. Kollektivs were so designed as to be able to account for all statistical properties of finite sequences and they do so perfectly. To that end, a certain amount of idealisation, in particular the consideration of infinite sequences turned out to be convenient. But the consideration of infinite sequences was not an end in itself and von Mises certainly had no intention whatsoever to model infinite random

"phenomena".
The only criterion for accepting or rejecting properties of infinite Kollektivs was their use in solving the finitary problems of probability theory and for that purpose, assuming invariance under place selections suffices. Now objection (b_2) claims that in fact there *does* exist another criterion: satisfaction of asymptotic laws derived by measure theoretic methods. So let us now consider the second part of objection (b).

Ad (b_2). As we have seen in 2.2.3, this objection does not make sense on the strict frequency interpretation of probability, i.e. von Mises' own interpretation. Limiting relative frequencies in Kollektivs do not owe their existence to the law of large numbers. Neither are they invariant under admissible place selections because place selections are measure preserving (lemma 2.5.2.2). Similarly, the fact that the law of the iterated logarithm has been derived (for infinite sequences) does not in itself entail that Kollektivs should satisfy it.
On the propensity interpretation, objection (b_2) makes sense, although in that case it is less clear at whom the objection is directed, since infinite Kollektivs then have no role to play in the theory of probability.
An adherent of the propensity interpretation may study Kollektivs for their own sake, as models for the deductions of probability theory, but to give a "good" definition becomes a fairly hopeless task: since one can't have satisfaction of all properties of probability one, it is necessary to choose, but what are the guiding principles for such a choice?
Note that, although von Mises' theory might seem to be plagued by the same problem (which set of place selections do we choose to define Kollektivs?) it is in reality less vulnerable: you need assume only that amount of invariance which allows you to perform a (successful) computation and if the computation fails to produce the right answer, you know the assumption of invariance was wrong.
No such empirical check exists for definitions of random sequences based on the propensity interpretation, such as those of Martin-Löf and Schnorr considered in the next chapter.

Another way to state von Mises' viewpoint on the relationship between Kollektivs in 2^ω and strong limit laws (considered as subsets of 2^ω) is the following.
If μ_p is considered as just a measure, there is no relationship at all. If μ_p is a veritable probability distribution, then there exists some Kollektiv $\xi \in (2^\omega)^\omega$ such that P_ξ defined by

$$P_\xi(A) := \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_A(\xi_k)$$

conincides with μ_p on some reasonably large algebra $\mathbb{E}$ of events $A \subseteq 2^\omega$.(Von Mises briefly considered this set-up in [75,101]. Interestingly, he attributes it to Doob [20], although it is doubtful whether Doob would have been happy with this attribution[7].) Now even if $P_\xi(A) = 1$,

this statement has no immediate bearing on Kollektivs in 2^ω; it tells us only that "most" ξ_k are in A. For reasonable definitions of Kollektivs in $(2^\omega)^\omega$, the ξ_k are themselves Kollektivs in 2^ω; but we see that there is no reason whatsoever why *all* Kollektivs in the sequence $\xi = (\xi_k)$, much less all Kollektivs in 2^ω, should satisfy A .
If μ_p is considerd as a probability measure, it describes the situation of picking points from 2^ω at random; a situation which is very different from that of picking zeros and ones at random to generate a sequence in 2^ω .
The latter procedure is evidently more constructive; and this was clearly one of the reasons why Borel preferred his own theory of "probabilités dénombrables", based on assumptions 1 and 2 as cited in 2.3.2, to measure theoretic probability [8,195], thus perhaps for the first time introducing free choice sequences (see Troelstra's survey of the history of choice sequences [97]).

Digression Another reason for Borel's preference was his conviction that the practical continuum (consisting of elements which can really be defined) is countable. So he states in the introduction to [8] that "dénombrable" refers to the cardinality of the sample space. Curiously, later authors, including Fréchet [28,53], thought that "dénombrable" refers to σ-additivity, in spite of Borel's statements to the contrary! Now Borel's conviction necessitated a new approach to probability theory, not based on measure theory, since an approach based on the latter seemed to require that the continuum be uncountable. The only measures *he* could think of were (what came to be called:) Lebesque-measure and measures defined from Lebesgue measure via densities; and all of these assign measure zero to countable sets. This point has been completely overlooked by Novikoff and Barone [79], who keep wondering about the "curious oversight" of Borel not to notice that probability theory *is* measure theory. This is not to say that Borel's reasoning is free of muddles; it is possible to do measure theory in a countable continuum, as Bishop [5] has shown. (End of digression.)

Lastly, we come to objection (c): von Mises' formalisation of gambling strategies (as place selections) is not the most general possible, since one can construct a strategy (a so-called Martingale) which may win unlimited amounts of money on the type of sequence constructed in 1. For the present discussion, one need not know precisely what a Martingale is; suffice it to say that it is given by a function $V: 2^{<\omega} \to \mathbb{R}^+$, where V(w) denotes the capital which the gambler , having played according to the strategy, possesses after w has occurred. The full definition will be given in Chapter 3. Ville exhibits a Martingale V such
that for the sequence x constructed in 1., $\limsup_{n\to\infty} V(x(n)) = \infty$; but, obviously, since x is

a Kollektiv, no gambling strategy in the sense of von Mises can win unlimited amounts of money on x. This objection is undoubtedly correct, but not very serious.

The purpose of von Mises' axioms is not to formalise the concept of an infinite sequence for which no successful gambling strategy exists. Rather, the purpose of the axioms is to lay down properties which allow the derivation of probabilistic laws. These properties are indeed justified by an appeal to the (empirical) "principle of the excluded gambling strategy" and perhaps this principle sanctions stronger axioms. For instance, in Chapter 3 we shall study definitions of randomness which take this principle as basic. But stronger axioms are necessary only if the given axioms do not suffice for the derivation of probabilistic laws.
At first sight it might seem that von Mises' theory cannot derive the characteristic properties of Martingales, e.g. the following:

(*) if V is a Martingale (w.r.t. λ), $\lambda\{x \in 2^\omega \mid \limsup_{n\to\infty} V(x(n)) = \infty\} = 0$.

But the situation here is completely analogous to that of the law of the iterated logarithm. There is no trouble in deriving the properties of Martingales in so far as they pertain to finite sequences (e.g. the Martingale inequality, from which (*) can be derived). The extension to infinite sequences is then, again, a matter of convention.
Conversely, we know by now that the derivation of (*) does not justify the requirement that for each Kollektiv y, $\limsup_{n\to\infty} V(x(n)) < \infty$.

But, one might argue, although Kollektivs such as x do not imperil the derivability of probabilistic laws, they may lead to wrong predictions. The following story illustrates what may go wrong and is at the same time an informal exposition of the results that will be obtained in 4.6.

Consider a casino, in which bets are placed on the outcomes of coin tosses. If the outcome is 1, the casino wins, otherwise the gambler wins. Beginning with the foundation of the establishment, the house issues each day a new coin with which the games have to be played. The management of the house, however, is thoroughly corrupt and issues coins which are false: the coin issued on the n^{th} day is such that the probability of heads on this day is $p_n = \frac{1}{2}(1 + (n+1)^{-\frac{1}{2}})$ (so that $p_n > \frac{1}{2}$, but $\lim_{n\to\infty} p_n = \frac{1}{2}$). The reason behind this devious procedure is the following.
A state inspector checkes the honesty of the casino by tossing a coin once a day, jotting down the outcome and testing at the end of the year (say) whether the sequence so obtained is Church random. The management of the house knows that, with the above choice of the p_n, there is a very large probability that the sequence in the inspector's notebook is indeed Church random (lemma 4.6.2). One day, however, the inspector learns of the definition of randomness given by Martin-Löf (Chapter 3), which is a (at least extensionally) a refinement of that of Church, and decides to check, after a year, whether the sequence of outcomes is Martin-Löf

random. Unfortunately for the management, there is also a very high probability that this sequence is not Martin-Löf random (theorem 4.6.1). However, after consulting the relevant literature (corollary 4.6.5), they change the value of p_n to $\frac{1}{2}(1+ (n+1)^{-1})$. To his satisfaction, the inspector notes that the sequences produced are (approximately) Martin-Löf random. The management is also satisfied, since no definition of randomness, however strong, can force them to change the value of p_n to a value which is less advantageous to them.

The moral of this tale is that, for each w, the inspector's prediction for the relative frequency of the occurrence of w on a specific day is false, regardless of whether a Church- or a Martin-Löf random sequence is used for the prediction (and that is the reason why the establishment is so profitable to its owners). Doesn't it follow that von Mises' theory fails in this case? No; the inspector could, on the basis of his data, only predict the relative frequencies of the outcomes of the experiment which consists of grouping (say) n days together and tossing a coin each day. The data are not relevant for the experiment which consists of taking a single day and grouping together the outcomes of n tosses with the coin isssued that day.

This concludes our review of the objections brought forward by Fréchet. These objections do not necessitate a revision either of strict frequentism or of the definition of Kollektivs; but we do not, of course, wish to claim that such objections are logically impossible.

2.7 Conclusions Two themes have occupied us in the preceding pages: the interpretation of probability and the definition of Kollektivs.

1. The great merit of von Mises' theory lies in the rigorous version of the frequency interpretation it presents. This interpretation, strict frequentism, is perhaps not the ultimate truth; but its main rival among the objective interpretations of probability, the propensity interpretation, has not yet arrived at a comparable stage of development, no one having investigated its consequences and assumptions as thoroughly as von Mises did for strict frequentism.

This is not to say that henceforth measure theoretic probability theory should be abandoned in favour of von Mises' theory. We view the relation between the first and the latter much as the relation between classical and constructive mathematics; there is nothing objectionable in doing classical mathematics, but if you really want to know what your results mean, you have to translate them in constructive terms, a translation which is sometimes impossible. Similarly, a deduction in measure theoretic probability theory should ideally be accompanied by a translation in terms of frequencies and Kollektivs; and this translation is not always trivial, as was demonstrated using the law of the iterated logarithm.

2. Von Mises' theory shows very clearly the assumptions that underlie any application of probability theory, in particular the necessity of the assumption that the mass phenomena to which probability theory is applicable be Kollektivs.

The older theory consists of two parts: invariance under place selections as an instrument for deductions and an explanation of invariance via admissibility.
The explanatory part has strong intuitive appeal, but is rather difficult to formalize; although the formalisation implicitly adopted in the alleged proof of inconsistency is blatantly not the one intended by von Mises.
We could distinguish two approaches toward formalisation: identifying admissible selections with lawlike selections and a contextual approach. For various reasons the identification of lawlikeness and admissibility leads to a much too restricted notion of the latter, and in particular leaves out the physical aspects.
Von Mises himself favoured the contextual approach, which means renouncing the attempt to define Kollektivs, but assuming in each specific instance the amount of invariance needed. To justify invariance, one may appeal to admissibility, but it does not occur anymore in the theory.
However, to study the question why probability theory is applicable to certain phenomena it seems best to follow the lines of the older theory and to make precise its basic idea: probabilistic computations are successful when they correspond to admissible place selections. In subsequent chapters we present a piecemeal approach to this problem: different formalisations of admissibility which embody different aspects.
Lastly, we saw that, on the strict frequency interpretation, it suffices to define Kollektivs using place selections only. The demand that truly random sequences satisfy all strong limit laws proved by probability theory stems from a misinterpretation of the condition that limiting relative frequencies in a Kollektiv exist; such a demand can be justified at most on the propensity interpretation of probability.

Nevertheless, the objections voiced by Fréchet were almost universally accepted. Attempts to define Kollektivs became rare. A renewal of interest in the subject occurred only after Kolmogorov emphasized the necessity of Kollektivs for the frequency interpretation. For technical reasons, however, we start, not with Kolmogorov's own proposal, but with a later development: Martin-Löf's definition.

Notes to Chapter 2

1. Kolmogorov's *Grundbegriffe* contains a paragraph on "Das Verhältnis zur Erfahrungswelt" in which he says

> In de Darstellung der notwendigen Voraussetzungen für die Anwendbarkeit der Wahrscheilichkeitsrechnung auf die Welt der reellen Geschehnisse folgt der Verfasser im hohen Maße den Ausführungen von Herrn von Mises [44,3].

But his *condition B* is slightly awkward from a strict frequentist point of view:

> B. Ist P(A) sehr klein, so kann man praktisch sicher sein, daß bei einer einmaligen Realisation der Bedingungen [which determine the occurrence of A or its complement] das Ereignis A nicht stattfindet [44,4].

This condition contains a vestige of the propensity interpretation and does not harmonize very well with von Mises' views on the meaning of probability zero. However, even in von Mises-Geiringer [74,110] we read:

> Hence we assume that *in certain known fields of application the frequency limits are approached fairly rapidly*. We also assume that certain "privileged" sequences (to be expected by the law of large numbers) appear right from the beginning and not only after a million of trials.

Apparently, the part of [74] where this passage occurs was not written by von Mises (see the preface to [74]); I know of no comparable passage in von Mises' own works ([67] to [73]).
2. But note that von Mises' axioms for Kollektivs go much further and attempt to capture the *independence* of the successive casts, using asymptotic properties in a way which is anathema to the intuitionist. See also note 5.
3. For simplicity, we call *independent* what von Mises calls *independent and combinable* [74,31].
4. We saw in 2.4 that lawlike selections do not suffice for this purpose.
5. In our overview of the history of Kollektivs, we did not consider objections inspired by various forms of constructivism. But it will be clear that, for those who hold that the mathematical universe consists of lawlike objects only, Kollektivs are equally impossible. For in this case, if x is a purported Kollektiv, the set $\{n \mid x_n = 1\}$ is itself lawlike (see Reichenbach [85]). Other objections were based on the conviction that the convergence of the relative frequency postulated of Kollektivs had to be uniform; see, e.g., the lecture notes "Grondslagen der Waarschijnlijkheidsrekening" [Foundations of Probability] by D. van Dantzig (library of the Mathematical Institute, University of Amsterdam).
6. Wald's suggestion occurs in [101,98]. He defines a place selection to be *singular* (with respect to Lebesgue measure) if its domain has Lebesgue measure zero. A sequence x is a *Kollektiv in the strong sense* (with respect to $(\frac{1}{2},\frac{1}{2})$ and some countable set of place selections $\mathfrak{H}$) if it is a Kollektiv in the old sense and is, moreover, not contained in the domain of a singular place selection in $\mathfrak{H}$. Now given any countable set of probabilistic laws (with respect to Lebesgue measure) one can construct a set of place selections $\mathfrak{H}$, such that a Kollektiv in the strong sense with respect to $\mathfrak{H}$ satsfies these laws. For by the regularity of Lebesgue measure, the set of sequences not satisfying a probabilistic law is contained in a G_δ set. However, the domain of a place selection is also a G_δ set and it is easy to construct a place

selection whose domain is a given G_δ. Since the complement of a probabilistic law has measure zero, place selections so constructed are ipso facto singular.
Because that part of the law of the iterated logarithm which is of interest to us, is itself a G_δ set, we could use a simpler construction.
7. As will be clear from the discussion of the meaning of independence in 2.4, the measure $\mu_{\frac{1}{2}}$ refers to the following experimental set-up: each time you want to toss a coin, you take a *new* fair coin. In von Mises' theory this situation is to be distinguished from that of repeatedly tossing the same coin: in this case the productrule is provable. Apparently, von Mises considered the possibility of dropping this feature: see his references to the "Tornier-Doob frequency theory" in [75,101]. Tornier's theory is explained in Feller [23], von Mises-Geiringer [74] and Martin-Löf [63].

3 A New Start: Martin-Löf's Definition

3.1 Introduction At the close of the Geneva conference on probability theory (see 2.6) it became clear that von Mises' axiomatisation of the probability calculus had lost the day. Although *sub specie aeternitatis* almost none of the objections brought against von Mises was cogent, Kolmogorov's measure theoretic formalism, which did not attempt to define probability explicitly, was henceforth universally accepted.
With the acceptance of a measure theoretic foundation of probability theory, the necessity of providing a rigorous definition of randomness disappeared. Consequently, from the publication of Ville' s book [99] in 1939 to 1963, interest in the problem dwindled. In 1963, however, Kolmogorov came to the conclusion that the frequency interpretation stood in need of a precise formulation after all. He published a definition of randomness for finite sequences [47] which contains the germ of *Kolmogorov–complexity* (defined in Chapter 5). Martin-Löf, investigating sequences with high Kolmogorov–complexity, gave a definition of randomness [62] involving a particular type of statistical test, namely, *significance tests*. This definition is nowadays the one most generally accepted. In this chapter we introduce Martin-Löf's definition and several variants and discuss their respective merits.

As a consequence of the criticism voiced by Fréchet and Ville, the problem of defining randomness was now conceived as follows: a random sequence (with respect to some probability measure) should satisfy all probabilistic laws for that measure; in other words, the set of random sequences should be the intersection of all properties of probability one. Of course, in this form, the demand is impossible to satisfy, since the required intersection is empty. Hence we have to choose among the properties of probability one; and Martin-Löf's definition is one such choice.

The main result of the previous chapter is that this way of introducing Kollektivs has not much more than the name in common with von Mises' ideas. For one thing, it completely reverses the attitude von Mises expressed in the slogan "Erst das Kollektiv, dann die Wahrscheinlichkeit". What's more, for von Mises a Kollektiv x in 2^ω induces a probability distribution on $\{0,1\}$, not on 2^ω itself; so from his point of view, there is no immediate relation between properties of probability one in 2^ω and Kollektivs x in 2^ω.
Speaking mathematically, a distribution (1–p,p) on $\{0,1\}$ determines a *measure* $\mu_p = (1–p,p)^\omega$ on 2^ω, but this measure is a *probability* only if it is induced by a Kollektiv $\xi \in (2^\omega)^\omega$. To be sure, such a measure can be extremely helpful in proving existence theorems; for instance, in this way we proved that the set of Church random sequences C(p) has μ_p– measure one

(theorem 2.5.2.3). But this result should not be construed as implying that a "true" random sequence should at least be Church random (*because* Church randomness is a property of probability one).
Another consequence of strict frequentism is that the distribution (1–p,p) on {0,1} in no way determines a unique distribution on 2^ω, to wit, μ_p. Indeed, the distribution on (1–p,p) would lead uniquely to μ_p if it were a property of each coordinate, as in the propensity interpretation. But, according to strict frequentism, a Kollektiv x in 2^ω allows no such conclusion: p is really only a limiting relative frequency. It follows that all measures which, in a sense to be made precise in Chapter 4, determine the same limiting relative frequency p, should be treated on equal footing, and existence theorems should not be sensitive to which measure (from the class of measures which determine the same relative frequencies) we choose. Some notation we introduced in Chapter 2 was intended to reflect this point: e.g. the set of Church random sequences with parameter p was denoted C(p), to emphasize the fact that only the limiting relative frequency p is relevant. In Chapter 4 we shall show that, roughly speaking, C(p) has measure one for measures which determine the same p.
The randomness notions which we shall introduce in this chapter are, on the other hand, very sensitive to the underlying measure. This is emphasized by the notation $R(\mu)$, meaning "the set of sequences random with respect to the measure μ". Exactly *how* sensitive to the choice of a measure these notions are, will be investigated in Chapter 4.

Although we may have so far given the impression that the definition of randomness of Martin-Löf and its variants, being conceived in sin, are ipso facto unsatisfactory, this is not our purpose. The preceding chapter should have convinced the reader that randomness defined as the satisfaction of "all" properties of probability one is anathema to the strict frequentist. It is not, however, implied that such a definition does not make sense on any view of probability. In particular, if you subscribe to some variant of the propensity interpretation, which views probability primarily as a physical property of an experimental set-up, it does make sense to have randomness defined with respect to some unique probability distribution on 2^ω.
Indeed, the widespread belief that Kollektivs should satisfy the law of the iterated logarithm, and that probability zero of an outcome should exclude that this outcome occurs infinitely often (at least for a discrete sample space), probably testifies to an instinctive acceptance of the propensity interpretation. Accordingly, the mathematical differences between the two definitions, investigated in detail in Chapter 4, may be seen as a contribution towards the study of the philosophical differences between these two interpretations of probability.

This chapter is organized as follows. In 3.2 we introduce the definitions of randomness of Martin-Löf [62] and Schnorr [88] and we prove some recursion theoretic properties of these definitions (3.2.2-3).

Although most of the results occur already in Schnorr's book, the proofs have been simplified, e.g. by using the so-called Basis Theorem from recursion theory. Apart from added elegance, we thus introduce a technique that will be helpful in Chapter 5.
Having thus prepared the ground, we turn to some problems not usually treated in the literature. For one thing, there is a notable lack of concrete examples of properties which random sequences satisfy. E.g. in Schnorr's book, only the validity of the law of large numbers is verified, not even that of the law of the iterated logarithm. This fact is slightly ironical, since the non-validity of the law of the iterated logarithm for von Mises' Kollektivs was the main impetus behind the new approach.
One of the goals of this thesis is, therefore, to exhibit more examples of properties of random sequences. For a start we prove in 3.3 effective versions of the Borel-Cantelli lemmas, which allow one to show that random sequences satisfy the usual probabilistic laws.
So far, random sequences are considered only from the point of view of probability theory. Martin-Löf's original introduction of random sequences proceeded slightly differently: a sequence was defined to be random with respect to some statistical hypothesis H if it is not rejected by some (effective) statistical test for H at arbitrarily small levels of significance. From this perspective, it is not immediately clear that Martin-Löf's definition is the correct one to use, since there is some controversy surrounding the notion of significance test employed in the definition.
To set the stage for the discussion, we introduce *Martingales* in 3.4. Martingales were first mentioned in 2.6.2, in connection with Ville's construction, as formalisations of gambling strategies. We shall briefly examine this aspect of Martingales, but our main interest lies in their statistical meaning, as *likelihood ratios.* In 3.5 we explain the controversy surrounding significance tests and we discuss some alternatives to Martin-Löf's definition. A conclusion follows in 3.6.
The relation between Martin-Löf's definition and that of von Mises is discussed in Chapter 4, which is considerably more technical than Chapter 3.

3.2 The definitions of Martin-Löf and Schnorr

3.2.1 Randomness via probabilistic laws Ville ended his book [99] on a note of resignation: a random sequence should satisfy all properties of probability one; that's impossible, so which probabilistic laws should we choose? Ville had shown that, in a sense, any probabilistic law can be represented by a Martingale (see lemma 3.4.7 below), so the question could equivalently be posed as: which gambling strategies should one choose? Any choice seemed to be arbitrary, thus causing the definition of random sequences to be arbitrary as well. Of course Ville didn't mind, not being a strict frequentist.

In [62], Martin-Löf proposed a canonical choice for the class of probabilistic laws: the class of those laws which can be proved effectively. To explain this notion of effectiveness, we must look at proofs of probabilistic laws.
A probabilistic law, according to the usual interpretation, is a statement of the form:

$$\mu\{x \in 2^{\omega} | A(x)\} = 1,$$

where A is some formula. The discussion in 2.4.3 should have made clear that this is not von Mises' concept of a probabilistic law; but we are in a different circle of ideas now.
Typically, a proof of such a statement proceeds in either of the two following ways (examples will be given in 2.3):
(i) One constructs a sequence (O_n) of open sets such that (a) $\{x|A(x)\}^c \subseteq O_n$ for all n, (b) $\mu O_n \leq 2^{-n}$ (or any other recursive function of n which decreases to 0), (c) the O_n are recursively enumerable unions of cylinders, or at least unions recursively enumerable in μ and (d) similarly, the function which associates to each n a Gödelnumber for O_n is recursive in μ.
(ii) One uses the two Borel-Cantelli lemmas (Feller [25,200-2]):
(a) if (A_n) is a sequence of sets such that $\sum_n \mu A_n < \infty$, then

$$\mu \bigcap_n \bigcup_{m \geq n} A_m = 0$$

(b) if (A_n) is a sequence of independent events such that $\sum_n \mu A_n = \infty$, then

$$\mu \bigcap_n \bigcup_{m \geq n} A_m = 1.$$

Usually such a sequence (A_n) satisfies properties analogous to (c) and (d) in (i).

Roughly speaking, a probabilistic law is effective if it can be proved according to (i) or (ii). Not all probabilistic laws are effective in this sense; the ergodic theorem (see 7.4) may be a case in point[1].
Martin-Löf's definition of randomness may be seen as a formalisation of procedure (i). Procedure (ii) will receive separate treatment in 3.3.
Let us first introduce two notions of a measure being computable.

3.2.1.1 Definition The probability measure μ on 2^{ω} is called *computable* if there exists a recursive function $g: 2^{<\omega} \times \omega \to \mathbb{Q}$ such that for all w,k: $|\mu[w] - g(w,k)| < 2^{-k}$.
Note that if μ is a computable measure, then the following sets are Σ_1:
$W_> := \{\langle w,a\rangle \in 2^{<\omega} \times \mathbb{Q}^+ | \mu[w] > a\}$ and $W_< := \{\langle w,a\rangle \in 2^{<\omega} \times \mathbb{Q}^+ | \mu[w] < a\}$.
A slightly stronger concept of computability for measures results if we demand that these sets be Δ_1: a measure μ is *strongly computable* if the associated sets $W_<$, $W_>$ are Δ_1.
Evidently a strongly computable measure is computable, but not conversely: strong computability excludes measures μ such that it cannot be decided whether $\mu[w]$ is rational, a

case not very likely to occur in practice. In section 3.4 we have to introduce still another notion of computability for measures, this time weaker than those above.

For computable measures, the clauses "recursive in" in (c) and (d) of (i) can be replaced by "recursive" pure and simple. We shall now formally introduce procedure (i) under the name of "recursive sequential test". This name, coined by Martin-Löf, reflects the statistical origin of these sets, statistical rather than probabilistic. The statistical view will be explained in 3.5.

3.2.1.2 Definition Let μ be a computable measure. $N \subseteq 2^\omega$ is a *recursive sequential test* with respect to μ if N can be written as a $\prod_2$ set $\bigcap_n O_n$, where $O_n \in \sum_1$, the function $n \to O_n$ is recursive, $O_{n+1} \subseteq O_n$ and $\mu O_n \leq 2^{-n}$.

We shall see below that probabilistic laws such as the law of the iterated logarithm or the law of large numbers can indeed be proven by constructing recursive sequential tests covering the sets of sequences not satisfying these laws. In fact, these proofs usually show something more: with the notation as in the preceding definition, one usually has that the μO_n are computable *uniformly in* n, i.e. that for some recursive function $f: \omega\times\omega \to \mathbb{Q}$,

$$\forall n,k\ |\mu O_n - f(n,k)| < 2^{-k}$$

This added feature is present in Schnorr's definition of *total recursive sequential test* [88,63].

3.2.1.3 Definition With the notation of 3.2.1.2: N is a *total recursive sequential test* with respect to μ if μO_n is computable uniformly in n.

Schnorr's reasons for preferring this definition will be examined in 3.2.3 and 3.4. In 3.2.3 we shall see that indeed some recursive sequential tests are not total.
Abstractly, we may now introduce definitions of randomness as follows:

3.2.1.4 Definition Let μ be a computable measure. $x \in 2^\omega$ is *random* with respect to μ (denoted $x \in R(\mu)$) if for all recursive sequential tests N with respect to μ, $x \notin N$.

3.2.1.5 Definition Let μ be a computable measure. $x \in 2^\omega$ is *weakly random* with respect to μ (denoted $x \in R_w(\mu)$) if for all total recursive sequential tests N with respect to μ, $x \notin N$. (Schnorr calls *hyperzufällig* what we call random, and *zufällig* what we call weakly random.)
3.2.1.6 Lemma $R(\mu) \subseteq R_w(\mu)$ and $\mu R(\mu) = \mu R_w(\mu) = 1$.

Proof Each (total) recursive sequential test has measure zero and there are only countably many of them. □

These definitions are very abstract, much more so than that of von Mises. For example, while a probabilistic law gives rise to a (total) recursive sequential test, via procedure (i) on p. 58, the converse does not seem to be obvious: does every recursive sequential test correspond to a bona fide probabilistic law? In order to answer such questions, one must have some kind of representation or classification of recursive sequential tests. Sections 3-5 of this chapter, and also Chapter 4, contain some efforts in this direction. The rest of 3.2 develops some recursion theoretic properties of the above definitions and settles a question left open by lemma 3.2.1.6, namely: is every weakly random sequence also random?

3.2.2 Recursive sequential tests A surprising property of recursive sequential tests is:

3.2.2.1 Lemma (Martin-Löf [62]) Let μ be a computable measure. (a) The collection of recursive sequential tests with respect μ to is recursively enumerable. (b) There exists a universal recursive sequential test with respect to μ, i.e. a test U such that for all recursive sequential tests N with respect to μ, $N \subseteq U$.

A curious consequence of the preceding lemma is that $R(\mu)$ and, a fortiori $R_w(\mu)$, have elements which are rather simple. Although neither set contains recursive sequences if μ is non-atomic (for if x is recursive, $\bigcap_n[x(n)]$ is a total recursive sequential test with respect to any non-atomic computable μ; cf. remark 3.2.3.11), $R(\mu)$ does contain Δ_2-definable sequences. This is a consequence of the following

3.2.2.2 Basis Theorem (Soare [92,109]) Any non-empty $\prod_1$ subset of 2^ω has a Δ_2-definable element.

Proofsketch A $\prod_1$ subset of 2^ω can be viewed as the set of infinite paths through a recursive binary tree T. Call $w \in T$ *admissible* if $\forall n>|w| \exists v \in 2^n (w \subseteq v \ \& \ v \in T)$. (By König's Lemma, w is admissible iff there is an infinite branch of T through w.) The set of admissible words is $\prod_1$. Since the subset is non-empty, T has an infinite branch. The *leftmost* infinite branch can be constructed recursively in the set of admissible words, which is $\prod_1$; hence this branch must itself be Δ_2. □

3.2.2.3 Lemma Let μ be a non-atomic computable measure. Then $R(\mu)$ contains Δ_2-, but no Δ_1-, definable sequences.

Proof (See also Schnorr [88, 56].) By 3.2.2.1, $R(\mu)$ is a $\sum_2$ set of measure 1. Pick a $\prod_1$ set $A \subseteq R(\mu)$ such $\mu A > 0$ and apply the Basis Theorem. If x is recursive and μ computable and

non–atomic, then $\bigcap_n[x(n)]$ is a total recursive sequential test with respect to any non-atomic computable μ; cf. remark 3.2.3.11. □

Although Δ_2 sequences may thus possess all statistical properties associated with randomness, in another sense they can be completely deterministic.

$$\lim_{k\to\infty} (\xi_k)_n.$$

In words: Δ_2 sequences x can be produced by Turing machines if the machine is allowed to correct itself a finite number of times per x_n. This is a far cry from the usual mechanisms that produce random sequences: indeterministic systems such as those of quantum mechanics, or deterministic systems that have been subject to coarse graining (see Chapter 5). The finer tools of Kolmogorov complexity will allow us to distinguish between Δ_2 definable random sequences and those which are not so simply definable.

3.2.3 Total recursive sequential tests The requirement of uniform computability of the μO_n is strong; to prove that a recursive sequential test is in fact total sometimes demands considerable effort. Fortunately, nullsets bearing a strong resemblance to total recursive sequential tests were already known in constructive mathematics, so we can draw upon the large reservoir of proof techniques developed there (see, e.g., the books by Bishop [5], Bridges [9] and Bishop-Bridges [6]) Although not every total recursive function is acceptable in constructive mathematics (since the proof that the function is in fact total must itself be constructively valid), arguments involving constructive functions usually carry over directly to recursive functions; when the result is simple we shall not bother to write down proofs. For instance, we shall often have occasion to use the following comparison principle:

3.2.3.1 Lemma (See [5,30].) Let (a_n), (b_n) be recursive sequences of computable reals such that $0 \leq a_n \leq b_n$ and $\sum_n b_n < \infty$ is computable. Then $\sum_n a_n$ is also computable.

To compute the measure of a $\sum_1$ set, it is often helpful to have such sets presented in normal form, namely as a disjoint union of sets of the form [w]. For if A in $\sum_1$ is brought in such a form, i.e. $A = \bigcup_i [w^i]$, then $\mu A = \sum_i \mu[w^i]$.

3.2.3.2 Definition A subset S of $2^{<\omega}$ is called *prefixfree* if for distinct $w,v \in S$: neither $w \subseteq v$ nor $v \subseteq w$.

If S is prefixfree, the open set determined by S, namely $[S] = \{x \mid \exists n(x(n)) \in S)\}$ can be written as

$$[S] = \bigcup^{\perp}_{w \in S} [w] \quad \text{(where } \bigcup^{\perp} \text{ denotes disjoint union).}$$

3.2.3.3 Lemma For every $\sum_1$ set $A \subseteq 2^{\omega}$, one can effectively determine a recursively enumerable prefixfree set $S \subseteq 2^{<\omega}$ such that $A = [S]$.

Proof A is of the form [T], $T \subseteq 2^{<\omega}$ r.e. Generate T. S is obtained as a union $\bigcup_n S_n$, $S_n \subseteq S_{n+1}$. Suppose S_n has been constructed. Consider the $n+1^{th}$ word w in T. (a) If w is a prolongation of some v in S_n, put $S_{n+1} = S_n$. (b) If w is an initial segment of some v in S_n, replace w by all its prolongations of length |v| and apply (a) and (b) to each of these prolongations. This process comes to a halt; let S_n be the union of S_n and the finite list thus obtained and proceed. (c) In all other cases, put $S_{n+1} = S_n \cup \{w\}$.[2] □

Using this lemma one can easily show

3.2.3.4 Lemma Let μ be a computable measure on 2^{ω}; A, B $\sum_1$ subsets of 2^{ω} with μA, μB computable. Then $\mu(A \cup B)$, $\mu(A \cap B)$ are computable.

Proof We do the first case only. We may suppose that A is written as a *disjoint* union $\bigcup_n [w^n]$; let B = [v]. Then $\mu(A \cup B) = \sum_n \mu([w^n] \cup [v])$ and we may apply lemma 3.2.3.1 with $a_n = \mu([w^n] \cup [v])$ and $b_n = \mu[w^n] + \mu[v]$. For the general case, write B as a disjoint union $\bigcup_m [v^m]$; then $\mu(A \cup B) = \sum_m \mu(A \cup [v^m])$. Apply 3.2.3.1 with $a_m = \mu(A \cup [v^m])$ (which is compuable by the first part of the proof) and $b_m = \mu A + \mu[v^m]$. □

We now come to an essential feature of $\sum_1$ sets O such that μO is computable. If O is just $\sum_1$, it may be the case that all recursive sequences are contained in O; this is for instance true of the levels U_n of a universal recursive sequential test U. Not so for $\sum_1$ sets O with μO computable:

3.2.3.5 Lemma Let μ be a computable measure, O in $\sum_1$ and μO computable. Then for any word w such that $\mu([w] \cap O) < \mu[w]$, there exists a *recursive* x in $[w] \cap O^c$.

Proof This is just a formalisation of an old intuitionistic result; see e.g. Schnorr [88,64-5]. Alternatively, one could show that, if μO is computable, it can be written as a recursive union of cylinders [w] and then apply the lemma proved in footnote 2. □

3.2.3.6 Corollary Let μ be a computable measure which is positive on open sets, A a $\prod_1$ set without recursive elements. Then either $\mu A = 0$ or μA is not computable (both cases occur).

For our purpose the most important consequence is

3.2.3.7 Corollary (a) Let μ be a computable measure. If N is a total recursive sequential test with respect to μ, there exists a recursive $x \notin N$. (b) If μ is non-atomic, there exists no universal total recursive sequential test with respect to μ.

Proof (a) Write $N = \bigcap_n O_n$ as in definition 3.2.1.3. Observe that $\mu O_1 < 1$ and apply lemma 3.2.3.5. (b) Otherwise, by (a), there would exist a recursive sequence outside this universal test. □

Schnorr sees in the preceding lemma a mark of the superiority of total recursive sequential tests over recursive sequential tests. The construction of a recursive x outside N implies that we can construct a model of the probabilistic law corresponding to N, so that we can visualize the property stated by the law (von Mises considered this use of recursive "Kollektivs" in [69]). This is indeed not an unreasonable requirement for probabilistic laws which purport to be effective. But the requirement is satisfied by other types of tests as well (see footnote 2 and section 3.4). Furthermore, the existence of recursive sequences satisfying a probabilistic law does not imply visualizability of that law in any real, practical, sense: there must exist recursive absolutely normal numbers (i.e. numbers which are normal to every base), but there are no examples of absolutely normal numbers which are as easily described as the example of a normal number in lemma 2.5.1.5. It therefore seems more correct to say that, whenever a probabilistic law can be associated with a *total* recursive sequential test, the possibility of a visualizable model for that law is at least not excluded.
We now state a technical lemma which, besides being useful later, will imply that the collection of total recursive sequential tests (with respect to a given measure) is not r.e.

3.2.3.8 Lemma (Schnorr [88,65]) Let μ be a computable measure and $(N_k)_k$ a recursively enumerable collection of total recursive sequential tests with respect to μ. Then $\bigcup_k N_k$ is a contained in a total recursive sequential test M with respect to μ.

Proof Let $N_k = \bigcap_n O_{k,n}$. Put $M = \bigcap_n \bigcup_k O_{k,(n+k)}$. M is a recursive sequential test with respect to μ.

To compute $\mu \bigcup_k O_{k,(n+k)}$, note that for $n+1 < i < j$:

$$\mu\bigcup_{k=1}^{j} O_{k,(n+k)} - \mu\bigcup_{k=1}^{i} O_{k,(n+k)} \leq \sum_{k=i}^{j}\mu O_{k,(n+k)} \leq \sum_{k=i}^{j} 2^{-k-n}$$

hence lemma 3.2.3.4 implies that

$$\left(\mu\bigcup_{k=1}^{j} O_{k,(n+k)}\right)_{j\in\mathbb{N}}$$

is a recursive sequence of computable reals which is recursively Cauchy, so converges to a computable real (see [5,27]). □

3.2.3.9 Corollary Let μ be a non-atomic computable measure. The collection of total recursive sequential tests with respect to μ is not r.e.

Proof Otherwise the M constructed in lemma 3.2.3.8 would be universal. □

We now come to the main result of this section: that $R(\lambda) \subset R_w(\lambda)$. This observation is due to Schnorr [88,77], whose proof uses Martingales and a detour via a different randomness concept.

3.2.3.10 Theorem Let μ be a computable measure. Then there exists a sequence which is weakly random, but not random, with respect to μ.

Proof Let $(N_k)_{k\in\mathbb{N}}$ be an enumeration of the collection of total recursive sequential tests with respect to μ. By lemma 3.2.3.8, we may assume that each N_k is of the form $\bigcap_n O_{k,n}$, where

$$O_{k,n} = \bigcup_{i=1}^{k-1} O_{i,(n+i)}.$$

We construct a weakly random, but non–random x as a pointwise limit of a sequence $(\xi_k)_{k\in\mathbb{N}}$, where $\xi_k \in 2^\omega$. Let $U = \bigcap_n U_n$ be the universal recursive sequential test with respect to μ.
By lemma 3.2.3.5, we can construct a recursive ξ_1 not contained in $O_{1,1}$. Since μ is non-atomic, U contains all recursive sequences. Determine k_1 such that $[\xi_1(k_1)] \subseteq U$. Since $[\xi_1(k_1)]\cap(O_{1,1})^c \neq \varnothing$, there exists a recursive ξ_2 such that $\xi_2(k_1) = \xi_1(k_1)$ and ξ_2 not contained in $O_{2,1}$. Determine $k_2 > k_1$ such that $[\xi_2(k_2)] \subseteq U$. Proceeding inductively we construct recursive ξ_k not contained in $O_{k,1}$. Put $x_n = \lim_{k\to\infty} (\xi_k)_n$. We show that for all k,

$x \notin O_{k,k+1}$. For if $x \in O_{k,k+1}$, say $[x(m)] \subseteq O_{k,k+1}$, we can determine k'>k such that $\xi_{k'}(m) = x(m)$. Since $x_{k'}$ is not contained in $O_{k',1}$ and

$$O_{k',1} = \bigcup_{i=1}^{k'-1} O_{i,i+1},$$

$\xi_{k'}$ is not contained in $O_{k,k+1}$, a contradiction. □

3.2.3.11 Remark If M is a total recursive sequential test with respect to μ, $M = \bigcap_n O_n$, then the conventional upper bound on μO_n is 2^{-n}. This requirement may be relaxed. For if $M = \bigcap_n O_n$ is a $\prod_2$ μ-nullset and each μO_n is computable, then M is contained in a total recursive sequential test N: since for each k, $\mu \bigcap_{n \leq k} O_n$ is computable (uniformly in k) by lemma 3.2.3.2, there exists a total recursive $g: \omega \to \omega$ such that for all m,

$$\mu \bigcap_{n \leq g(m)} O_n \leq 2^{-m};$$

if we then put

$$O'_m = \bigcap_{n \leq g(m)} O_n ,$$

$N := \bigcap_m O'_m$ is the required recursive sequential test.

3.2.4 An appraisal and some generalisations Do the definitions of Martin-Löf and Schnorr really amount to a canonical choice of a class of probabilistic laws, thus providing an *absolute* concept of randomness? Martin-Löf must have had his doubts, since he later proposed to define the set of random sequences as the intersection of all hyperarithmetical sets of measure one [64], the reason being that "the specific Borel sets considered [in probability theory] are always obtained by applying the Borelian operations to recursive sequences of previously defined sets, which means precisely that they are hyperarithmetical" [64,74]. Nor is it clear that this is really the end: why not consider all Borel sets of measure one with codes in some admissible set, the theory of admissible sets being the natural generalisation of recursion theory?
Even if we assume that a random sequence should satisfy all "effective" laws of probability theory, still "effectiveness" is an open-ended notion, so we can't expect to arrive at some definitive notion of randomness in this way. The question is, whether we would be much happier with such a definition.

We believe that the alleged "problem of the relativity of randomness" is a pseudo-problem, born from an excessive concern with abstract things. The fundamental concept of mathematics, set, is relative (with respect to axioms and models for set theory), but that doesn't imply that the notion is useless; only that we should stick to those properties which are uncontroversial, whenever possible. Very few mathematicians are willing to forego sets, just because the contours of the universe of sets are hazy. Some, notably Kreisel, even believe that philosophical analysis of the notion of set may help to enlarge the charted domain.

The situation with respect to random sequences is different in so far as it is quite possible to do mathematics without them; and one is of course much less willing to bear with a problematic concept if one can forego it. We have seen in the previous chapter, however, that random sequences are necessary for a frequentist foundation of probability and in particular that random sequences should minimally be invariant under admissible place selections. Invariance under place selections also suffices to explain the applicability of probability theory, so that Martin-Löf's definition is threatened by relativity only because it disregards the function of random sequences in von Mises' probability theory. The propensity interpretation does nothing to remove this relativity.

We therefore propose to investigate the modern definitions of randomness, not with a view to single out one as *the* definition, but rather to establish reasonable (or just interesting) properties of random sequences. This attitude entails that we do not introduce sets which are more complex (in the sense of the arithmetical hierarchy) than those occurring in definitions 3.2.1.2-3, unless we are forced to do so (see below). We wish to remain agnostic about the exact boundary of the set of properties a random sequence has to satisfy (when these sequences are not considered in their role as foundation for probability theory). The fact that we shall almost never consider sets which are more complex than those in definitions 3.2.1.2-3 does *not* imply that we believe that all (total) recursive sequential tests are reasonable probabilistic laws, since it depends on one's views on, e.g., statistics (does significance testing make sense in the absence of an alternative hypothesis? what exactly *is* an alternative hypothesis?) which properties of random sequences to accept. All in all, then, we regard the definitions of Martin-Löf and Schnorr as convenient way-stations, as technically elegant, concise descriptions of probabilistic laws. But we think that, in their present form, these definitions are too abstract and that questions such as "Is Martin-Löf's definition the right one to use?" do not make sense. Moreover, worrying about the recursive aspects of the definition might easily lead to a neglect of its more urbane questionable aspects.

We shall now examine possible reasons for enlarging the framework. Up till now, we have considered only computable measures. What happens if, for some reason or other, we wish to consider measures which are not computable? A moment's reflection on how a measure μ occurs in a probabilistic law (or a glance at section 3.3) will show that the most useful concept in this context is 3.2.1.2. with "$\prod_2$" replaced by "$\prod_2$ in μ". Most theorems hold for the new concept if we put in "recursive in μ" in the appropriate places; section 3.3 will provide illustrations of this point. Consequently, allowing non-computable measures does not really amount to a generalization.

We *do* get a generalization if we drop the requirement in 3.2.1.2 that μO_n be bounded by 2^{-n};

that leaves us with just a bare $\prod_2$ μ-nullset. Once we're on this slippery slope, we could replace the $\prod_2$ set by a $\prod_n$ set, for arbitrary n. This is indeed what happens in Gaifman and Snir [34]. They introduce

3.2.4.1 Definition Let μ be a computable measure. x is n–*random* with respect to μ (Notation: $x \in R_n(\mu)$) if for all $\prod_n$ μ-nullsets N, $x \notin N$.

It will turn out (in Chapter 4) that the concept is actually most useful for strongly computable measures, which were defined in 3.2.1.1. Again, if we wish to consider arbitrary measures μ, it is best to replace "$\prod_n$" by "$\prod_n$ in μ".

It is doubtful whether we really do need this generality. I know of one probabilistic law which may not be effective in the sense introduced in 3.2.1: the ergodic theorem (which is stated in the appendix, 6.4). In this case, e.g. the set

$$\{ x \in 2^{\omega} | \ \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k > \mu[1] > \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_k \}$$

is $\sum_3$ in μ, i.e. a countable union of $\prod_2$ μ-nullsets; so here at least is some use for 2–randomness.
Let us therefore in conclusion of this part compare 2–randomness (definition 3.2.4.1) with randomness (definition 3.2.1.4).

3.2.4.2 Lemma Let μ be a non-atomic computable measure. (a) There is no universal $\prod_2$ μ-nullset. (b) There exist sequences which are random, but not 2–random, with respect to μ.

Proof (a) Suppose U were a universal $\prod_2$ μ-nullset. Then $\mu U^c = 1$ and U^c is $\sum_2$. It then follows from the Basis Theorem (3.2.2.2) that U contains a Δ_2 definable sequence x. But then $\{x\}$ is a $\prod_2$ set and $\mu\{x\} = 0$ by non-atomicity of μ. (b) If not, then $R(\mu)^c$ would be a universal $\prod_2$ μ-nullset. □

In fact, as an application of the techniques developed in Chapter 4 we shall show in 4.7 that for some continuous measure μ: $\mu(R(\lambda) \cap R_2(\lambda)^c) = 1$.

3.3 Probabilistic laws After these abstract considerations, let us now exhibit some concrete examples of probabilistic laws which are satisfied by (weakly) random sequences. The main technical tools here are effective versions of the two Borel-Cantelli lemmas (Feller [25,200-2]).

3.3.1 Lemma Let μ be a computable measure, $(A_n)_{n\in\mathbb{N}}$ a recursive sequence of Σ_1 sets in 2^ω such that each μA_n is computable (uniformly in n) and $\sum_n \mu A_n$ converges recursively[3]. Then $N := \bigcap_n \bigcup_{k\geq n} A_k$ is a total recursive sequential test with respect to μ.

Proof Obviously N is $\prod_2$. $\mu \bigcup_{k\geq n} A_k$ is computable since for $m_2 > m_1$,

$$\mu \bigcup_{k=n}^{m_2} A_k - \mu \bigcup_{k=n}^{m_1} A_k \leq \sum_{k=m_1}^{m_2} \mu A_k$$

and decreasing to 0 since $\sum_n \mu A_n$ converges. Now apply remark 3.2.3.11. □

Seeing that one automatically obtains a *total* recursive sequential test, starting from the natural condition that $\sum_n \mu A_n$ converges constructively, one might wonder whether there exists some condition which yields only recursive sequential tests. There is, namely:

$$\text{for some total recursive } f\colon \omega \to \omega, \text{ for all n: } \sum_{k\geq f(n)} \mu A_k \leq 2^{-n};$$

but, in practice, whenever in an application of the first Borel-Cantelli lemma the latter condition is satisfied, so is the more exacting condition of lemma 3.3.1. This illustrates a general phenomenon: it is hard to come up with *natural* examples of recursive sequential tests which are not total (they may come from the theory of Martingales, to which the next section is devoted). Nevertheless, it will become clear in the sequel and especially in Chapter 5, that Martin-Löf's concept has immense technical advantages.

Likewise we have the following effective analogue of the second Borel-Cantelli lemma:

3.3.2 Lemma Let μ be a computable measure, $(A_n)_{n\in\mathbb{N}}$ a recursive sequence of independent Σ_1 sets in 2^ω such that $\sum_n \mu A_n$ diverges and μA_n is computable (uniformly in n). Then $\bigcup_n \bigcap_{k\geq n} A_k^c$ is contained in a total recursive sequential test with respect to μ.

Proof By the second Borel-cantelli lemma (Feller [25,201]), $\mu \bigcap_{k\geq n} A_k^c = 0$, for each n. $\bigcap_{k\geq n} A_k^c$ is a $\prod_1$ set, which by remark 3.2.3.11, can be taken to be a total recursive sequential test. Now apply lemma 3.2.3.8. □

As an application of the preceding material, we shall now prove the strong law of large numbers for (weakly) random sequences. The probabilistic argument is copied from Feller [25,259], but we have to complicate the construction to ensure computability.

3.3.3 Theorem Let $\mu = \prod_n(1\text{-}p_n,p_n)$ be a computable product measure. For a recursive and dense (in (0,1)) set of computable reals ε, the sets

$$\left\{ x \in 2^\omega \mid \forall m \exists n \geq m \left| \frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} p_k \right| > \varepsilon \right\}$$

are contained in a total recursive sequential test with respect to μ.

Proof Choose > 0 and rational. Let

$$A_k := \left\{ x \in 2^\omega \mid \exists n (2^{k-1} < n \leq 2^k \;\&\; \left| \frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} p_k \right| > \varepsilon) \right\}$$

The obvious candidate for a total recursive sequential test is $\bigcap_n \bigcup_{k\geq n} A_k$, but there is a slight problem here: μA_k need not be computable, even if ε is rational; for we might not be able to decide whether

$$\frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} p_k = \varepsilon$$

for pathological μ. One may circumvent this problem by restricting μ to be strongly computable (definition 3.2.1.1) or by choosing ε such that we know *in advance* that this situation cannot occur. Now every number

$$\left| \frac{1}{n}\sum_{k=1}^{n} x_k - \frac{1}{n}\sum_{k=1}^{n} p_k \right|$$

is of the form

$$\left| \frac{m}{n} - \frac{1}{n}\sum_{k=1}^{n} p_k \right| =: a_{mn}, \text{ where } m \leq n.$$

Obviously each a_{mn} is computable and the sequence $(a_{mn})_{m,n\in\mathbb{N}}$ is recursive. By repeated diagonalisation one may then construct a recursive sequence of computable reals $(\varepsilon_j)_{j\in\mathbb{N}}$ such that $\lim_{k\to\infty} \varepsilon_j = 0$ and for all j, n and m: $\varepsilon_j \neq a_{mn}$.

Now if we set, in the definition of A_k, ε equal to ε_j, we do have that A_k is $\sum_0$, (A_k) is recursive and μA_k is computable (uniformly in k). (A similar argument occurs in 4.4, where we need an effective version of the Baire Category Theorem to effect the iterated diagonalisation.) The argument then follows familiar probabilistic lines: if s_n is the variance of μ at the n^{th} coordinate, then $s_n = p_n \cdot (1 - p_n)$ and since for all n, $s_n \cdot n^{-2} \leq n^{-2}$, $\sum_n s_n \cdot n^{-2} \leq \sum_n n^{-2} = \pi^2/6$ converges constructively by lemma 3.2.3.1. By Kolmogorov's inequality (Feller [25,234]),

$$\mu A_k \leq 4\cdot\varepsilon_j^{-2}\cdot s_{2^k}\cdot 2^{-2k}$$

hence

$$\sum_k \mu A_k \leq 4\cdot\varepsilon_j^{-2}\cdot\sum_k 2^{-2k}\cdot\sum_{n=1}^{2^k} s_n = 4\cdot\varepsilon_j^{-2}\sum_n s_n\cdot\sum_{2^k\geq n} 2^{-2k} \leq 8\cdot\varepsilon_j^{-2}\sum_n s_n\cdot n^{-2}.$$

Now apply lemma 3.3.1. □

The law of the iterated logarithm can be proved similarly, this time using both effective Borel-Cantelli lemmas and the proof of the law of the iterated logarithm in Feller [25,205]. In Chapter 4 we shall construct examples of probabilistic laws not hitherto considered in the literature.

In conclusion of this section, let us investigate what happens if we drop the requirement in lemma 3.3.3, that the product measure μ be computable. Since there is now no sense in requiring the μA_k to be computable, we may choose rational $\varepsilon > 0$. We then have that the sequence (A_k) is recursive in μ and that the upper bounds on μA_k are given by a recursive function of k, by the inequality $\sum_n s_n\cdot n^{-2} \leq \sum_n n^{-2}$. This illustrates our claim in 3.2.4, that the most useful concept of effective probabilistic law for *arbitrary* μ is obtained if we replace in definition 3.2.1.2, "Π_2" by "Π_2 in μ".

3.4 Martingales As a technical prelude to 3.5, where we examine Martin-Löf's original way of introducing random sequences, we present a different a characterisation of random sequences, using Martingales, Ville's formalisation of the concept of a gambling strategy.

Von Mises' axioms for Kollektivs were stated in terms of admissible place selections and did not mention gambling strategies. The second axiom, however, was explained informally as the "principle of the excluded gambling strategy"; so it is natural to ask whether all gambling strategies can be represented as place selections. As we have seen in 2.6.2, Ville [99] showed that such is not the case. He argued that place selections left one essential element of gambling strategies out of consideration: the possibility to vary one's stakes from one bet to the next. We now give a rapid introduction to the definition and main properties of gambling strategies with variable stakes, so-called *Martingales*, and afterwards discuss their interpretation.

The stakes are given by functions $B_0, B_1: 2^{<\omega} \to \mathbb{R}^+$ as follows: we bet $B_0(w)$ on the event that w is followed by 0 and $B_1(w)$ on the event that w is followed by 1. If V(w) denotes our capital after the sequence w has occurred, we must have (we exclude loans): $B_0(w) + B_1(w) \leq$

V(w). We say that the game played with strategy V is *fair* if, for each n, the expected capital after the n+1th trial is equal to the capital after the nth trial. To formalize this condition of fairness we need a probability measure μ on 2^ω. Having a probability measure, we may then define Martingales.

3.4.1 Definition Let μ be a measure on 2^ω. V: $2^{<\omega} \to \mathbb{R}^+$ is a (positive) Martingale with respect to μ if $V(\langle\,\rangle) < \infty$ and for all w:

$$V(w) = \frac{\mu[w0]}{\mu[w]} \cdot V(w0) + \frac{\mu[w1]}{\mu[w]} \cdot V(w1).$$

The relation to the usual probabilistic concept (see e.g. Feller [26] and Neveu [77]) should be clear: let B_n denote the algebra generated by the cylinders of length n, V_n: $2^{<\omega} \to \mathbb{R}^+$ the function defined by $V_n(x) = V(x(n))$, then the sequence (V_n) is a Martingale (in the usual sense) with respect to μ and the filtration (B_n).

We say that a Martingale V is *successful* on a sequence x if $\limsup_{n\to\infty} V(x(n)) = \infty$. The following lemma, called Kolmogorov's inequality for Martingales by Feller [26,242], but which occurs already in Ville [99,100], shows that Martingales (with respect to μ) are almost never (again with respect to μ) successful.

3.4.2 Lemma Let V be a Martingale with respect to μ, then for $a \in \mathbb{R}^+$

$$\mu\{x \in 2^\omega \mid \exists n\,(V(x(n)) > a)\} \leq \min\left(\frac{V(<>)}{a}, 1\right).$$

As a consequence,

$$\mu\{x \in 2^\omega \mid \limsup_{n\to\infty} V(x(n)) = \infty\} = 0.$$

3.4.3 Examples

1. Let Φ be a place selection (see definition 2.5.1.1). Choose $p,q \in (0,1)$. Define a Martingale V_q with respect to the measure μ_p by

(i) $V_q(\langle\,\rangle) = 1$

(ii) if $\phi(w) = 0$, let $V_q(w) = V_q(w0) = V_q(w1)$

(iii) if $\phi(w) = 1$, put $V_q(w0) = V_q(w)\cdot(1-q)/(1-p)$ and $V_q(w1) = V_q(w)\cdot q/p$.

Then V_q is a Martingale with respect to μ_p, and one can show that $\Phi(x) \notin$ LLN(p) iff for some q, $\limsup_{n\to\infty} V_q(x(n)) = \infty$ (see Schnorr [88,78-82]). (For the definition of LLN(p), see 2.3.2.3.) So Martingales are indeed generalisations of place selections.

2. Likelihood ratios. Let μ_0, μ_1 be probability measures on 2^ω. Put $V(w) = \mu_0[w]/\mu_1[w]$, then V is called the *likelihood ratio* of μ_0 and μ_1 and V is a Martingale with respect to μ_1:

$$\frac{\mu_1[w0]}{\mu_1[w]} \cdot \frac{\mu_0[w0]}{\mu_1[w0]} + \frac{\mu_1[w1]}{\mu_1[w]} \cdot \frac{\mu_0[w1]}{\mu_1[w1]} = \frac{\mu_0[w]}{\mu_1[w]}.$$

Note that some of the Martingales V defined in 1. are also of this form: if the place selection Φ is the identity, $V_q(w) = \mu_q[w]/\mu_p[w]$. In fact, any Martingale in the sense of definition 3.4.1 can be written in the form of a likelihood ratio: if V is a Martingale with respect to μ with $V(\langle\,\rangle) = 1$, and if we define $\mu'[w] := V(w)\cdot\mu[w]$, them μ' determines a probability measure and V is the likelihood ratio of μ' and μ.

In order to obtain a rich supply of recursive sequential tests, we now introduce some computability considerations, in particular a weak notion of computability for measures.

3.4.4. Definition A measure μ on is called *subcomputable* if the set

$$\{ \langle w,a \rangle \in 2^{<\omega} \times \mathbb{Q} \mid \mu[w] > a \}$$

is Σ_1. A Martingale V is called *subcomputable* if the set

$$\{ \langle w,a \rangle \in 2^{<\omega} \times \mathbb{Q} \mid V(w) > a \}$$

is Σ_1.

These concepts are not very natural from the point of view of probability theory, but the representation of recursive sequential tests in terms of Martingales will make clear why they are useful. The following two lemmas can be found in Schnorr [88, 38-44], but, stripped of their recursive content, they go back to Ville [99,87-93].

3.4.5 Lemma Let V be a subcomputable Martingale with respect to some measure μ. Then $\{x| \forall k \exists n V(x(n)) > 2^k \}$ is a recursive sequential test with respect to μ.

Proof By subcomputability, the set $\{x| \forall k \exists n V(x(n)) > 2^k \}$ is $\prod_2$. Without loss of generality we may assume $V(\langle\,\rangle) \leq 1$; then by lemma 3.4.2, $\mu\{x| \exists n V(x(n)) > 2^k \} \leq 2^{-k}$. □

3.4.6 Example Likelihood ratios. Let μ_0, μ_1 be computable measures on such that μ_1 is not absolutely continuous with respect to μ_0. Then there exists a recursive sequential test N with respect to μ_0 such that $\mu_1 N > 0$. Indeed, put $N = \{x| \forall k \exists n V(x(n)) > 2^k\}$, where $V(w) = \mu_1[w]/\mu_0[w]$. By the preceding lemma, N is a recursive sequential test with respect to μ_0. The Lebesgue decomposition of μ_1 with respect to μ_0 can be written as

$$\mu_1 = \int \lim_{n\to\infty} V(x(n))d\mu_0(x) \quad + \quad 1_N d\mu_1,$$

so that if μ_1 is not absolutely continuous with respect to μ_0, then $\mu_1 N > 0$.
We now prove a converse to lemma 3.4.5.

3.4.7 Lemma Let N be a recursive sequential test with respect to some computable measure μ. Then there exists a subcomputable Martingale V with respect to μ such that $N \subseteq \{x| \forall k \exists n V(x(n)) > 2^k\}$.

Proof Write $N = \bigcap_n O_n$ as in definition 3.2.1.2. Put $V(w) := \sum_n n\cdot\mu([w]\cap O_n)\cdot\mu[w]^{-1}$. Then V is a Martingale with respect to μ:

$$\frac{\mu[w0]}{\mu[w]} \cdot V(w0) + \frac{\mu[w1]}{\mu[w]} \cdot V(w1) =$$

$$\frac{\mu[w0]}{\mu[w]} \sum_n n\cdot\mu([w0]\cap O_n)\cdot\mu[w0]^{-1} + \frac{\mu[w1]}{\mu[w]} \sum_n n\cdot\mu([w1]\cap O_n)\cdot\mu[w1]^{-1} = V(w).$$

Furthermore, $V(\langle\rangle) = \sum_n n\cdot\mu O_n \leq \sum_n n\cdot 2^{-n} < \infty$. V is subcomputable since for any set O in Σ_1, $\{\langle w,a\rangle \in 2^{<\omega} \times \mathbb{Q} \mid \mu([w]\cap O) > a\}$ is itself Σ_1.
Lastly, $N \subseteq \{x| \forall k \exists n V(x(n)) > 2^k\}$:
if $x \in \bigcap_n O_n$, then $\forall n \exists m \geq n \forall m' \geq m\ (\mu([x(m')]\cap O) = \mu[x(m')])$, which implies
$\forall n \exists m \geq n \forall m' \geq m\ (V(x(m')) \geq n)$ and this in turn implies $\lim_{n\to\infty} V(x(n)) = \infty$. □

The preceding lemmas may be combined to obtain a characterisation of random sequences along the lines suggested by Ville, namely as sequences which do not admit a successful gambling strategy (where the latter are taken to be Martingales):

3.4.8 Lemma Let μ be a computable measure. Then $x \in R(\mu)$ iff for all subcomputable Martingales V with respect to μ: $\limsup_{n\to\infty} V(x(n)) < \infty$. (Note that, as a consequence of the proof, the latter condition is in turn equivalent to: for all subcomputable Martingales V with respect to μ: $\lim_{n\to\infty} V(x(n)) < \infty$.)

We may now give a more precise discussion of Ville's objection, that not all gambling strategies can be represented as place selections. Recall that Ville could construct $x \in 2^\omega$ which satisfy (where $C(\frac{1}{2})$ is the set of Church-random sequences defined in 2.5.1.7.):

$x \in C(\frac{1}{2})$ and for all n, $\frac{1}{n}\sum_{k=1}^{n} x_k \geq \frac{1}{2}$.

The second property is in contradiction with the law of the iterated logarithm. By the results in section 3.3, the set of sequences not satsfying the law of the iterated logarithm (for the measure λ) is a (total) recursive sequential test with respect to λ. The last lemma then implies that for some Martingale V with respect to λ: $\lim(\sup)_{n\to\infty} V(x(n)) = \infty$. This Martingale V cannot be obtained from a place selection (in contradistinction to the Martingales V_q defined in example 3.4.3). Hence, to give a precise formulation of the "principle of the excluded gambling strategy", one should define Kollektivs using Martingales, not just place selections.

We do not think that this result is a problem for von Mises, who after all does not require that there is no successful gambling strategy, *of whatever kind,* on a Kollektiv. Furthermore, Ville's argument assumes without further ado that Martingales constitute a good formalisation of fair games and indeed that the notion of fairness is itself clear and unproblematic. But that may not be so.

We formulated fairness as follows: a game is fair if, for each n, the expected capital after the $n+1^{th}$ trial is equal to the capital after the n^{th} trial. But taking expectations requires some probability measure; and which probability measure should one consider? Adopting the standpoint of strict frequentism, one might be inclined to say that expectations have to be computed with respect to the measures P_n on 2^n, induced by Ville's Kollektiv x via combination as explained in 2.4 (so that in this case the measures P_n are uniform distributions on 2^n). In other words, one might think that the pay-offs for a game *on x* should be determined by the limiting relative frequencies *in x*. Ville's example shows that, when two people agree to play a game according to *this* concept of fairness, one of them may have a successful gambling strategy on Kollektivs of the type constructed by Ville. What's more, in Chapter 4 we shall show that there exist product measures $\mu = \prod_n(1-p_n,p_n)$ with $\mu C(\frac{1}{2}) = 1$, but $\mu\{x| \limsup_{n\to\infty} V(x(n)) = \infty\} = 1$, for some computable Martingale V (for instance, one may take $p_n = \frac{1}{2}(1 + (n+1)^{-\frac{1}{2}})$). Thus, the first tentative "operational" definition of fairness apparently has to be rejected: although it applies for games with fixed stakes (i.e. place selections), it is not applicable to games with variable stakes. However, it does not seem to follow from the strict frequency interpretation that this is the *only* way in which fairness can be defined.

The intuitive idea behind fairness seems to be that it makes sense to speak of "probability of heads at the n toss". This notion of fairness is clear on the propensity interpretation (or perhaps

one should say: not less clear than the propensity interpretation), so it is not surprising that Ville has no qualms about fairness. But, as we have seen in the previous chapter, from the point of view of strict frequentism one may speak of probabilities at specific coordinates only with reference to Kollektivs $\xi \in (2^{\omega})^{\omega}$. In particular, one must consider infinitely many (infinite) runs of the mechanism that produces the Kollektivs (with which the game has to be played) and then count the limiting relative frequencies *in each coordinate*; and *these* probabilities must determine the pay-offs. Now with this definition, a Martingale with respect to the uniform distribution would no longer be considered fair for a game played with Kollektivs of Ville's type: if each ξ_k is of this type, then the probability of 1 at the n^{th} coordinate will be larger than $\frac{1}{2}$.

In conclusion, we may say that Ville's argument is not relevant for the question how to define Kollektivs, but rather for the examination of the probabilistic assumptions that go into the intuitive notion of a fair game. For games with variable stakes, fairness seems to involve a reference to probabilities at some specified coordinate. An adherent of the propensity interpretation will have no difficulty recognizing such probabilities, but the strict frequentist can only introduce them using a Kollektiv of Kollektivs. If for some reason or other his data consist in only one Kollektiv $x \in 2^{\omega}$, in other words, if his data consist only in a distribution over {0,1}, he cannot decide whether some proposed game is in fact fair. To some, the strict frequentist conception of fairness may seem artificial; but this seeming unnaturalness serves to confirm the impression that the *instinctively* adopted interpretation of probability is the propensity interpretation. Interestingly, the only reference to Martingales that I could find in von Mises' published works expresses his incomprehension:

> Jusqu'ici je n'ai pu encore saisir l'idée essentielle qui serait à la base de la notion de "martingale" et de toute la théorie de M. Ville. Mais je ne doute point que, une fois son livre paru, on s'apercevra à quel point il aurait réussi à concilier les fondements classiques du calcul des probabilités avec la notion moderne du collectif [72,67].

Needless to say, there are no technical obstacles to a treatment of Martingales in von Mises' theory; as for the interpretation of the results obtained, we need not repeat here the observations made in 2.4.3 à propos of the strong limit laws.

We now continue our discussion of the technical aspects of the relationship between randomness and Martingales. In section 3.5 we need more detailed information on the Martingale constructed in lemma 3.4.7. This construction has the following analytical meaning:

3.4.9 Corollary Let N be a recursive sequential test with respect to μ and let V be the Martingale constructed in the proof of lemma 3.4.7. If we put $\mu'[w] := V(w) \cdot \mu[w]$, then μ' is

absolutely continuous with respect to μ.

Proof Put

$$f(x) := \sum_n n \cdot 1_{O_n}(x)$$

then f is in $L^1(\mu)$ and f is the density of μ' with respect to μ:

$$\mu'[w] = \sum_n n \cdot \mu([w] \cap O_n) = \int_{[w]} f d\mu.$$

If μ is Lebesgue measure, one can show that the distribution function of μ' has derivative equal to $+\infty$ at all points of N. □

The parallel theory for *total* recursive sequential tests is considerably less smooth.

3.4.10 Definition The Martingale V is *computable* if for some recursive function g: $2^{<\omega} \times \omega \to \mathbb{Q}$: $\forall n \forall w\ |V(w) - g(w,n)| < 2^{-n}$.

Inspecting the proof of lemma 3.4.7 we see that

3.4.11 Lemma Let μ be a computable measure and let N be a total recursive sequential test with respect to μ. Then there exists a computable Martingale V such that N is contained in $\{x | \forall k \exists n\ V(x(n)) > 2^k\}$.

Proof Write $N = \bigcap_n O_n$ as in definition 3.2.1.3 and define V as in lemma 3.4.7. It suffices to show that the expression $\sum_n n \cdot \mu([w] \cap O_n)$ is computable uniformly in w. Since $n \cdot \mu([w] \cap O_n) \leq n \cdot \mu O_n$ and $\sum_n n \cdot \mu O_n$ is computable, this follows from lemma 3.2.3.1. □

In this case the converse, namely

> If V is a computable Martingale with respect to a computable measure μ, then $N = \{x | \forall k \exists n\ V(x(n)) > 2^k\}$ is a total recursive sequential test with respect to μ,

causes some trouble. Obviously N, so defined, is a recursive sequential test; but we also need to show that $\mu\{x | \exists n\ V(x(n)) > 2^k\}$ is computable (uniformly in k). The obvious way to do this, is to use lemma 3.2.3.1 and first passage times: $\mu\{x | \exists n\ V(x(n)) > 2^k\} = \sum_m \mu\{x | V(x(m-1)) \leq 2^k < V(x(m))\}$; and one could hope that there is some recursive sequence of computable reals (a_m) such that $\mu\{x | V(x(m-1)) \leq 2^k < V(x(m))\} \leq a_m$ and $\sum_m a_m$ converges recursively.

However, it is impossible to choose such a sequence (a_m) independent of the Martingale under consideration, since for each m, one may construct a Martingale V' such that $\mu\{x|V'(x(m-1)) \leq 2^k < V'(x(m))\} = 2^{-k}$. Hence, knowledge of the specific structure of the Martingale is necessary. This is the reason why the Martingale convergence theorem in Bishop [5,225] has to be proven under additional assumptions on the Martingales.

In order to circumvent this problem, Schnorr [88,70-7] proposed a different definition of the total recursive sequential tests associated with computable Martingales.

3.4.12 Definition Let $f: \mathbb{N} \to \mathbb{R}^+$ be a computable function, V a computable Martingale. The set $N = \{x| \limsup_{n\to\infty} V(x(n))\cdot f(n)^{-1} > 0\}$ is called the *nullset of order* f *associated to* V.

In other words, only those sequences are put into the nullset on which V can grow sufficiently fast. With the help of the following lemma one may then show that N is indeed contained in a total recursive sequential test.

3.4.13 Lemma (Schnorr [88,72]) Let V be a computable Martingale. For any rational $\varepsilon > 0$, one can construct a *recursive* Martingale $V': 2^{<\omega} \to \mathbb{Q}^+$ such that for all w, $V'(w) \geq V(w)$ and $V'(w) - V(w) \leq \varepsilon$.

3.4.14 Lemma Let V be a computable Martingale with respect to μ and let be N as in definition 3.4.12. Then N is contained in a total recursive sequential test with respect to μ.

Sketch of proof The total recursive sequential test can be defined by

$$M = \{x|\forall k\exists n\ (V'(x(n)) > 2^k\cdot V'(\langle\ \rangle)\ \&\ V'(x(n)) > f(n))\},$$

where V' is the Martingale constructed in the previous lemma. For a verification that M is indeed a total recursive sequential test, see Schnorr [88,73][4]. □

Although Schnorr claims that the concept of randomness itself suggests consideration of Martingales together with order functions (a sequence should be non-random only if we can detect the non-randomness sufficiently fast [88,70]), we think that definition 3.4.12 is interesting only in those cases in which it follows *from the definition* of a Martingale V that it must grow with speed f on some given nullset. Schnorr has established some results of this kind (see chapter 10 of [88]). In other cases, Schnorr's way out seems to be adhoc.

The considerations of this section therefore suggest a concept of randomness which might be different from that of Schnorr.

3.4.15 Definition Let μ be a computable measure. x is called *Martingale-random* with respect to μ (notation: $x \in R_M(\mu)$) if for all computable Martingales V with respect to μ:
$\limsup_{n\to\infty} V(x(n)) < \infty$.

By lemma 3.4.11, $R_M(\mu) \subseteq R_w(\mu)$; it is difficult to say whether we in fact have equality.

In conclusion of this section we point out that tests of the form $\{x|\forall k \exists n\ V(x(n)) > 2^k\}$, for computable Martingales V, share one of Schnorr's desiderata with total recursive sequential tests: the existence of recursive sequences outside these sets (cf. corollary 3.2.3.7 and the discussion which follows it).

3.4.16 Lemma Let V be a computable Martingale. Then for some recursive x:
$\limsup_{n\to\infty} V(x(n)) < \infty$.

Proof Let V' be the Martingale constructed in lemma 3.4.13. Choose rational $\delta > 0$ and define a recursive binary tree T by $T := [w|\ V'(w) < V'(\langle\ \rangle) + \delta\}$. For every $w \in T$, $w0 \in T$ or $w1 \in T$ by the Martingale property, and we can decide which by the computability of V. Consequently the leftmost infinite branch of T is recursive. □

3.5 Randomness via statistical tests

Originally, Martin-Löf [62] introduced the set of random sequences $R(\lambda)$ as follows: a sequence is random wirh respect to if λ it is not rejected at arbitrarily small levels of significance by any (effective) statistical test for λ. Since this way of introducing randomness raises some interesting problems of its own, we shall now give it a separate treatment. To do so, we must first recall some elementary notions concerning statistical tests. As always, we consider an experiment (or measurement) with two outcomes, 0 and 1.

3.5.1 Types of statistical tests We want to test the hypothesis H_0, that the probability of the outcome 1 of an experiment equals p. We may divide tests of H_0 into two classes:
(a) We may distinguish between tests of H_0 which refer to some alternative hypothesis, the so-called *hypothesis tests*, and *significance tests*, which reject H_0 when an outcome sequence is observed which has sufficiently low probability under the hypothesis H_0, without consideration of alternative hypotheses;

(b) We may also distinguish between tests which use a *fixed sample size*, i.e. tests where the number of repetitions of the experiment is fixed before the execution of the experiment, and tests which are *sequential*, where the data themselves decide how large the sample is to be.

We now proceed to a detailed description. Let us first assume that we have a fixed sample size, say n; hence the set of possible outcomes, the sample space, is 2^n. Under the hypothesis H_0 an outcome sequence w in 2^n is assigned probability $\mu_p[w]$[5]. In essence, a significance test for the hypothesis H_0 consists in a partition of the sample space 2^n in disjoint pieces S_0 and S_1. Observation of an outcome sequence w in S_0 leads to rejection of H_0. Observation of w in S_1 does *not* lead to rejection of H (in practice, this will mean that H_1 is given the benefit of the doubt). S_0 is often called a *critical region*. The probability of S_0 under H_0, namely

$$\sum_{w \in S_0} \mu_p[w]$$

is called the *size* of the test and can be interpreted as the relative frequency of unwarranted rejections of H_0 were this test to be executed very often. Obviously we want the size to be small; how small depends on the importance we attach to the hypothesis.
Usually S_0 and S_1 are determined via a *test statistic*, a function $t: 2^{<\omega} \to \mathbb{R}^+$ which can be seen as a measure of the discrepancy between hypothesis and data. Accordingly, the critical region S_0 is of the form:

$$S_0 = \{ w \in 2^n \mid t(w) > a \}$$

where a is adjusted so as to have, for some preassigned *significance level* α,

$$\sum_{t(w)>a} \mu_p[w] \leq \alpha.$$

How should we choose such test statistics? Obviously not every S_0 of small probability can reasonably be interpreted as a critical region for H_0; e.g. for n = 1000 and p = ½, the set of words in 2^n with 500 ones has very small probability, but to take this set for our critical region would be a silly choice indeed.
This line of reasoning shows that the choice of a test statistic is a delicate matter, and it is still a subject of lively debate whether this choice can be effected at all without the consideration of hypotheses alternative to H_0. In the survey by Cox [18], the issue is stated as follows:

> The central philosophical point concerns whether it is sensible to find evidence against a hypothesis solely because an outcome of relatively low probability has occurred, and without regard to possible alternative explanations. If the labelling of the sample points in the sample space is totally arbitrary and no other information is available, there seems to be no option but to use the absolute test [i.e. significance test in the sense defined above]; such situations do, however, seem quite exceptional in applications [18,53].

Cox' first question is answered with an emphatic *no* by the founding fathers of modern statistical theory, Neyman and Pearson:

> It is indeed obvious, upon a little consideration, that the mere fact that a particular sample may be expected to occur very rarely in sampling from [a certain population] would not in itself justify the rejection of the hypothesis that it had been so drawn [from that population], if there were no other more probable hypothesis conceivable [78,4].

It is clear from Martin-Löf's statistical work [65;66] that he rejects this view (or perhaps one should say that his concept of "alternative hypothesis" is much wider than that of Neyman and Pearson); but let us first expound the view of Neyman and Pearson.

To eliminate the possibility of disastrous choices of the test statistic, Neyman and Pearson propose to introduce the consideration of alternative hypotheses. In the simplest case, we have only one alternative H_1 to H_0, where H_1 states that the outcome 1 has a different probability $q \neq p$. A test for H_1 against H_0 is again specified by a partition (S_0,S_1) of 2^n: S_0 corresponds to rejection of H_0 (and acceptance of H_1) and S_1 corresponds to acceptance of H_0 (and rejection of H_1).
In this case, there are two possibilities for wrong decisions: rejecting H_0 when it is true (*type I error* ; the probability of type I error is called the *size* of the test) and accepting H_1 when it is in fact false (*type II error*; 1 – the probability of type II error is called the *power* of the test) . As in the case of a significance test, the probability of type I error is equal to

$$\sum_{w \in S_0} \mu_p[w].$$

But whereas it makes no sense to speak of type II error for a significance test, for lack of an alternative hypothesis, we may compute the probability of type II error here as

$$\sum_{w \in S_1} \mu_q[w].$$

The interpretation of power is the same is that of size: it measures the performance of the test were it used a large number of times.

The distinction between type I and type II errors allows us to discredit the test defined on p. 81, which rejects H_0: $p = \frac{1}{2}$, upon observation of an outcome sequence of length 1000 with 500 ones. Clearly, in this case, for any $q \neq \frac{1}{2}$,

$$\sum_{w \in S_1} \mu_q[w]$$

is large; and it will be required of a good test that both types of errors are simultaneously small

(they are of course not independent).

Call a test of H_0 against H_1 *most powerful* of level α if

$$\sum_{w \in S_1} \mu_q[w]$$

is as small as is compatible with

$$\sum_{w \in S_0} \mu_p[w] \leq \alpha.$$

In this particular situation, most powerful tests exist and can even be given explicitly; this is the content of the

Neyman-Pearson Lemma[6] For suitable constant c (depending on α, the sample size n as well as on the hypotheses involved): if a partition (S_0,S_1) of 2^n is defined by

$$S_0 = \left\{ w \in 2^n \middle| \frac{\mu_q[w]}{\mu_p[w]} > c \right\}, \; S_1 = 2^n - S_0,$$

then (S_0,S_1) is the most powerful level α test of H_0 against H_1.

The preceding exposition of significance tests and hypothesis tests proceeded on the assumption of a fixed sample size. We now relax this assumption and generalize the description to situations in which the sample size is not fixed beforehand. The following description of sequential tests is borrowed from Wald [102,22].

An essential feature of the sequential test, as distinguished from the [fixed sample size test] is that the number of observations required by the sequential test depends on the outcome of the observations and is therefore not predetermined but a random variable. The sequential method of testing a hypothesis H may be descibed as follows. A rule is given for making one of the following decisions at any stage of the experiment (at the m^{th} trial for each integral value of m):

(1) to accept the hypothesis H

(2) to reject the hypothesis H

(3) to continue the experiment by making an additional observation.

Thus, such a test procedure is carried out sequentially. On the basis of the first observation, one of the aforementioned three decisions is made. If the first or the second decision is made, the process is terminated. If the third decision is made, a second trial is performed [...]. The process is continued until either the first or the second decision is made. The number n of observations required by such a test procedure is a random variable, since the value of n depends on the outcome of the observations.

Formally, a sequential test for *hypothesis* testing may be described as follows. We have a test statistic t: $2^{<\omega} \to \mathbb{R}^+$and constants A,B such that

(1) if $t(w) > A$ and for all $v \subset w$, $B \leq t(v) \leq A$, reject H_0 (accept H_1);

(2) if $t(w) < B$ and for all $v \subset w$, $B \leq t(v) \leq A$, reject H_1 (accept H_0);

(3) if for all $v \subseteq w$, $B \leq t(v) \leq A$, go on testing.

If the measure μ_p corresponds to H_0, and μ_q to H_1, the probabilities of type I and type II errors can be computed as follows:

$$\text{size} = \mu_p\{x \mid \exists n\, (t(x(n)) > A \;\&\; \forall m<n\, (B \leq t(x(m)) \leq A))\}$$

$$1 - \text{power} = \mu_q\{x \mid \exists n\, (t(x(n)) < B \;\&\; \forall m<n\, (B \leq t(x(m)) \leq A))\}.$$

Obviously, we want both types of errors to be as small as possible. Again, for the simple situation of testing one hypothesis against another, there is an optimum result: for given significance level α, one can determine constants A and B such that the *likelihood ratio test* defined by putting $t(w) := \mu_q[w]/\mu_p[w]$ in the decision rules above, is the most powerful test of significance level α.

In a sequential *significance* test we are concerned with one hypothesis H_0 only. In this case the set-up is as follows: we have a test statistic t and a constant A such that H_0 is rejected on the basis of data w if $t(w) > A$; otherwise we go on testing. Of course A is adjusted so as to achieve a prescribed significance level α.

The difficulties we pointed out for fixed sample size tests seem to be even more severe in the sequential case. Not only does the choice of a test statistic present a problem in the absence of an alternative hypothesis; but there seems to be no rational basis for a decision to give H_0 the benefit of the doubt, since there does not appear to be a non-arbitrary way to determine a constant B such $t(w) < B$ entails the decision to stop testing.

So it seems that sequential significance tests are useful for rejecting hypotheses, rather than for accepting them. This point should be borne in mind when we discuss Martin-Löf's definition of randomness via statistical tests.

3.5.2 Effective statistical tests It is now easy to view definition 3.2.1.2 as a formalisation of sequential (significance and hypothesis) tests. Let μ be a computable measure on 2^ω. μ need not be of the form μ_p, since we also wish to study tests applicable in situations not involving independent repetitions of the same experiment. We interpret μ as the nullhypothesis to be tested. Typically, μ contains information about the underlying model (Markov chain, independent repetitions) as well as about the parameters of the model. We are interested in arbitrarily small levels of significance; we may take these levels to be of the form 2^{-k}, $k \in \mathbb{N}$.

Now, in practice, a test statistic t: $2^{<\omega} \to \mathbb{R}^+$ will be a computable function. This implies that the set $\{x \mid \exists n\, (t(x(n)) > A \;\&\; \forall m<n\, (t(x(n)) \leq A))\}$ is $\sum_1$ for suitable choices of A, namely for

those computable A which do not occur in the range of t. As in section 3.3.3, we may construct a recursive and dense (in $\mathbb{R}^+$) set of such A's by iterated diagonalisation.
Clearly, if $(A_k)_{k\in\mathbb{N}}$ is a recursive set of computable reals which do not occur in the range of t, the set $\{x|\forall k\exists n\ (t(x(n)) > A_k\ \&\ \forall m<n\ (t(x(m)) \leq A_k))\}$ is $\prod_2$.

If the sequence (A_k) is such that $\mu\{x|\exists n\ (t(x(n)) > A_k\ \&\ \forall m<n\ (t(x(m)) \leq A_k))\} \leq 2^{-k}$, we have arrived at a recursive sequential test with respect to μ. This, in a nutshell, is the statistical motivation of definition 3.2.1.2. Note that the test statistics are subject only to restrictions of a recursion theoretic nature.

3.5.3 Discussion Seeing that every effective statistical test corresponds to a recursive sequential test, we may now ask for a converse: does every recursive sequential test determine an acceptable statistical test? To settle this question,we have to investigate the influence of the reservations concerning significance tests, expressed above, on the proposed definition of randomness. In essence, these reservations come down to this: it is impossible to construct good test statistics without consideration of alternative hypotheses. "Good" here means: the test based on the statistic should not reject the hypothesis when it is intuitively true.
This danger can be avoided if we require that the critical region is in a sense minimal: only reject the nullhypothesis on the basis of data w if w is more plausible on some other hypothesis. In Lévy's words

> Si donc en présence d'une suite remarquable nous excluons la première hypothèse [of the random origin of the data] ce n'est pas que le hasard ait *a priori* moins de chance de la produire qu'une autre; c'est qu'une cause autre que le hasard a plus de chance de la produire [57,92].

Does the alternative hypothesis necessarily have to be of probabilistic origin, stating a different value of a parameter, or perhaps a different model? In other words, should the condition "if w is more *plausible* on some other hypothesis" be interpreted as "if w is more *probable* on some other hypothesis"? The talk of *chance* in the above quotation strongly suggests so and, as we have seen, this was certainly the view of Neyman and Pearson.

If this is indeed the case, we may be led to a notion of randomness which is likely to be different from that of Martin-Löf (or Schnorr), depending upon the definition of "alternative hypothesis" in this abstract setting. The function of the alternative hypothesis ν is to assign a high probability to events to which μ assigns a low probability. If we take "high" and "low" in an absolute sense, so that "high" means "close to 1" and "low" "close to 0", we may regard ν as an alternative to μ if $\nu\perp\mu$.

3.5.3.1 Definition Let μ be a computable measure. Put

$$R_H(\mu) := \left\{ x \mid \text{for all subcomputable measures } \nu\text{: } \limsup_{n\to\infty} \frac{\nu[x(n)]}{\mu[x(n)]} < \infty \right\}.$$

3.5.3.2 Remark $R_{MH}(\mu)$ may be defined as $R_M(\mu)$, except that we require the measures to be computable. This is obviously the more natural concept, but in this case we have *trivially* $R(\mu) \subset R_{MH}(\mu)$, since, by lemma 3.4.16, the diagonalisation argument of Theorem 3.2.3.10 goes through as well in this case. To guard oneself against a trivial solution of the problem, whether a restriction to hypothesis tests enlarges the class of random sequences, one must therefore allow the alternatives to μ to be subcomputable only. The subscript "H" refers to "hypothesis testing"; $R_{MH}(\mu)$ should be interpreted as "the analogue of $R_M(\mu)$ (definition 3.4.15) when we consider hypothesis tests only". For reasons expounded at length in section 3.4, $R_w(\mu)$ probably has no analogue in this sense.

The following lemma shows that sequences in $R_H(\mu_p)$ and $R_{MH}(\mu_p)$ have some reasonable randomness properties:

3.5.3.3 Lemma If $p \in (0,1)$ is a computable real, then $R_{MH}(\mu_p) \subseteq LLN(p)$; moreover, $R_{MH}(\mu_p)$ is invariant under recursive place selections whose domain has full measure.

Proof Consider for $q \in (-1,1)\cap\mathbb{Q}$ the Martingale V defined by

$$V_q(w) := \frac{\mu_q[w]}{\mu_p[w]}.$$

In chapter 10 of [88] Schnorr shows that $N_q := \{x \mid \forall k \exists n V_q(x(n)) > 2^k\}$ is a total recursive sequential test with respect to μ_p and that $x \in LLN(p)^c$ iff for some q, $x \in N_q$. Obviously $\mu_q \perp \mu_p$. If Φ is a recursive place selection whose domain has full measure, then $\mu_q\Phi^{-1} = \mu_q$, so $\mu_q\Phi^{-1}$ is also singular with respect to μ_p. Now apply Schnorr's result with Φx instead of x. □

By lemma 3.4.5, we have $R(\mu) \subseteq R_H(\mu)$ and it is likely that in fact $R(\mu) \subset R_H(\mu)$. To prove equality, for each recursive sequential test N with respect to μ, one must be able to construct a computable measure $\nu \perp \mu$, such that N is contained in

$$\left\{ x \mid \limsup_{n\to\infty} \frac{\nu[x(n)]}{\mu[x(n)]} = \infty \right\}.$$

This is probably impossible; but in section 3.4 we showed that recursive sequential tests, which were introduced by Martin-Löf as significance tests, can always be represented via a likelihood ratio of measures ν and μ, *if we allow that* ν *be absolutely continuous with respect to* μ (lemma 3.4.7 and corollary 3.4.9). The meaning of the condition

$$\limsup_{n\to\infty} \frac{\nu[x(n)]}{\mu[x(n)]} = \infty$$

for absolutely continuous ν with respect to μ, is that neighbourhoods of x have probabilities under ν which are *relatively* much larger than their probabilities under μ; in an absolute sense, however, both probabilities may be small. If *this* concept of alternative hypothesis is reasonable, then so is Martin-Löf's definition of randomness (*modulo* the propensity interpretation). We leave this question open.

3.6 Conclusion Using recursion theory, Martin-Löf has provided a definition of (effective) statistical test and of randomness of great generality. How good a definition of randomness this is, depends, among else, on
– the interpretation of probability
– the interpretation of statistical tests.
We need not here repeat at length the remarks on the foundations of probability made in Chapter 2 and in the introduction to this chapter. For the sake of argument, we shall assume the propensity interpretation and the idea that randomness should be defined as satisfaction of certain statistical laws; let us see how far Martin-Löf succeeds in formalizing this idea.

As regards the interpretation of statistical tests, the very generality of Martin-Löf's definition presents a problem. There is a glaring contrast between the careful, piecemeal discussion of statistical tests in the literature (see for instance Cox [18] and Barnett [3]) and Martin-Löf's sweeping generalisation. It seems to me that there is no use in trying to establish once and for all *all* properties of random sequences if we cannot survey this totality and if there are no *general* arguments for the choice of a particular class of properties. In this case, these arguments would have to be supplied by recursion theory. Now the prospects for such general arguments look bleak: without too much effort we could devise several alternatives to the definitions proposed by Martin-Löf and Schnorr.

If these general arguments do not exist, the use of recursion theory may be rather inessential here. After the discovery of a statistical law which should be true of random sequences, we may determine its recursion theoretic structure; but this structure seems to be rather accidental. It is open to doubt whether there really exists such an intimate connection between randomness and recursion theory. Martin-Löf and Schnorr never seem to question this assumption. We saw in Chapter 2 that the only argument given in favour of such an intimate connection, the identification of admissible and lawlike place selections, is defective and that other concepts, such as entropy, seem to be more relevant. In general, hierarchies which have proved to be useful and natural in recursion theory or mathematical logic, might be unnatural

or even misleading elsewhere. But if that holds true in this case, a definition of randomness should be founded on principles which are less formal and are more concerned with the content of probabilistic laws than those of Martin-Löf.

Also, if more and more concrete examples pile up, there is no guarantee that they will always fit in the straitjacket of definitions 3.2.1.2 and 3.2.1.3. Our remarks on the ergodic theorem (in 3.2.4) and on Martingales (in 3.4) provide cases in point. We don't have much sympathy either for attempts, reviewed in 3.2.4, to fix an upper bound on the arithmetical complexity of statistical tests which is so large that it is inconceivable that it will ever be attained; and even if it were attained, we might have included *too many* properties, witness the discussion on statistical tests.

We conclude that Martin-Löf's definition provides nothing in the way of a *canonical* choice of properties of randomness. We shall therefore take definitions 3.2.1.2 and 3.2.1.3 with a grain of salt and certainly not as the ultimate truth concerning randomness. If, in the sequel, we shall nonetheless use these definitions, it is because they provide a convenient formalisation of a view which is diametrically opposed to that of von Mises; and as such they will be investigated in Chapter 4.

Notes to Chapter 3

1. For an argument to the effect that the ergodic theorem is not constructively valid, see Bishop [5,233].
2. Schnorr's claim [88,37] that S can be chosen to be recursive is false; the universal recursive sequential test provides a counterexample. This is a consequence of the following **Lemma** Suppose the $\sum_1$ set $O \neq 2^\omega$ can be written as the union of a *recursive* set of cylinders [w]. Then there exists a recursive sequence in O^c.

Proof Let $O = \bigcup_n [w^n]$. We may assume that the recursive set $\{w^n \mid n \in \mathbb{N}\}$ is *sequential*, i.e. that every prolongation of some w^n occurs among the w^n. O^c is a non-empty $\prod_1$ set, which is given by a recursive binary tree T. Determine a recursive subtree T' of T by throwing out all the w^n. No infinite branch of T is lost in this process, since no infinite branch of T passes through a w^n. Now every word of T' is admissible in the sense of (the proof of) theorem 3.2.2.2: for if no infinite branch of T' passes through a word v, this means that every infinite branch starting with v must belong to O; but then v must be one of the w^n. Since the set of admissible words of T' is recursive, the leftmost infinite branch of T' is recursive. □

Note that a $\sum_1$ set may be the union of a recursive set of cylinders without having, say, computable Lebesgue measure.

3. A sequence $(a_n)_n$ of computable reals *converges recursively* to a computable real a if there exists a total recursive function g: $\omega \to \omega$ such that for all k,n: $n \geq g(k)$ implies $|a - a_n| < 2^{-k}$. This is the usual constructive definition of convergence couched in recursion theoretic terminology.

4. Since Schnorr wants to consider only Martingales together with some function indicating growth, he must show that every total recursive sequential test is contained in a set of the form defined in 3.4.12. His Satz (9.5) [88,74] purports to establish this, but the proof contains a mistake.

5. This is so by definition if we assume von Mises' concept of probability. Otherwise, we have to add that the repetitions of the experiment are assumed to be independent.

6. We disregard subtleties having to do with randomization at the boundary to achieve the exact significance level α.

4 Place Selections Revisited

4.1 Introduction Now that we have definitions of randomness based on two entirely different ideas, to wit, place selections (Chapter 2) and statistical tests (Chapter 3), we must investigate the relations between these definitions. The main *philosophical* differences are summarized in the Introduction to Chapter 3 and we shall not repeat them here. In this chapter, we shall be interested primarily in the *extensional* relation between von Mises' proposal and that of Martin-Löf. Prima facie, an obstacle to a mathematical investigation of this relation is that, as it stands, von Mises' definition is not formal and does not lead to a well-defined set of random sequences, whereas Martin-Löf's definition does determine such a set. We therefore cannot in any literal sense determine the extensional relationship, but we may ask, for example, how one could introduce admissible place selections in Martin-Löf's framework (note that Martin-Löf's definition as such accords no privileged position to place selections). We shall do so in two steps: sections 4.2-5 contain a quantitative study of the behaviour of random sequences under place selections and 5.6 adds admissibility.

It is perhaps best to view these investigations along the following lines: we take some mathematical model for Kollektivs, in this case random sequences (according to any of the definitions of Chapter 3) and we investigate their adequacy for the expression of von Mises' ideas. In a similar vein, Kamae [40] chooses as a formalisation of Kollektivs the Bernoulli sequences (definition 2.5.1.3) and investigates how these sequences behave under a special class of admissible place selections, the entropy zero sequences (see section 5.6). We do not claim finality for any of these formalisations; we are interested in constructing mathematical models for some of von Mises' ideas, even if these models are only partial or in some respects defective.

While the results of 4.5 show that random sequences share many of the desiderata of Kollektivs, section 4.6 elaborates on Ville's theorem (2.6.2.2) and shows that there are some properties of random sequences which need not be satsfied by Kollektivs, when these are defined using some countable set of place selections. The law of the iterated logarithm is one such property, but not the only one. The novelty of the argument of 4.6 is mainly that it is based as directly as possible on the *philosophical* differences between strict frequentism and the propensity interpretation uncovered in Chapter 2.

We now give an outline of the contents of this chapter. In sections 4.2-5 we state precisely and prove the "principle of homogeneity" first mentioned in 2.5: *if* x *is a Kollektiv with respect to* (1–p,p), *so is almost every subsequence of* x. The main result is Theorem 4.5.2, the version of the principle adapted to Martin-Löf's definition of randomness. The really hard part is 4.4, where we prove various effective versions of Fubini's theorem. In section 4.6 we give a new proof of Ville's theorem, which says that for any countable set of place selections $\mathfrak{H}$, one can

construct a Kollektiv x with respect to $\mathfrak{H}$ which approaches its limiting relative frequency $\frac{1}{2}$ from above, thus contradicting the law of the iterated logarithm. The philosophical significance of Ville's theorem was discussed at length in 2.6.2.2. The idea of the new proof is to construct a non-atomic measure μ on 2^ω such that $\mu(C(\frac{1}{2}) \cap R_w(\frac{1}{2})^c) = 1$, where $C(\frac{1}{2})$ denotes Church-randomness (with parameter $\frac{1}{2}$). We then have at one stroke continuously many Church-random sequences which are not (weakly) random, but the main advantage of the proof is that it also provides an explanation of this phenomenon.

4.2 Place selections from a modern perspective The starting point of our investigations is proposition 2.3.2.2 (von Mises [67,58]):

> *An admissibly chosen subsequence of a Kollektiv is again a Kollektiv, with the same distribution.*

Using recursive place selections one obtains countably many subsequences of a Kollektiv which are themselves Kollektivs, but we noted in 2.5.2 that a "true" Kollektiv was likely to satisfy a stronger property, dubbed the "principle of homogeneity":

> *If* x *is a Kollektiv with respect to* (1–p,p)*, then so is almost every subsequence of* x.

To put the conjecture in a form susceptible to mathematical analysis, we recall some notation from Chapter 2.

4.2.1 Definition Let $x, y \in 2^\omega$ and suppose that y contains infinitely many ones. Then $^x/_y \in 2^\omega$ is determined by

$$(^x/_y)_k = x_m \text{ if } m \text{ is the index of the } k^{\text{th}} \text{ 1 in y.}$$

One may now state the principle of homogeneity as follows:

> *If* x *is a Kollektiv with respect to distribution* (1–p,p), *then* $\mu_p\{x|\ x/y \text{ is a Kollektiv w.r.t. } \mu_p\} = 1$.

This statement is still only semi-formal, since we have not said what we mean by "Kollektiv". We now examine two possible formalizations.

It seems that the first attempt to prove a principle of homogeneity was Steinhaus' [94,305]. He showed (curiously, without mentioning either von Mises or Kollektivs):

4.2.2 Theorem $x \in LLN(p)$ iff for all $q \in (0,1)$: $\mu_q\{y|\ ^x/_y \in LLN(p)\} = 1$.

While this interesting in itself and will be useful to us later, it is defective as a formulation of the principle of homogeneity. It would be satisfactory only if LLN(p) could be replaced by, say, $C(\mathfrak{H},p)$, for arbitrary countable sets of place selections $\mathfrak{H}$; but the proof does not yield this. Hence typical Kollektiv-like behaviour is not incorporated in the theorem. Indeed, we

know of no probabilistic proof which accomplishes this (except for the slightly differently oriented work of Kamae).

In Martin-Löf's set-up, we identify Kollektivs with random sequences and we may prove the principle of homogeneity in the following form (Theorem 4.5.2):

> *Let* $p \in (0,1)$ *be computable and suppose that* ν *is a non-atomic computable measure on* 2^ω. *Then for* $x \in R(\mu_p)$, $\nu\{y/\,{}^x/y \notin R(\mu_p)\} = 0$.

The "almost all" clause in the principle of homogeneity thus refers, not to some specific measure, but to all computable non-atomic measures, indicating (at least for the constructivist) the extreme smallness of the set of subsequences which are not themselves Kollektivs.
But note that the theorem itself does not speak of admissibility (unless we *define*: y is admissible with respect to x if ${}^x/y \in R(\mu_p)$); it has a purely quantitative character. A direct formulation of admissibility must wait until 5.6, when we have at our disposal the notion of Kolmogorov-complexity. There, the techniques used in proving the above theorem will be helpful. One final remark on the principle of homogeneity: it will be observed that the principle states a necessary condition for randomness, whereas Steinhaus' theorem (4.2.2) states a necessary and sufficient condition. We comment on the difference in 4.5.

For completeness' sake, we prove the principle of homogeneity not only for (Martin-Löf) randomness, but for all notions of randomness introduced in Chapter 3. In the case of weak randomness this leads to considerable complexities, but this part of 4.4 can be skipped: section 4.3, lemma 4.4.1 and Theorem 4.4.4 suffice to understand the proof of the main theorem (4.5.2).

4.3 Preliminaries Eventually, in section 4.5, we shall prove

> Let $p \in (0,1)$ be computable and suppose that ν is a non-atomic computable measure on 2^ω. Then for $x \in R(\mu_p)$, $\nu\{y/\,{}^x/y \notin R(\mu_p)\} = 0$.

Here $R(\mu_p)$ refers to Martin-Löf's definition of randomness (3.2.1.4), but the result holds as well if we replace $R(\mu_p)$ by $R_w(\mu_p)$ (definition 3.2.1.5). For the notions of Gaifman and Snir introduced in section 3.2.4 there is an analogous result if we replace "computable" by "strongly computable". In this section we present some preparatory lemmas and motivate the construction to follow.

The method used in the proof of the main theorem is based on the following observations. The first lemma was already mentioned in section 2.5.

4.3.1 Lemma (Doob [20]) Let $p \in (0,1)$. If $\Phi: 2^\omega \to 2^\omega$ is a place selection, A a Borel subset of 2^ω, then $\mu_p\{x|\ \Phi x \in A\} \leq \mu_p A$. If $\mu_p(\mathrm{dom}\Phi) = 1$, then we have in fact equality for all A.

Proof See Schnorr [88,23]. □

4.3.2 Lemma For all $p \in (0,1)$, for all non-atomic measures ν on 2^ω, for all Borel subsets A in 2^ω: $\mu_p \times \nu\{<x,y> |\ ^x/y \in A\} = \mu_p A$.

Proof If y contains infinitely many ones, /y: $2^\omega \to 2^\omega$ is a total place selection. Since ν is non-atomic, the set of y's having only finitely many ones has measure zero. We may therefore write, using the previous lemma and Fubini's theorem: $\mu_p \times \nu\{<x,y> |\ ^x/y \in A\} =$

$$= \int 1_{\{<x,y> |\ ^x/y \in A\}} d\mu_p \times \nu = \int \mu_p\{x |\ ^x/y \in A\} d\nu(y) = \mu_p A .$$ □

4.3.3 Lemma If $O \subseteq 2^\omega$ is $\sum_1$, then the set $\{<x,y> \in 2^\omega \times 2^\omega |\ ^x/y \in O\}$ is $\sum_1$, with Gödelnumber primitive recursive in the Gödelnumber for O.

Proof It suffices to prove the lemma for O = [w]. Now observe that the operation / is completely determined by the operation /': $\bigcup_n (2^n \times 2^n) \to 2^{<\omega}$, as follows:

$(^v/'u)_k = v_m$ if m is the index of the k^{th} 1 in u;

and /' is primitive recursive. □

Lemma 4.3.1 suffices to show that for computable $p \in (0,1)$, $R_w(\mu_p)$ is closed under the action of recursive place selections with domain of full measure. Let Φ be a recursive place selection and suppose $\mu_p(\mathrm{dom}\Phi) = 1$. If $N = \bigcap_n O_n$ is a total recursive sequential test with respect to μ_p, then $\Phi^{-1}N = \bigcap_n \Phi^{-1}O_n$ is $\prod_2$ and by lemma 4.3.1, $\mu_p\Phi^{-1}O_n = \mu_p O_n$, so that $\Phi^{-1}N$ is a total recursive sequential test with respect to μ_p. Obviously, for Martin-Löf's $R(\mu_p)$ we have also invariance under recursive place selections whose domain has measure less than one.

Now let μ,ν be computable measures on 2^ω. In Chapter 3 we defined (total) recursive sequential tests as subsets of 2^ω, but definitions 3.2.1.2-3 are easily generalized to the space $2^\omega \times 2^\omega$ and the measure $\mu \times \nu$. We may then state the most useful consequence of the preceding lemmas as follows:

4.3.4 Lemma Let $p \in (0,1)$ be computable and suppose ν is a computable measure on 2^ω. If N is a (total) recursive sequential test in 2^ω with respect to μ_p, then $\{<x,y> |\ ^x/y \in N\}$ is a (total) recursive sequential test with respect to $\mu_p \times \nu$. Similarly, for $n \geq 2$, if N is $\prod_n \mu_p$-nullset

in 2^ω, then $\{\langle x,y\rangle |\ {}^x/_y \in N\}$ is a $\prod_n \mu_p \times \nu$-nullset in $2^\omega \times 2^\omega$.

This lemma suggests the following strategy for proving the main theorem. Since $R(\mu_p)^c$ is a recursive sequential test with respect to μ_p, the last lemma implies that for any computable measure ν, $\{\langle x,y\rangle |\ {}^x/_y \in R(\mu_p)\}$ is a recursive sequential test with respect to $\mu_p \times \nu$. By Fubini's theorem, $\mu_p\{x |\ \nu\{y |\ {}^x/_y \in R(\mu_p)\} > 0\} = 0$. We are done if we can show that this set of x's is in fact contained in a recursive sequential test with respect to μ_p. That this is so, will be proven in the next section.

4.4 Effective Fubini theorems Let μ,ν be computable measures on 2^ω. This section addresses the following question: if $N \subseteq 2^\omega \times 2^\omega$ is a (total) recursive sequential test with respect to $\mu \times \nu$, is it possible to construct a (total) recursive sequential test M with respect to μ such that $\{x |\ \nu N_x > 0\} \subseteq M$? The answer is yes, but the construction is somewhat complicated, especially in the case of total recursive sequential tests. We also treat briefly the analogous question for $\prod_n \mu \times \nu$-nullsets.

In the following pages we shall often use the phrase "[a real] $b_{n,..}$ is computable, uniformly in (the parameter(s)) n,...". This phrase should be interpreted as: "There exists a total recursive function g such that g(n,...) is a Gödelnumber for an algorithm which computes $b_{n,...}$".

The first lemma is in essence due to Sacks (see Sacks [87] or Kechris [42]). For the definition of strongly computable measures, the reader is referred to 3.2.1.1.

4.4.1 Lemma (i) Let ν be a computable measure on 2^ω and suppose that A is a $\sum_0$ subset of $2^\omega \times 2^\omega$. Then the function $x \to \nu A_x$ is of the form

$$\nu A_x = \sum_{k=1}^{n} c_k \cdot 1_{C_k}(x),$$

where C_k is a $\sum_0$ subset of 2^ω and c_k is a computable real. In addition, if ν is strongly computable, then the sets $\{a \in \mathbb{Q} |\ c_k < a\}$ and $\{a \in \mathbb{Q} |\ c_k > a\}$ are recursive. (ii) Let ν be a computable measure on 2^ω and suppose that A is a $\sum_1$ subset of $2^\omega \times 2^\omega$. Then the set $\{\langle a,x\rangle \in \mathbb{Q} \times 2^\omega \mid \nu A_x > a\}$ is $\sum_1$. (iii) Let ν be a strongly computable measure on 2^ω. If $A \subseteq 2^\omega \times 2^\omega$ is $\sum_n$, then the set $\{\langle a,x\rangle \in \mathbb{Q} \times 2^\omega |\ \nu A_x > a\}$ is $\sum_n$. If A is $\prod_n$, then $\{\langle a,x\rangle \in \mathbb{Q} \times 2^\omega |\ \nu A_x > a\}$ is $\sum_{n+1}$.

Proof (i) Using if necessary a suitable tiling of A, we may write A as a *disjoint* union

$$A = \bigcup_{i=1}^{m} ([w^i] \times [v^i])$$

such that all w^i have the same length n (hence the $[w^i]$ are either disjoint or identical). Then we have, for all x

$$\nu A_x = \sum_{x(n) = w^i} \nu[v^i].$$

If we define for $k \leq 2^n$, $C_k := [u]$ for the k^{th} word u in 2^n and

$$c_k := \sum_{C_k = w^i} \nu[v^i] \text{ (where } \sum_{\emptyset} \nu[v^i] = 0),$$

then c_k has the required properties and

$$\nu A_x = \sum_{k=1}^{2^n} c_k \cdot 1_{C_k}(x).$$

(ii) Let $A = \{\langle x,y\rangle | \exists n\, R(n,x,y)\}$, where R is a recursive relation. Write $A^m := \{\langle x,y\rangle | \exists n \leq m\, R(n,x,y)\}$, then A^m is Σ_0. We have

$$\{\langle a,x\rangle | \nu A_x > a\} = \{\langle a,x\rangle | \exists m\, (\nu A^m_x > a)\},$$

and the result follows by (i).
(iii) If A is Π_1, then $A = \{\langle x,y\rangle | \forall n\, R(n,x,y)\}$ for some recursive relation R. Put $A^m := \{\langle x,y\rangle | \forall n \leq m\, R(n,x,y)\}$, then A^m is Σ_0 and we may write

$$\{\langle a,x\rangle | \nu A_x > a\} = \{\langle a,x\rangle | \exists \delta \in \mathbb{Q}^+ \forall m\, (\nu A^m_x > a+\delta)\},$$

and for strongly computable measures ν this set is Σ_2, by (i). The result now follows by induction on n. □

4.4.2 Theorem Let μ,ν be strongly computable measures on 2^ω. Suppose that $N \subseteq 2^\omega \times 2^\omega$ is a Π_n μ×ν-nullset. Then $\{x | \nu N_x > 0\}$ is a Σ_{n+1} μ-nullset.

Proof $\{x | \nu N_x > 0\} = \{x | \exists a \in \mathbb{Q}^+ (\nu N_x > a)\}$ is Σ_{n+1} by lemma 4.4.1 and a μ-nullset by Fubini's theorem. □

Theorem 4.4.2 is slightly unsatisfactory, in that one would like to have "Π_n" instead of "Σ_{n+1}" in the conclusion of the theorem. We do not know whether the above estimate is exact. We can show, however, that in general "Σ_{n+1}" cannot be replaced by "Σ_n". Namely, we construct a Π_2 λ×λ-nullset in $2^\omega \times 2^\omega$ such that $\{x | \lambda N_x > 0\}$ is not contained in a Σ_2 λ-nullset. Let M be a total recursive sequential test (with respect to λ) which contains $LLN(\frac{1}{2})^c$ (see section 3.3). Consider $N := \{\langle x,y\rangle | {}^x/_y \in M\}$. By lemma 4.3.3, N is Π_2 and by lemma 4.3.2, $\lambda \times \lambda N = 0$. Suppose $\{x | \lambda N_x > 0\}$ were contained in a Σ_2 set B with $\lambda B = 0$. If $x \in LLN(\frac{1}{2})^c$, then by Theorem 4.2.2, $\{y | {}^x/_y \in LLN(\frac{1}{2})^c\} = 1$; hence $\lambda N_x = 1$ and thus $x \in B$. Therefore $B^c \subseteq$

LLN($\frac{1}{2}$). But this is impossible since LLN($\frac{1}{2}$) is first category while B^c is residual: the first statement is obvious and the second statement follows since B^c is a G_δ set which is dense by $\lambda B^c = 1$.

In what follows, we shall often refer to *computable* real-valued functions on 2^ω, the recursion-theoretic analogue of the *continuous* real-valued functions of constructive analysis (see e.g. Bishop–Bridges [6,38]). We therefore introduce

4.4.3 Definition $f: 2^\omega \to \mathbb{R}$ is *computable* if it is recursively uniformly continuous, i.e. if for some total recursive $h: \omega \to \mathbb{Q}$:

for all n, for all x,y: if $|x - y| < h(n)$, then $|f(x) - f(y)| < 2^{-n}$.

The first part of lemma 4.4.1 implies that if ν is a computable measure and $A \subseteq 2^\omega \times 2^\omega$ is $\sum_0$, then the function $x \to \nu A_x$ is computable.

The effective Fubini theorem for recursive sequential tests can fortunately be obtained easily by formalizing the proof of Theorem 14.1 in Oxtoby [80].

4.4.4 Theorem Let μ,ν be computable measures on 2^ω and suppose that $N \subseteq 2^\omega \times 2^\omega$ is a recursive sequential test with respect to $\mu\times\nu$. Then $\{x|\ \nu N_x > 0\}$ is contained in a recursive sequential test with respect to μ.

Proof Let $N = \bigcap_n O_n \subseteq 2^\omega \times 2^\omega$ be a recursive sequential test with respect to $\mu\times\nu$. Uniformly in n, we construct $\sum_1$ sets $B_n \subseteq 2^\omega$ such that $\mu B_n \leq 2^{-n}$ and $\{x|\ \nu N_x > 0\} \subseteq B_n$. Choose n. Clearly $\mu\times\nu \bigcup_{k>n} O_k \leq 2^{-n}$. $\bigcup_{k>n} O_k$ is of the form $\bigcup_i [w^i]\times[v^i]$ and the sequence $([w^i]\times[v^i])_i$ covers N infinitely often, that is, each $\langle x,y\rangle \in N$ is contained in infinitely many cylinders $[w^i]\times[v^i]$ of the sequence.
Define a sequence of functions f_k, $k \geq 0$, by

$$f_0(x) = 0 \text{ for all } x$$

$$f_k(x) = \sum_{\{i \leq k |\, x \in [w^i]\}} \nu[v^i], \text{ for } k \geq 1.$$

f_k is a computable stepfunction, $f_k: 2^\omega \to [0,1]$, $f_k \leq f_{k+1}$ and

$$f_{k+1}(x) - f_k(x) = \begin{cases} \nu[v^{k+1}] & \text{if } x \in [w^{k+1}] \\ 0 & \text{otherwise.} \end{cases}$$

Clearly

$$\int f_k d\mu = \sum_{i=1}^{k} \int (f_i - f_{i-1}) d\mu = \sum_{i=1}^{k} \mu[w^i] \cdot \nu[v^i] \leq 2^{-n}.$$

Define $B_n := \{x | \exists k\ f_k(x) > 1\}$ (remember that the f_k depend implicitly on n!). Obviously, B_n is Σ_1, uniformly in n. Moreover, $\{x | \nu N_x > 0\} \subseteq B_n$: choose x such that $\nu N_x > 0$, then a fortiori for some y, $\langle x,y \rangle \in N$. Hence for infinitely many i: $\langle x,y \rangle \in [w^i] \times [v^i]$. Let (i') be the sequence of indices for which $x \in [w^{i'}]$. For any $y \in N_x$, for infinitely many i': $y \in [v^{i'}]$. Hence the sequence $([v^{i'}])_{i'}$ covers N_x infinitely often, so $\sum_{i'} \nu[v^{i'}]$ must diverge (otherwise, we could cover N_x with open sets of arbitrarily small ν–measure). It follows that, still for this particular x, $\lim_{k \to \infty} f_k(x) = \infty$ and thus, for some k, $f_k(x) > 1$, i.e. $x \in B_n$.

Clearly then, $\bigcap_n B_n$ is the required recursive sequential test if we can show that $\mu B_n \leq 2^{-n}$. Now if we put $A_m := \{x | \exists k \leq m\ f_k(x) > 1\}$, B_n is the limit of the A_m. Since $f_k \leq f_{k+1}$,

$$\mu A_m = \int 1_{A_m} d\mu < \int f_m d\mu \leq 2^{-n} \text{ for all m,}$$

and so $\mu B_n \leq 2^{-n}$. □

4.4.5 Corollary Let μ,ν be computable measures on 2^ω. Suppose that U is the universal recursive sequential test with respect to $\mu \times \nu$ and that U' is the universal recursive sequential test with respect to μ. Then $U' = \{x | \nu U_x > 0\}$.

Proof By the preceding theorem, $\{x | \nu U_x > 0\} \subseteq U'$. On the other hand, $U' \times 2^\omega \subseteq U$. □

Consequently, if N is a recursive sequential test, $\{x | \nu N_x > 0\}$ need not be contained in a *total* recursive sequential test, since such a test cannot be universal, as we have seen in Chapter 3. This fact necessitates a separate effective Fubini theorem for total recursive sequential tests. The reader not especiallly interested in total recursive sequential tests is free to stop here and may proceed directly to section 4.5.

Our next object is to prove

4.4.6 Theorem Let μ,ν be computable measures on 2^ω. Let $N \subseteq 2^\omega \times 2^\omega$ be a total recursive sequential test with respect to $\mu \times \nu$. Then $\{x | \nu N_x > 0\}$ is contained in a total recursive sequential test with respect to μ.

This theorem can presumably be proved by formalizing proofs of Fubini's theorem from

constructive analysis. However, since we allowed ourselves the use of classical logic and mathematics, a more direct approach is possible. The key of the proof consists in the following observation:

> If $O \subseteq 2^\omega \times 2^\omega$ is a $\sum_1$ set such that $\mu \times \nu O$ is computable and if the image measure π is defined by $\pi[0,s] := \mu\{x \mid \nu O_x \leq s\}$, for $0 \leq s \leq 1$, then the set of points of continuity of π has a $\prod_2$ definition.

Since the set of points of continuity is dense, it follows from an effective version of the Baire Category Theorem, that π has a recursively enumerable dense set of computable points of continuity. From then on, the going is easy.

Our proof strategy is fairly opportunistic: whenever possible, we borrow the requisite algorithms from constructive analysis (e.g. the functions $g(u,v,\cdot)$ defined below, are taken from Bishop and Cheng [7]); but the proofs that these algorithms are in fact total are entirely classical (e.g. lemma 4.4.12).

We now proceed to the proof of Theorem 4.4.6. Write $N = \bigcap_n O^n$, $O^{n+1} \subseteq O^n$, $O^n \in \sum_1$, $\mu \times \nu O^n$ computable (uniformly in n) and $\leq 2^{-n}$. Define on [0,1] the image measure π_n as follows:

$$\pi_n[0,s] := \mu\{x \mid \nu O^n_x \leq s\}, \quad 0 \leq s \leq 1.$$

π_n need not be a computable measure, but nevertheless, as we shall see, some integrals with respect to π_n are computable. We use this fact to compute $\pi_n[0,s]$ for a recursively enumerable dense set of computable reals s.

4.4.7 Definition For $u,v \in [0,1] \cap \mathbb{Q}$, $u<v$, we determine a function $g(u,v,\cdot)$ as follows:

$$g(u,v,t) = \begin{cases} 1 & t<u \\ (v-t)/(v-u) & u \leq t \leq v \\ 0 & v<t. \end{cases}$$

Let $u_0<v_0<u_1<v_1$ be rationals. The functions $f(u_0,v_0,u_1,v_1,\cdot)$ are defined by

$$f(u_0,v_0,u_1,v_1,t) := \min\{1 - g(u_0,v_0,t),\ g(u_1,v_1,t)\}.$$

Before we can motivate the introduction of these auxiliary functions, we need a lemma.

4.4.8 Lemma The integrals

$$\int_{[0,1]} g(u,v,t)d\pi_n(t)\,, \quad \int_{[0,1]} f(u_0,v_0,u_1,v_1,t)d\pi_n(t)$$

are computable uniformly in the parameters n,u,v and n,u_0,v_0,u_1,v_1 respectively.

Proof For this lemma we are indebted to constructive analysis, and in particular to the constructive theory of integration developed in [6], [7] and [9]. Observe that

(i) $\int_{[0,1]} g(u,v,t)d\pi_n(t) = \int_{2^\omega} g(u,v,\nu O^n_x)d\mu(x)$;

(ii) $g(u,v,\nu O^n_x) = \min\left\{1, \frac{(v - \min(\nu O^n_x, v))}{v-u}\right\}$;

(iii) there exists a recursive family of Σ_0 sets $C^{n,k}$ such that each O^n can be written as a *disjoint* union $O^n = \bigcup_k C^{n,k}$. We then have, for all x:

$\nu O^n_x = \sum_{k=1}^{\infty} \nu C^{n,k}_x$ and $\sum_{k=1}^{\infty} \int_{2^\omega} \nu C^{n,k}_x d\mu(x) = \int_{2^\omega} \nu O^n_x d\mu(x) = \mu\times\nu O^n$ is

computable, uniformly in n.

(iv) the function $x \to \nu C^{n,k}_x$ is computable (by lemma 4.4.1) and

$\int_{2^\omega} \nu C^{n,k}_x d\mu(x)$ is computable, both uniformly in n and k.

Call a function h *integrable* (with respect to μ) if there exists a sequence (h_m) of computable functions such that $h = \sum_m h_m$ μ-a.e. and $\sum_m \int h_m d\mu$ is computable (cf. [6,226]). Then the function $x \to \nu O_x$ is integrable (by (iii) and (iv)) and Theorem 2.18 of Bishop-Bridges [6,230] may be translated to our recursion–theoretic setting to show that the operation $\min(\cdot,\cdot)$ preserves integrability. Hence f and g are integrable (by (i) and (ii)). □

Now consider a computable real s and rationals u_0,v_0,u_1,v_1 such that $u_0<v_0<s<u_1<v_1$. Obviously, $\int g(u_0,v_0,t)d\pi_n(t) \le \pi_n[0,s] \le \int g(u_1,v_1,t)d\pi_n(t)$, and by the preceding lemma the terms on the left hand side and on the right hand side are computable. What remains to be done, is to find a computable estimate of the difference

$$\int g(u_1,v_1,t)d\pi_n(t) - \int g(u_0,v_0,t)d\pi_n(t).$$

For certain s, this can be achieved using the functions $f(u_0,v_0,u_1,v_1,t)$.

4.4.9 Definition $s \in [0,1]$ is an *atom* of π_n if $\pi_n\{s\} > 0$. $s \in [0,1]$ is a *point of continuity* of π_n (abbreviated: s is p.c. of π_n) if $\pi_n\{s\} = 0$.

The key of the proof of Theorem 4.4.6 is that the set of p.c.'s of the π_n has a $\prod_2$ definition.

4.4.10 Lemma $s \in [0,1]$ is p.c. of all π_n iff

$$(*)\ \forall n\ \forall \varepsilon{>}0\ \exists \delta{>}0\ \exists u_0, v_0, u_1, v_1\ \left(v_0{<}s{-}\delta{<}s{+}\delta{<}u_1\ \&\ \int_{2^\omega} f(u_0, v_0, u_1, v_1, t)\, d\pi_n(t) < \varepsilon\right),$$

where the quantifiers "$\forall\varepsilon$" and "$\exists\delta$" range over the rationals. Moreover, (*) is a $\prod_2$ statement.

Proof The first statement is obvious as soon as we realize that the condition "$v_0{<}s{-}\delta{<}s{+}\delta{<}u_1$" in (*) means that $f(u_0,v_0,u_1,v_1,t)$ equals 1 on $(s-\delta, s+\delta)$. The second statement follows from lemma 4.4.8. □

The $\prod_2$ definition of the property "s is p.c. of all π_n" enables us to apply the following effective version of the Baire Category Theorem:

4.4.11 Lemma Let G be a dense $\prod_2$ subset of [0,1]. Then G contains a recursively enumerable dense subset of computable reals.

Proof Formalize a proof of the Baire Category Theorem (e.g. Oxtoby [80,2]). □

Combining these lemmas, we get

4.4.12 Lemma There exists a recursively enumerable dense set D of computable points of continuity of all π_n.

Proof By lemma 4.4.10 the set of p.c. of all π_n has a $\prod_2$ definition. This set is dense in [0,1], since the set of s which are an atom for some π_n is countable (this argument is non-constructive). Now apply the preceding lemma. □

We are now almost done.

4.4.13 Lemma Let $s \in [0,1]$ be a computable point of continuity of all π_n. Then $\pi_n[0,s]$ is computable, uniformly in n.

Proof Choose $\varepsilon > 0$. We must effectively determine $u < v < u' < v'$ such that

(1) $\int g(u,v,t)\, d\pi_n(t) \le \pi_n[0,s] \le \int g(u',v',t)\, d\pi_n(t)$

(2) $\int g(u',v',t)d\pi_n(t) - \int g(u,v,t)d\pi_n(t) < \varepsilon$.

Choose recursively enumerable sequences of rationals (b_k), (c_k) such that for all k, $b_k < s < c_k$ and $c_k - b_k < 2^{-k}$. By lemma 4.4.10 there exist (for this particular ε) $\delta > 0$ and rationals $u_0 < v_0 < u_1 < v_1$ such that $v_0 < s-\delta < s+\delta < u_1$ and $\int f(u_0,v_0,u_1,v_1,t)d\pi_n(t) < \varepsilon$. Choose k large enough so that $s - b_k < \delta/4$ and $c_k - s < \delta/4$.
Define $u := b_k - \delta/4$, $v := b_k$, $u' := c_k$ and $v' = c_k + \delta/4$. Then $v_0 < u < v < s < u' < v' < u_1$, hence (1) holds and $\int g(u',v',t)d\pi_n(t) - \int g(u,v,t)d\pi_n(t) \leq \int f(u_0,v_0,u_1,v_1,t)d\pi_n(t) < \varepsilon$. □

Now let D be the set constructed in lemma 4.4.12. Theorem 4.4.6 follows if we can show that

$$\bigcup_{s\in D} \bigcap_n \{x|\ \nu O^n_x > s\}$$

is contained in a total recursive sequential test with respect to μ.
By lemma 4.4.1, for $s \in D$,

$$\{x|\ \nu O^n_x > s\} \in \Sigma_1.$$

Moreover, since

(i) $\mu\{x|\ \nu O^n_x > s\}$ is computable, uniformly in n (by lemma 4.4.13)

(ii) $\mu \bigcap_n \{x|\ \nu O^n_x > s\} = 0$ by Fubini's theorem,

we can determine a recursively enumerable infinite sequence (n_k) of natural numbers such that for all k

$$\mu\{x|\ \nu O^{n_k}_x > s\} < 2^{-k}.$$

Because $O^{n+1} \subseteq O^n$

$$\bigcap_n \{x|\ \nu O^n_x > s\} = \bigcap_k \{x|\ \nu O^{n_k}_x > s\};$$

and

$$\bigcap_k \{x|\ \nu O^{n_k}_x > s\}$$

is a total recursive sequential test with respect to μ. By lemma 3.2.3.8, the union of these tests over D is contained in a total recursive sequential test with respect to μ. But this union equals $\{x|\ \nu N_x > 0\}$. This concludes the proof of Theorem 4.4.6. □

4.5 Proof of the principle of homogeneity Classically, a subset E of 2^ω has *absolute measure zero* if for every finite non-atomic measure μ on 2^ω we can find a Borelset A such that $E \subseteq A$ and $\mu A = 0$. Hausdorff constructed an example of such a set of cardinality $\aleph_1$ (and

this is the best possible result).
This concept can be transferred to the constructive realm as follows: $E \subseteq 2^\omega$ is *recursively small* if for every *computable* finite non-atomic measure μ on 2^ω, we can find a Borelset A such that $E \subseteq A$ and $\mu A = 0$.
Theorems 4.5.2-3 will show that if $x \in R(\mu_p)$ ($R_w(\mu_p)$), then the set $\{y| {}^x/_y \notin R(\mu_p)\}$ ($\{y| {}^x/_y \notin R_w(\mu_p)\}$) is recursively small. (In another sense, these sets are quite large, since they are residual.) For completeness' sake, we begin with the corresponding result for n-randomness.
Strongly computable measures were defined in 3.2.1.1. We say that $p \in (0,1)$ is *strongly computable* if the sets $\{a \in \mathbb{Q}| a > p\}$ and $\{a \in \mathbb{Q}| a < p\}$ are both Δ_1. If $p \in (0,1)$ is strongly computable, then μ_p is a strongly computable measure.

4.5.1 Theorem Let ν be a non-atomic strongly computable measure on 2^ω and let $p \in (0,1)$ be strongly computable. For $n \geq 2$, if x is n–random with respect to μ_p, then $\nu\{y| {}^x/_y$ is not n–random with respect to $\mu_p\} = 0$.

Proof It suffices to show that for each $\prod_n \mu_p \times \nu$-nullset N, $\{x| \nu\{y| {}^x/_y \in N\} > 0\}$ is contained in a $\sum_{n+1} \mu_p$-nullset. By lemma 4.3.4, $\{\langle x,y\rangle| {}^x/_y \in N\}$ is a $\prod_n \mu_p \times \nu$-nullset. Since μ_p is strongly computable, we may now apply Theorem 4.4.2. □

4.5.2 Theorem Let $p \in (0,1)$ be computable. If $x \in R(\mu_p)$, then $\{y| {}^x/_y \notin R(\mu_p)\}$ is recursively small.

Proof Let ν be a non-atomic computable measure. Since $R(\mu_p)^c$ is a recursive sequential test with respect to μ_p, lemma 4.3.4 implies that $\{\langle x,y\rangle| {}^x/_y \notin R(\mu_p)\}$ is a recursive sequential test with respect to $\mu_p \times \nu$. Now apply Theorem 4.4.4. □

4.5.3 Theorem Let $p \in (0,1)$ be computable. If $x \in R_w(\mu_p)$, then $\{y| {}^x/_y \notin R_w(\mu_p)\}$ is recursively small.

Proof Let N be a total recursive sequential test with respect to μ_p. Let ν be a non-atomic computable measure. By lemma 4.3.4, $\{\langle x,y\rangle| {}^x/_y \in N\}$ is a total recursive sequential test with respect to $\mu_p \times \nu$. Now apply Theorem 4.4.6. □

4.5.4 Remarks (i) The principle of homogeneity thus holds true for a wide class of definitions of randomness based on probabilistic laws, although we needed three different proofs to show this. The common core of these proofs is that the operation / is measure-preserving and also preserves arithmetical structure; the differences result from the fact that the Fubini-property needs a separate verification in each case.

(ii) Looking back on what we have accomplished, we see that, at least in a quantitative sense, von Mises' intuitions can be salvaged: if we provisionally identify Kollektivs with random sequences (in Martin–Löf's sense), then the set of subsequences of a Kollektiv which are not themselves Kollektivs is exceedingly small. Alternatively, we might say that Martin-Löf's definition and its variants capture at least some of von Mises' intentions. Observe that, from von Mises' point of view, the preceding theorems should not be interpreted as a result on the extremely small *probability* of the set $\{y|\ {}^x/y \notin R(\mu_p)\}$.

(iii) If we compare Theorem 4.2.2 with the preceding theorems, we see that the latter state a necessary condition for randomness, whereas the first is a necessary and sufficient condition for satisfying the law of large numbers. We doubt whether the preceding theorems admit a converse. Perhaps there is a converse if we strengthen the consequences using compositions of recursive selections and random selections, in the following sense.

If Φ is a recursive place selection (with generating function ϕ as in definition 2.5.1.1) such that $\mu_p dom\Phi = 1$, define $/^\Phi$ by

$({}^x/^\Phi y)_k = x_m$ if m is the index of the k^{th} 1 in y and $\phi(x(m-1)) = 1$.

Since $/^\Phi$ satisfies lemmas 4.3.1-3, the preceding theorems hold with / replaced by $/^\Phi$.

Various other theorems on the operation / can be derived along these lines, the most interesting of which is perhaps the following. Let ${}^x/^\circ y$ be defined as ${}^x/y$, except that we now look at the zeros of y. Hence, when viewed as sets of natural numbers, ${}^x/y \cup {}^x/^\circ y = \mathbb{N}$.

4.5.4 Theorem Let $p \in (0,1)$ be computable. If $x \in R(\mu_p)$, then the set $\{y| <{}^x/y, {}^x/^\circ y> \notin R(\mu_p \times \mu_p)\}$ is recursively small.

Proof Let ν be a computable non-atomic measure. We show first that $\mu_p \times \nu\{<x,y>| <{}^x/y, {}^x/^\circ y> \in A \times B\} = \mu_p A \cdot \mu_p B$. As in lemma 4.3.2, it suffices to show that for fixed y, $\mu_p\{x|\ {}^x/y \in A, {}^x/^\circ y \in B\} = \mu_p A \cdot \mu_p B$. We need only verify this equality for A = [w], B= [v]. But $\{x|\ {}^x/y \in [w], {}^x/^\circ y \in [v]\} = [u]$, where $|u| = |w| + |v|$ and u consists of w and v intertwined. Hence $\mu_p[u] = \mu_p[w] \cdot \mu_p[v]$. From here on, the argument is entirely similar to the arguments above. □

To interpret this theorem, recall that we defined two Kollektivs z^0, z^1 to be *independent* if the pair $<z^0, z^1>$ is a Kollektiv with respect to the product distribution (cf. 2.4.1). Having formalized Kollektivs as random sequences, it seems reasonable to formalize a pair of independent Kollektivs as an element of $R(\mu_p \times \mu_p)$ (such pairs are invariant under recursive place selections, they satisfy the law of the iterated logarithm etc.). We saw in 2.4.1 that a *lawlike* partition of a Kollektiv into two (or more) Kollektivs yields provably independent Kollektivs and we remarked that this feature reflects the assumed independence of successive

tosses. We now see that also in this context a principle of homogeneity obtains: "almost every" partition, whether lawlike or not, produces independent Kollektivs.

4.6 New proof of a theorem of Ville In 2.6.2.2, we stated Ville's theorem as follows:

Given a countable set $\mathcal{H}$ of place selections Φ: $2^\omega \to 2^\omega$ we can construct $x \in 2^\omega$ such that

(i) $x \in \mathrm{dom}\Phi$ implies $\Phi x \in \mathrm{LLN}(\frac{1}{2})$, for all $\Phi \in \mathcal{H}$

(ii) for all n $\frac{1}{n}\sum_{k=1}^{n} x_k \geq \frac{1}{2}$.

C(p), the set of Church-random sequences with parameter p, was defined in 2.5.1.7. Since property (ii) contradicts the law of the iterated logarithm and all (weakly) random sequences satisfy the law of the iterated logarithm (as was shown in section 3.3), we have as a consequence $C(\frac{1}{2}) \cap R(\lambda)^c \neq \emptyset$ (although of course $R(\lambda) \subseteq C(\frac{1}{2})$). Thus $C(\frac{1}{2})$ and $R(\lambda)$, which have very different philosophical justifications, differ also extensionally.
We need not repeat here the discussion on the philosophical significance of Ville's theorem given in 2.6.2.2; in the present section we are concerned only with its proof. Ville's argument [99,55-69] has a combinatorial character and consists roughly speaking in replacing $\mathcal{H}$ by a different set $\mathcal{H}'$ of place selections Ψ such that if $\Psi, \Psi' \in \mathcal{H}'$, then Ψ and Ψ' are "disjoint". This notion of disjointness is best illustrated by means of an example. Let (p_n) be an enumeration of the prime numbers and let Ψ_n be the place selection that chooses all indices which are a power of p_n. Then no two Ψ_n choose the same index and in this case it is very easy to construct an x which satisfies specifications (i) and (ii). By adroitly manipulating place selections, Ville is able to reduce the general case to something very like the above example.

Without denying the ingenuity of Ville's construction, it seems worthwhile to try to derive the theorem from first principles, that is, as an expression of the philosophical differences between strict frequentism and the propensity interpretation uncovered in Chapter 2. In other words, we want to show that the different interpretations of probability underlying the definitions of Church-random sequences and (Martin-Löf) random sequences, namely probability as relative frequency and coordinate-wise probability respectively, *themselves* imply that $C(\frac{1}{2}) \cap R(\lambda)^c \neq \emptyset$.

In the introduction to Chapter 3 we observed that, from the point of view of strict frequentism, the distribution (1–p,p) on {0,1} should not be associated with a unique measure on 2^ω, to wit, μ_p, but rather with a whole class of measures, namely all those which in a certain sense determine the same limiting relative frequencies 1–p and p. Existence theorems should not be affected when we replace one measure from this class by another. We may therefore state

Conjecture 1 Let $\pi = \prod_n(1-p_n,p_n)$ be a product measure such that $\lim_{n\to\infty} p_n = p$ and let $\mathcal{H}$ be a countable set of place selections. Then not only $\mu_p C(\mathcal{H},p) = 1$, as was shown in Theorem 2.5.2.3, but also $\pi C(\mathcal{H},p) = 1$.

On the other hand, the definition of (weakly) random sequences at first sight seems to involve a unique measure, namely μ_p. This impression is confirmed by the discussion of Martingales in 3.4, where it was seen that their definition seemed to require (constant) probabilities at individual coordinates. One way to state this seeming dependence upon the underlying measure is as follows:

Conjecture 2 If $\pi = \prod_n(1-p_n,p_n)$ with $\lim_{n\to\infty} p_n = p$ but $p_n \neq p$ for all n, then $\pi R(\mu_p) = 0$. By the 0–1 law, $\pi R(\mu_p)$ is either one or zero and the first case seems to be excluded by the above argument.

Both conjectures, taken together, would give us the required proof of Ville's theorem from first principles, for if π satisfies the hypothesis of Conjecture 2, we would have $\pi(C(p)\cap R(\mu_p)^c) = 1$. But, although Conjecture 1 can indeed be proven (see corollary 4.6.3), Conjecture 2 is false. First impressions notwithstanding, $R(\mu_p)$ is not *that* sensitive to the choice of the underlying measure: there exist $\pi = \prod_n(1-p_n,p_n)$ such that $\lim_{k\to\infty} p_n = p$ and $p_n \neq p$ for all n, for which $\pi R(\mu_p) = 1$.

On the other hand, the idea that the extensional difference between C(p) and $R(\mu_p)$ is due to a difference in sensitivity to the choice of the measure is correct, but it should be formulated more carefully. Although for a computable product measure $\pi = \prod_n(1-p_n,p_n)$, $\pi R(\mu_p) = 1$ does not imply that $p_n = p$ for all n, it *does* imply that $\sum_n(p-p_n)^2 < \infty$, in other words, that p_n converges to p rather *fast*.We then get a proof of Ville's theorem if we take a computable product measure π for which the marginals p_n converge *slowly* to p, for in that case $\pi(C(p)\cap R(\mu_p)^c) = 1$ (Theorem 4.6.1).

The result we derive in this way differs from Ville's original formulation in two minor respects:

– not only is $C(p)\cap R(\mu_p)^c$ non-empty, it has the cardinality of the continuum;

– on the other hand, the proof does not yield that for *every* π such that $\pi(C(p)\cap R(\mu_p)^c) = 1$, already $\pi(C(p)\cap LIL(\mu_p)^c) = 1$, where $LIL(\mu_p)$ is the set of sequences which satisfy the law of the iterated logarithm for μ_p. Indeed, the proof *cannot* yield such a result, since it is false for some π with $\pi(C(p)\cap R(\mu_p)^c) = 1$. But for some very slowly converging π, we do have that $\pi(C(p)\cap LIL(\mu_p)^c) = 1$, thus strengthening Ville's theorem in its original formulation.

The reader may wonder why we persistently formulate these results for C(p) instead of for C($\mathcal{H}$,p), for arbitrary countable sets $\mathcal{H}$ of place selections. The answer is that R(μ_p), due to its recursion theoretic structure can only be reasonably compared with C(p). For very slowly converging π, however, we have, for arbitrary $\mathcal{H}$, $\pi(C(\mathcal{H},p) \cap LIL(\mu_p)^c) = 1$.

This section is organized as follows. We first prove Ville's theorem along the lines sketched above (Theorem 4.6.1) and we comment on the significance of the proof (Corollary 4.6.6 and following discussion). The reader may then proceed to Chapter 5; the rest of the section generalizes Corollary 4.6.6 to measures which are not product measures and is not essential to the main argument.

We prove Ville's theorem in the following form:

4.6.1 Theorem Let $p \in (0,1)$ be a computable real. There exists a non-atomic computable measure π such that $\pi(C(p) \cap R_w(\mu_p)^c) = 1$. A fortiori, $\pi(C(p) \cap R(\mu_p)^c) = 1$ and $C(p) \cap R_w(\mu_p)^c$ has the cardinality of the continuum.

The measure will be a computable product measure $\pi = \prod_n(1-p_n,p_n)$ such that $\lim_{k\to\infty} p_n = p$

and $\pi \perp \mu_p$. In fact, the proof will show that for *any* such measure π, $\pi(C(p) \cap R_w(\mu_p)^c) = 1$.

4.6.2 Lemma Let $\pi = \prod_n(1-p_n,p_n)$ be a computable product measure. Then $\pi C(p) = 1$ iff $\lim_{n\to\infty} p_n = p$.

Proof $\Rightarrow$ Suppose not. Then for some rational $\varepsilon > 0$, at least one of the sets $\{n|\ p_n > p+\varepsilon\}$, $\{n|\ p_n < p-\varepsilon\}$ is infinite, say the first set. By the computability of π this set is recursively enumerable, hence contains an infinite recursive subset. Using this subset, we can define a recursive place selection Φ such that $\pi\Phi^{-1}(LLN(p)) = 0$, a contradiction.
$\Leftarrow$ For this direction, no assumption of computability or recursiveness is needed. So let Φ be a place selection and π a measure of the form $\pi = \prod_n(1-p_n,p_n)$ such that $\lim_{n\to\infty} p_n = p$

and assume that $p_n \neq 0$ for all n (which is no essential restriction). We show that $\pi(dom\Phi) = \pi(dom\Phi \cap \Phi^{-1}(LLN(p)))$. Given Φ and its generating function ϕ (as in definition 2.5.1.1), we define a partial function θ: $2^\omega \times \omega \to \omega$ as follows:

(1) $\mathrm{dom}\theta = \mathrm{dom}\Phi$

(2) if $x \in \mathrm{dom}\Phi$, then $\theta(x,n) = 1 + \min\{k \mid n = \sum_{j=1}^{k}\phi(x(j))\}$.

Assume first that $\pi(\mathrm{dom}\Phi) = 1$ Define random variables Z_n: $2^\omega \to \mathbb{R}$ by

$$Z_n(x) = \frac{x_{\theta(x,n)}}{p_{\theta(x,n)}} \text{ for } x \in \mathrm{dom}\Phi \text{ and } Z_n(x) = 1 \text{ otherwise.}$$

Let B_n denote the algebra generated by the cylinders of length n. Let $\mathbb{E}_\pi$ denote the expectation with respect to π and $\mathbb{E}_\pi(\ldots|B_n)$ the conditional expectation with respect to π and B_n. We then have

(3) $\mathbb{E}_\pi(Z_n) = 1$ for all n: $\mathbb{E}_\pi(Z_n) = \sum_{k=n}^{\infty} \int_{\{x|\theta(x,n) = k\}} Z_n d\pi =$

$$= \sum_{k=n}^{\infty} p_k^{-1}\cdot\pi\{x|\ \theta(x,n) = k \ \&\ x_k = 1\} = \sum_{k=n}^{\infty} p_k^{-1}\cdot p_k\cdot\pi\{x|\ \theta(x,n) = k\} =$$

$$= \sum_{k=n}^{\infty}\pi\{x|\ \theta(x,n) = k\} = 1.$$

The third equality is a consequence of the fact that $\{x|\ \theta(x,n) = k\} \in B_{k-1}$ and $\{x|\ x_k = 1\} \in B_k$, so that these events are independent with respect to π. The last equality follows from the assumption that $\pi(\mathrm{dom}\Phi) = 1$.

(4) $\mathbb{E}_\pi(Z_n|B_{n-1})(x) = 1$ for all x: by definition, $\mathbb{E}_\pi(Z_n|B_{n-1})$ is B_{n-1}–measurable and satisfies for $B \in B_{n-1}$ $\int_B Z_n d\pi = \int_B \mathbb{E}_\pi(Z_n|B_{n-1})d\pi$.

Now $\int_B Z_n d\pi = \sum_{k=n}^{\infty} \int_{\{x|\theta(x,n) = k\}\cap B} x_k\cdot p_k^{-1} d\pi = \sum_{k=n}^{\infty} p_k^{-1}\cdot\pi\{x|\ \theta(x,n) = k \ \&\ x_k = 1\}\cap B =$

$= \sum_{k=n}^{\infty} p_k^{-1}\cdot p_k\cdot\pi\{x|\ \theta(x,n) = k\}\cap B = \pi B$. Since $\mathbb{E}_\pi(Z_n|B_{n-1})$ is constant on cylinders of length n–1, this implies that $\mathbb{E}_\pi(Z_n|B_{n-1})$ equals 1 everywhere.

(5) Since $\lim_{n\to\infty} p_n = p \in (0,1)$, there is $\delta \in (0,1)$ and $n_0 \in \mathbb{N}$ such that for $n \geq n_0$: $\delta < p_n < 1-\delta$. Then, again for $n \geq n_0$: $0 \leq Z_n \leq p_n^{-1} < \delta^{-1}$, hence the Z_n are uniformly bounded. By Theorem 3 in Feller [26,243]:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}Z_k(x) \;=\; \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{x_{\theta(x,k)}}{p_{\theta(x,k)}} \;=\; \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{\Phi(x)_k}{p_{\theta(x,k)}} \;=\; 1 \;\; \pi\text{–a.e.}$$

But, generally, if (a_n) and (b_n) are sequences of positive reals,

$$\text{if } \lim_{n\to\infty} b_n = p \text{ and } \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{a_k}{b_k} = 1, \text{ then } \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}a_k = p.$$

Hence, still under the assumption $\pi(\mathrm{dom}\Phi) = 1$:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\Phi(x)_k \;=\; p.$$

We now drop the assumption. Note that $\pi(\mathrm{dom}\Phi\cap\Phi^{-1}(\mathrm{LLN}(p))) = \pi(\mathrm{dom}\Phi)$ is equivalent to $\pi(\Phi^{-1}\mathrm{LLN}(p)|\mathrm{dom}\Phi) = 1$ (under the assumption $\pi(\mathrm{dom}\Phi) > 0$, but otherwise there is nothing to prove), so for the general case it suffices to replace in the above proof π by $\pi(\ldots|\mathrm{dom}\Phi)$. □

4.6.3 Corollary Let $\mathcal{H}$ be a countable set of place selections and $\pi = \prod_n(1–p_n,p_n)$ a product measure such that $\lim_{n\to\infty} p_n = p$. Then $\pi C(\mathcal{H},p) = 1$.

We next investigate the sensitivity of $R(\mu_p)$ to the underlying measure.

4.6.4 Lemma Let μ,ν be computable measures on 2^ω. $\mu\perp\nu$ is equivalent to either of the following statements: (i) there exists a total recursive sequential test N with respect to μ such that $\nu N = 1$; (ii) for each rational $\varepsilon > 0$, there exists a $\prod_1$ set A such that $\nu A > 1–\varepsilon$ and $\mu A = 0$.

Proof Trivially, (i) and (ii) imply $\mu\perp\nu$. For $\mu\perp\nu$ implies (i) we use the following equivalence

$$\mu\perp\nu \;\text{ iff }\; \forall\varepsilon{>}0\;\exists C{\in}\textstyle\sum_0\,(\nu C > 1–\varepsilon \;\&\; \mu C < \varepsilon)$$

and we take advantage of the $\prod_2$ statement on the right hand side. Let $f: \mathbb{Q}^+ \to\sum_0$ be a total recursive function which for each ε in $\mathbb{Q}^+$ gives $f(\varepsilon)$ in $\sum_0$ such that $\nu f(\varepsilon) > 1–\varepsilon$ and $\mu f(\varepsilon) < \varepsilon$. Such a function exists by the computability of μ and ν. Let $N = \bigcap_n\bigcup_i f(2^{-i-n-1})$. Obviously N is $\prod_2$. Since for each n and i, $\mu f(2^{-i-n-1}) < 2^{-i-n-1}$, $\mu\bigcup_i f(2^{-i-n-1})$ is computable (see the proof of the first effective Borel–Cantelli lemma (3.3.1)). Hence N is a total recursive sequential test with respect to μ. On the other hand, for each n and all i, $\nu\bigcup_i f(2^{-i-n-1}) \geq \nu f(2^{-i-n-1}) \geq 1–2^{-i-n-1}$, so $\nu\bigcup_i f(2^{-i-n-1}) = 1$. For (i) implies (ii), reverse the roles of μ and ν in (i), obtaining N =

$\bigcap_n O_n$ such that $\mu N = 1$, $\nu N = 0$ and each O_n in Σ_1; then some $(O_n)^c$ will do. □

The following beautiful criterion for singularity of product measures is due to Kakutani [39][1].

4.6.5 Lemma Let $\mu = \prod_n(1-p_n,p_n)$, $\pi = \prod_n(1-q_n,q_n)$ be product measures on 2^ω such that for some $\delta > 0$ and all n, $\delta < p_n,q_n < 1-\delta$. If $\sum_n(p_n - q_n)^2$ diverges, then μ and π are mutually singular; on the other hand, if $\sum_n(p_n - q_n)^2$ converges, then μ and π are equivalent.

It follows from the zero-one law that product measures on 2^ω are either singular or equivalent, but Kakutani's theorem provides us with a criterion to distinguish these cases and this is what we shall use to finish the proof of Theorem 4.6.1.

Let $p_n := p\cdot(1 + (n+1)^{-\frac{1}{2}})$, $\pi = \prod_n(1-p_n,p_n)$, then π is computable and since $\sum_n(p - p_n)^2 = \sum_n p^2\cdot n^{-1} = \infty$, $\pi \perp \mu_p$. By corollary 4.6.3, $\pi C(p) = 1$. By lemma 4.6.4, $\pi R(\mu_p) = 0$. This completes the proof of Theorem 4.6.1. □

We may extract the following information from the proof of Theorem 4.6.1:

4.6.6 Corollary Let $\pi = \prod_n(1-p_n,p_n)$ be a computable product measure, $p \in (0,1)$ a computable real.

(i) $\pi C(p) = 1$ iff $\lim_{n\to\infty} p_n = p$

(ii) $\pi R(\mu_p) = 1$ iff $\sum_n(p - p_n)^2$ converges.

(iii) $\pi(C(p)\cap R(\mu_p)^c) = 1$ iff $\lim_{n\to\infty} p_n = p$ but $\sum_n(p - p_n)^2$ diverges.

4.6.7 Remark We saw in 2.6.2.2 that there exist countably many recursive place selections Φ such that Kollektivs of Ville's type can never belong to the domain of Φ. But if $x \notin \mathrm{dom}\Phi$, then the statement "$x \in \mathrm{dom}\Phi$ implies $\Phi x \in LLN(p)$" is uninformative. (A failure is significant only when preceded by a serious effort.) Similarly, although we have formally proved that $\pi(C(p)\cap R(\mu_p)^c) = 1$ if π satisfies the right hand side of (iii), the theorem and its corollary are interesting only for a *subclass* of the recursive place selections, namely for those Φ for which $\pi(\mathrm{dom}\Phi) = 1$ if π is a product measure whose marginals converge to p.

The reader will have noticed undoubtedly that Ville's theorem in its original formulation uses the law of the iterated logarithm essentially, whereas it is absent from our proof. This leads to the following question: is the difference between C(p) and $R(\mu_p)$ due *entirely* to the law of the iterated logarithm, in the sense that each sequence in $C(p)\cap R(\mu_p)^c$ fails to satisfy it?

Interestingly, it is a corollary of Theorem 4.6.1 that this is not so: if π is, e.g., the product measure $\prod_n(1-p_n,p_n)$ with $p_n = p\cdot(1 + (n+1)^{-\frac{1}{2}})$, then π assigns measure one to the set of sequences which satisfy the law of the iterated logarithm (for μ_p). To see this, we need a general form of the

Law of the iterated logarithm (Kolmogorov [45])

Let $\mu = \prod_n(1 - q_n,q_n)$ be a product measure and define the variance s_n by $s_n := \sum_{k=1}^{n} q_n\cdot(1 - q_n)$.

Then

$$(1)\quad \text{for } \beta > 1, \text{ for } \mu\text{–a.a. } x:\ \exists m\ \forall n\geq m\ \Big|\sum_{k=1}^{n} x_k - \sum_{k=1}^{n} n\cdot q_k\Big| < \beta\sqrt{2s_n\text{loglogs}_n}$$

$$(2)\quad \text{for } \beta < 1,\ \text{for } \mu\text{–a.a. } x:\ \forall m\ \exists n\geq m\ \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} n\cdot q_k > \beta\sqrt{2s_n\text{loglogs}_n}$$

$$\text{for } \beta < 1,\ \text{for } \mu\text{–a.a. } x:\ \forall m\ \exists n\geq m\ \sum_{k=1}^{n} n\cdot q_k - \sum_{k=1}^{n} x_k > \beta\sqrt{2s_n\text{loglogs}_n}.$$

If all q_n are equal to p, we get back the form of the law stated in 2.6.2.2. Let $\text{LIL}(\mu_p)$ denote the set of sequences which satisfy the law for the measure μ_p. Let π be the product measure constructed above. We show that $\pi\text{LIL}(\mu_p) = 1$.

If for instance for some $\alpha < 1$,

$$\pi\{x|\ \exists m\ \forall n\geq m\ \sum_{k=1}^{n} x_k > p\cdot n\ -\ \alpha\sqrt{2n\cdot p\cdot(1-p)\text{loglogn}}\ \} = 1,$$

so that $\pi\text{LIL}(\mu_p) = 0$, then, by the general form of the law of the iterated logarithm, for $\beta > 1$ and n sufficiently large:

$$p\cdot n + p\cdot\sum_{k=1}^{n}\frac{1}{\sqrt{k+1}}\ -\ \beta\sqrt{2s_n\text{loglogs}_n}\ >\ p\cdot n -\ \alpha\sqrt{2n\cdot p\cdot(1-p)\text{loglogn}},$$

hence

$$p\cdot\sum_{k=1}^{n}\frac{1}{\sqrt{k+1}}\ >\ \beta\sqrt{2s_n\text{loglogs}_n}\ -\ \alpha\sqrt{2n\cdot p\cdot(1-p)\text{loglogn}};$$

but this is easily seen to be false, since the left hand side is $O(\sqrt{n})$, whereas the right hand side is $O(\sqrt{(n\text{loglogn})})$. An analogous argument for the upper bound then shows that $\pi\text{LIL}(\mu_p) = 1$.

On the other hand, it is possible to construct uncountably many Church-random sequences (with parameter p) which do not satisfy the law of the iterated logarithm (for μ_p) if we use product measures μ_p whose marginals converge to p slower than those of π. Choose a such that $-\frac{1}{2} < a < 0$ and put $q_n := p\cdot(1 + (n+1)^a)$, $\mu := \prod_n(1-q_n,q_n)$.
We now do have, for $\alpha < 1$,

$$\mu\{x|\ \exists m\ \forall n\geq m \sum_{k=1}^{n} x_k > p\cdot n - \alpha\sqrt{2n\cdot p\cdot(1-p)\log\log n}\ \} = 1;$$

by the general form of the law of the iterated logarithm, it suffices to show that for some $\beta > 1$ and all n sufficiently large:

$$p\cdot n + p\cdot\sum_{k=1}^{n}(k+1)^a - \beta\sqrt{2s_n\log\log s_n} > p\cdot n - \alpha\sqrt{2n\cdot p\cdot(1-p)\log\log n};$$

in other words, that

$$p\cdot\sum_{k=1}^{n}(k+1)^a > \beta\sqrt{2s_n\log\log s_n} - \alpha\sqrt{2n\cdot p\cdot(1-p)\log\log n}.$$

But now the left hand side is $O(n^{a+1})$, with $a+1 > \frac{1}{2}$ and the right hand side is still $O(\sqrt{(n\log\log n)})$. Hence not only $\mu(C(p)\cap R(\mu_p)^c) = 1$ (since $\mu\perp\mu_p$), but also $\mu(C(p)\cap LIL(\mu_p)^c) = 1$.

We may thus conclude that a part of, but *only* a part of, the difference between C(p) and R(μp) is caused by the law of the iterated logarithm. The proof of Theorem 4.6.1 shows that Church–random sequences may also fail to satisfy properties which are essentially different from the law of the iterated logarithm.

The rest of this section is rather technical: we investigate what remains of Corollary 4.6.6 if we drop the assumption that π be a product measure. We now obtain a theorem which connects the different concepts of randomness with different types of convergence of measures.

4.6.8 Definition Let μ and ν be measures on 2^ω and let $T: 2^\omega \to 2^\omega$ be the left shift. We say that the sequence of measures $(\mu T^{-n})_{n\in\mathbb{N}}$ *converges strongly* to ν if for all Borel sets A, $\lim_{n\to\infty} \mu T^{-n}A = \nu A$. We say that $(\mu T^{-n})_{n\in\mathbb{N}}$ *converges weakly* to ν if for all Borel sets A such that $\nu\partial A = 0$ (where ∂A is the boundary of A), $\lim_{n\to\infty} \mu T^{-n}A = \nu A$.

The next lemma considerably simplifies the last condition:

4.6.9 Lemma (See Billingsley [4].) $(\mu T^{-n})_{n\in\mathbb{N}}$ converges weakly to ν if for all cylinders [w]: $\lim_{n\to\infty} \mu T^{-n}[w] = \nu[w]$.

Part (i) of Corollary 4.6.6 can now be restated thus: $\pi C(p) = 1$ iff $(\pi T^{-n})_{n\in\mathbb{N}}$ converges weakly to μ_p. We shall see presently that one half of this result can be salvaged even without the assumption that π be a product measure.

4.6.10 Theorem Let μ be a measure such that for all place selections Φ recursive in μ, if $\mu(\mathrm{dom}\Phi) = 1$, then $\mu(\Phi^{-1}\mathrm{LLN}(p)) = 1$. Then $(\mu T^{-n})_{n\in\mathbb{N}}$ converges weakly to μ_p. In particular, if μ is computable and $\mu C(p) = 1$, then $(\mu T^{-n})_{n\in\mathbb{N}}$ converges weakly to μ_p.

Proof Suppose not; then there exists a smallest binary string s such that $\lim_{n\to\infty} \mu T^{-n}[s] \neq \mu_p[s]$. Without loss of generality we may suppose that for some rational $\varepsilon > 0$, for some sequence (N_i) recursive in μ and for all i:

$$\mu T^{-N_i}[s] > \mu_p[s] + \varepsilon.$$

Define *for this particular sequence* (N_i) and for *all* binary words v a place selection Ψ_v by

$$\psi_v(x(m)) = \begin{cases} 1 & \text{if } \exists i\,(N_i + |v| = m) \;\&\; \exists u\in 2^{<\omega}(uv = x(m)) \\ 0 & \text{otherwise.} \end{cases}$$

Recall that "$v\subset w$" means that v is a strict initial segment of w and that $\langle\rangle$ denotes the empty string.

Claim 1

$$\forall w\in 2^{<\omega}\,(\forall v\subset w \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\Psi_v(x)_k = p \longrightarrow \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[w]}(T^{N_i}x) = \mu_p[w]).$$

Proof of claim 1 We use induction on w. If $w = 1$, the hypothesis of the claim implies that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\Psi_{\langle\rangle}(x)_k = p,$$

which is by definition of $\Psi_{\langle\rangle}$ equivalent to

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[1]}(T^{N_i}x) = p.$$

Suppose the claim holds for w. Note that

$$\frac{\frac{1}{n}\sum_{i=1}^{n} 1_{[w1]}(T^{N_i}x)}{\frac{1}{n}\sum_{i=1}^{n} 1_{[w]}(T^{N_i}x)} = \frac{\sum_{k=1}^{|\Psi_w(x(N_n))|} \Psi_w(x)_k}{|\Psi_w(x(N_n))|}.$$

The hypothesis of the claim implies that the right hand side converges to p; the hypothesis of induction implies that the denominator of the left hand side converges to $\mu_p[w]$. It follows that the numerator of the left hand side must converge to $\mu_p[w1]$. This concludes the proof of claim 1.

Claim 2 Under the hypothesis of the theorem, for $\alpha = 0,1$:

$$\forall v \in 2^{<\omega}\, \mu(\mathrm{dom}\Psi_v) = 1 \;\&\; \mu\{x \mid \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[v\alpha]}(T^{N_i}x) = \mu_p[v\alpha]\} = 1.$$

Proof of claim 2 We use induction on v. Trivially, $\mu(\mathrm{dom}\Psi_{\langle\rangle}) = 1$ and hence for $\alpha = 0,1$:

$$\mu\{x \mid \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[\alpha]}(T^{N_i}x) = \mu_p[\alpha]\} = 1,$$

by claim 1 and the hypothesis of the theorem. Suppose the claim holds for $u \subset v$, then again by claim 1 and the hypothesis of the theorem:

$$\mu\{x \mid \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[v]}(T^{N_i}x) = \mu_p[v]\} = 1.$$

It follows that μ–a.e. v occurs infinitely often at coordinates starting with an index N_i+1; hence $\mu(\mathrm{dom}\Psi_v) = 1$. Then, as a consequence of claim 1 and the hypothesis of the theorem:

$$\mu\{x \mid \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[v\alpha]}(T^{N_i}x) = \mu_p[v\alpha]\} = 1.$$

This concludes the proof of claim 2.

Claim 2 implies that for the particular string s determined at the outset,

$$\mu\{x \mid \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{[s]}(T^{N_i}x) = \mu_p[s]\} = 1.$$

By the dominated convergence theorem,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mu T^{-N_i}[s] \;=\; \lim_{n\to\infty}\int_{2^\omega}\frac{1}{n}\sum_{i=1}^{n}1_{[s]}(T^{N_i}x)d\mu(x) \;=$$
$$=\; \int_{2^\omega}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}1_{[s]}(T^{N_i}x)d\mu(x) \;=\; \mu_p[s],$$

a contradiction. □

A converse to the theorem is not to be expected. Indeed, the conclusion of the theorem is probably too weak; it is plausible that the hypothesis implies some kind of asymptotic independence condition.

We next generalize the second part of Corollary 4.6.6 to arbitrary computable measures.

4.6.11 Lemma Let μ, ν be computable measures such that μ is not absolutely continuous with respect to ν. Then for some total recursive sequential test N with respect to ν, $\mu N > 0$.

Proof We showed in Example 3.4.6 that one can define a recursive sequential test N with respect to ν such that $\mu N > 0$, using the likelihood ratio $\mu[w]/\nu[w]$. For reasons explained at length in 3.4, it is difficult, if not impossible, to prove that N is a *total* recursive sequential test with respect to ν. We therefore borrow an idea of Gaifman and Snir [34,518]. Choose $\varepsilon > 0$. Since μ is not absolutely continuous with respect to ν, there exists a sequence (C_i) of $\sum_0$ sets such that $\mu\bigcap_i C_i > \varepsilon$ and $\nu\bigcap_i C_i = 0$. Let (D_k) be a recursive enumeration of the $\sum_0$ sets. Define

$$f(n) := \min\{k > n \mid \mu D_k > \varepsilon \;\&\; \nu D_k < 2^{-k}\}.$$

Let $N = \bigcap_n \bigcup_{m\geq n} D_{f(m)}$, then $\mu N > \varepsilon$. That N is a total recursive sequential test is shown by an argument similar to the proof of the effective first Borel–Cantelli lemma, 3.3.1. □

Gaifman [34,519] asks whether μ and ν can already be separated by a $\prod_1$ set. An affirmative answer would follow from lemma 4.6.4 in the unlikely event that the Lebesgue decomposition of μ with respect to ν, namely $\mu = \mu_0 + \mu_1$, where $\mu_0 << \nu$ and $\mu_1 \perp \nu$, can be achieved with computable μ_0, μ_1. It is more probable, however, that one can produce a counterexample to computable Lebesgue decomposition in this way.

4.6.11 Lemma μ is absolutely continuous with respect to μ_p iff $(\mu T^{-n})_{n\in\mathbb{N}}$ converges strongly to μ_p.

Proof $\Rightarrow$ T is *strongly mixing* with respect to μ_p, i.e. for f,g in $L^1(\mu_p)$, $\lim_{n\to\infty}\int(f\circ T^n)\cdot g d\mu_p$ $= (\int f d\mu_p)\cdot(\int g d\mu_p)$. Let g in $L^1(\mu_p)$ be a Radon–Nikodym derivative of μ with respect to μ_p, then for all Borel sets A: $\mu A = \int g\cdot 1_A d\mu_p$. Hence $\lim_{n\to\infty} \mu T^{-n}A = \lim_{n\to\infty} \int g\cdot(1_A\circ T^n)d\mu_p =$ $\mu_p A\cdot(\int g d\mu_p) = \mu_p A$.

$\Leftarrow$ (proof due to M.S. Keane) Suppose μ is not absolutely continuous with respect to μ_p. Let A be a Borel set with $\mu_p A = 0$ and such that $\mu A > 0$ is maximal. We construct a Borel set B such that $\mu_p B = 0$ and for all n, $\mu T^{-n}B = \mu A$. Let B_1: = TA. Claim: B_1 is also Borel. For we can split T into two homeomorphisms T_0: $[0] \to 2^\omega$, T_1: $[1] \to 2^\omega$ defined by $T_i(ix) = x$, for i = 0,1. Since the T_i are homeomorphisms, the sets $T_i(B\cap[i])$ are Borel; but $TB = T_0(B\cap[0])\cup T_1(B\cap[1])$. Clearly $\mu_p B = 0$. Since $T^{-1}B_1 \supseteq A$ and A was chosen to have maximal μ-measure, $\mu T^{-1}B_1 = \mu A$. For each n, repeat the above argument with T^n replacing T, yielding B_n. Put $B := \bigcup_n B_n$, then $\mu_p B = 0$ and $\mu T^{-n}B = \mu A$ for all n. □

4.6.12 Theorem Let μ be a computable measure. Then $\mu R(\mu_p) = 1$ iff $(\mu T^{-n})_{n\in\mathbb{N}}$ converges strongly to μ_p.

Proof By lemma 4.6.11, $\mu R(\mu_p) = 1$ implies that μ is absolutely continuous with respect to μ_p. The converse is trivial. Now apply the previous lemma. □

4.7 Digression: the difference between randomness and 2-randomness We are interested in the size of the difference between $R(\lambda)$ and $R_2(\lambda)$, the randomness notion that was defined in 3.2.4.1. We have seen in 3.2.4 that $R(\lambda)\cap R_2(\lambda)^c$ is non-empty. On the other hand, by lemma 4.6., there is no computable measure μ such that $\mu(R(\lambda)\cap R_2(\lambda)^c) = 1$: if $\mu R_2(\lambda) = 1$, then $\mu\perp\lambda$, which implies $\mu R(\lambda) = 0$. (Note that, for all we know, there might be a computable μ such that $\mu(R(\lambda)\cap R_2(\lambda)^c) > 0$.)

We now show, as an application of the techniques developed in 4.1-6, that $R(\lambda)\cap R_2(\lambda)^c$ is indeed large: there exists a non-atomic Δ_2 definable measure μ_x such that $\mu_x(R(\lambda)\cap R_2(\lambda)^c) = 1$.

To prove this, we need a random measure, that is, a family of measures $(\mu_x)_{x\in 2^\omega}$ defined as follows:

$$\mu_x = \prod_n (1-p_n^x, p_n^x), \text{ where } p_n^x = \begin{cases} 3/4 & \text{if } x_n = 1 \\ 1/4 & \text{if } x_n = 0. \end{cases}$$

It is easily shown that for each Borel set B, the mapping $x \to \mu_x B$ is measurable. Hence we

may define a measure μ on 2^ω by

$$\mu(A\times B) = \int_A \mu_x B d\lambda(x).$$

μ is obviously computable, hence $R(\mu)$ is well defined. Using a construction exactly parallel to the Fubini theorem for recursive sequential tests (Theorem 4.4.4), one can demonstrate that for all $x \in R(\lambda)$, $\mu_x R(\mu)_x = 1$. For this, it suffices to show that for each recursive sequential test N with respect to μ, $\{x|\ \mu_x N_x > 0\}$ is contained in a recursive sequential test with respect to λ. This can be done if we change slightly the definition of the functions f_k occurring in the proof of Theorem 4.4.4. We now put

$$f_0(x) = 0 \text{ for all } x$$

$$f_k(x) = \sum_{\{i\le k|\, x\in [w^i]\}} \mu_x[v^i], \text{ for } k\ge 1;$$

the rest of the proof then goes through almost literally.

We now show that for $x \in R(\lambda)$, $R(\mu)_x \subseteq R(\lambda)$. For this, it suffices to show that the mapping π_2: $2^\omega\times 2^\omega \to 2^\omega$ defined by $\pi_2<x,y> = y$ is such that for any recursive sequential test N with respect to λ, $\pi_2^{-1}N$ is a recursive sequential test with respect to μ, for in that case, $<x,y> \in R(\mu)$ implies $y \in R(\lambda)$. (Observe that $x \in R(\lambda)$ implies that $R(\mu)_x \neq \emptyset$.) Now $\pi_2^{-1}N$ is obviously $\prod_2$ and is a recursive sequential test, since for all Borel sets A: $\mu\pi_2^{-1}A = \lambda A$.

We thus have that for each $x \in R(\lambda)$, $\mu_x R(\lambda) = 1$. In particular, this is true of the Δ_2 sequence constructed in 3.2.2.3. Fix such a Δ_2 definable μ_x; this μ_x is then recursive in $\emptyset'$. It is not difficult to see that $\mu_x \perp \lambda$; either by Kakutani's theorem (4.6.5) or by observing that $R(\mu) \subseteq R(\lambda\times\lambda)^c$ and applying the Fubini theorem 4.4.4 to conclude that for $x \in R(\lambda)$, $\lambda R(\mu)_x = 0$.

Since our μ_x is singular to λ, we may perform the construction of lemma 4.6.4 (ii) recursively in $\emptyset'$, to obtain a Δ_2 definable sequence (C_n) of $\sum_0$ sets C_n, such that $\lambda\bigcap_n C_n = 0$ and $\mu_x\bigcap_n C_n > 0$. Now $\bigcap_n C_n$ is $\prod_2$, hence $\mu_x R_2(\lambda)^c > 0$ and since $R_2(\lambda)^c$ is a tailset and μ_x a product measure, we get in fact $\mu_x(R(\lambda)\cap R_2(\lambda)^c) = 1$.

Notes to Chapter 4

1. A simple proof of Kakutani's theorem has recently been published by S.D. Chatterji. See S.D. Chatterji, Martingale theory: An analytical formulation with some applications in

analysis, in: Letta (ed.), Probability and analysis, *Lecture Notes in Mathematics* **1206**, Springer-Verlag (1986).

5 Kolmogorov–complexity

Undoubtedly, the notion of Kolmogorov–complexity (sometimes called *descriptive*, as opposed to *computational* complexity), with its attendant complexity–based definition of randomness, is the most important development stimulated by von Mises' attempt to define Kollektivs. The virtues of Kolmogorov–complexity seem to reside in the fact that it allows a discussion of randomness at a more basic level. Indeed, the intuition behind its definition stems from a tradition, going back to Antiquity, which views the essence of chance as (objective) unpredictability or irregularity. So far, of course, we have been concerned with a form of randomness in which irregularity coexists with statistical regularity. In later life, Kolmogorov came to regard the relation between these two forms of chance as *the* problem for the foundations of probability.

> In everyday language we call random these phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking there is no ground to believe that a random phenomenon should possess any definite probability. Therefore we should have distinguished between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges a problem of finding the reasons for the applicability of the mathematical theory of probability to the real world [51,1].

Elsewhere, he writes

> In applying probability theory we do not confine ourselves to negating regularity, but from the hypothesis of randomness of the observed phenomena we draw definite positive conclusions [50,34].

Roughly speaking, irregular sequences are distinguishable from those which show irregularities *and* statistical regularities by the following property: in the latter type of sequences, the Kolmogorov–complexity of an initial segment divided by the length of that segment tends to stabilize. This phenomenon illustrates one of the technical advantages of Kolmogorov–complexity: not only does it classify sequences as random or otherwise, but it also assigns "degrees of randomness" to sequences. This is particularly useful when we study infinite sequences; it allows us, for instance, to discriminate between Δ_2 definable and "truly" random sequences. It must be admitted, however, that Kolmogorov himself considered infinite sequences to be irrelevant for the foundations of probability; indeed, his main motive for developing a measure of complexity for finite sequences was his conviction that only a frequency interpretation in terms of finite sequences is worthy of the name.

The themes introduced above determine the structure of this chapter. Sections 5.1–3 are

concerned with finite sequences. In 5.1 we define Kolmogorov–complexity and irregular sequences. It will turn out that a slight modification of Kolmogorov's definition, first proposed by Chaitin and Levin, has some conceptual and technical advantages. In 5.2 we discuss Kolmogorov's explanation of the applicability of probability theory. 5.3 collects some recursion theoretic properties of the complexity measures introduced in 5.1 and contains a critical discussion of Chaitin's claim that Kolmogorov–complexity sheds light on the incompleteness of formal systems. We then turn to the investigation of infinite sequences. In 5.4 we first characterize (Martin-Löf) randomness in terms of Chaitin's complexity measure, but the full power of this complexity measure (namely, as an indicator for the *degree* of randomness) is revealed only when we study complexity oscillations. Here, we meet various sources of unavoidable order in infinite sequences. The same theme, complexity as degree of randomness, dominates 5.5, where we compare complexity with more traditional measures of disorder, in particular (topological and metric) entropy. Lastly, in 5.6 we look back to Chapter 2 and define admissible place selections using Chaitin's complexity measure. The purpose of the first three sections is expository; apart from the critical discussions they do not contain any new material. The main novelty in 5.4 is that Δ_2 definable sequences always must have "low" complexity. This result allows a very simple proof of a theorem on complexity oscillations due to Martin-Löf. The results in 5.5 on the relation between complexity and topological entropy appear to be new.

5.1 Complexity of finite strings The intuition behind the definition of complexity of finite strings can be stated in various ways. One might say that if a sequence exhibits a regularity, it can be written as the output of a (simple) rule applied to a (simple) input. Another way to express this idea is to say that a sequence exhibiting a regularity can be *coded* efficiently, using the rule to produce the sequence from its code. Taking *rules* to be partial recursive functions from $2^{<\omega}$ to $2^{<\omega}$, we may define the *complexity* of a word w with respect to a rule A to be the length of a shortest input p such that A(p) = w. Sequences with low complexity (with respect to A) are then supposed to be fairly regular (with respect to A). In order to take account of all possible rules (i.e. partial recursive functions), we then use a *universal* machine. One obtains different concepts of complexity by imposing additional restrictions on the functions A. We begin with Kolmogorov–complexity, where no such restrictions are imposed.

5.1.1 Kolmogorov–complexity

5.1.1.1 Definition Let $A: 2^{<\omega} \to 2^{<\omega}$ be a partial recursive function with Gödelnumber $\ulcorner A \urcorner$. The *complexity* $K_A(w)$ of w with respect to A is defined to be

$$K_A(w) = \begin{cases} \infty & \text{if there is no p such that } A(p) = w \\ |p| & \text{if p is a shortest input such that } A(p) = w. \end{cases}$$

A universal machine U is said to be *asymptotically optimal* if it is specified by the requirement that on inputs of the form $q = 0^{\ulcorner A \urcorner}1p$ (i.e. a sequence of $\ulcorner A \urcorner$ zeroes followed by a one, followed by a string p), U simulates the action of A on p. Fix a Gödelnumbering and an asymptotically universal machine U and put $K(w) := K_U(w)$. K is called the *Kolmogorov–complexity* of w (Kolmogorov [48–51]). Inputs will also be called *programs*.

The fundamental properties of Kolmogorov–complexity are stated in the papers by Kolmogorov cited above, in the survey article by Levin and Zvonkin [54] and, in a slightly different form, in Chapter 15 of Schnorr's [88]. Clearly, we have

5.1.1.2 Lemma (a) For any partial recursive $A: 2^{<\omega} \to 2^{<\omega}$ and for all w, $K(w) \leq K_A(w) + \ulcorner A \urcorner + 1$; (b) for some constant c and for all w, $K(w) \leq |w| + c$.

Before we put the above definition to work, let us remark that complexity measures are not restricted to finite words over the alphabet {0,1}; *any* alphabet n = {0,.., n–1} will do. We only have to replace the functions $A: 2^{<\omega} \to 2^{<\omega}$ by functions which have as their range n^{ω}. Identifying a natural number with its binary representation, it makes sense to speak of the complexity of natural numbers. Similarly, given some recursive bijection $2^{<\omega} \to 2^{<\omega} \times 2^{<\omega}$, it makes sense to speak of the complexity of a *pair* of binary strings.

We now embark upon the promised definition of regular and irregular sequences. First suppose that $K(w) << |w|$; then for some algorithm A and input p such that both A and |p| are small compared to |w|, A(p) = w. In this case, we say that w exhibits a (simple) regularity. How small K(w) has to be is a matter of taste. Since we shall consider regularity only in connection with infinite sequences (cf. section 5.5), we shall not be precise here. On the other hand, it is worthwhile to develop a theory of *irregularity* for finite sequences. Recall that for some c, $K(w) \leq |w| + c$. We wish to say that w is irregular if it is maximally complex. Formally:

5.1.1.3 Definition Fix some natural number m. A binary string w is called *irregular* if $|w| > m$ and $K(w) > |w| - m$.

The definition of irregularity is relative to the choice of m, but this is inessential for our (highly theoretical) purposes.

A note on terminology What we call *irregular* is usually called *random*. The reason that we prefer the term "irregular" over "random", is that we have used randomness so far in a *stochastic* sense; but the intuition behind Kolmogorov's definition is *combinatorial* rather than stochastic. This will become particularly clear when we generalize this intuition to irregularity for binary words known to belong to a recursively enumerable *subset* of $2^{<\omega}$. It *is* possible to put a condition on the complexity of a word w which implies that w is approximately a Kollektiv with relative frequency (of 1) equal to p. However, this condition is stochastic from the outset, in the sense that it explicitly mentions a measure (cf. 5.2). Only when the measure is Lebesgue measure is the condition for stochastic randomness identical to the condition for irregularity; but this reflects the fact that Lebesgue measure is a so–called maximum entropy measure for the system $(2^{\omega},T)$. We shall come back to this topic in 5.5. In 5.1.4 the two aspects of definition 5.1.1.3, the combinatorial and the stochastic, will be separated; in 5.4. and 5.5 we investigate the corresponding definitions of randomness.

A simple counting argument will show that infinitely many irregular sequences exist. In the sequel, the expression "#A" always stands for the cardinality of the (finite) set A.

5.1.1.4 Lemma (a) $\#\{w\in 2^n \mid K(w) \leq n–m\} \leq 2^{n–m+1}–1$; (b) $\#\{w\in 2^n \mid K(w) > n–m\} > 2^n\cdot(1 – 2^{–m+1})$

Proof (a) The number of programs on U of length ≤ n–m is $\leq 2^{n–m+1}–1$. Hence (b) at least $2^n – 2^{n–m+1} = 2^n\cdot(1 – 2^{–m+1})$ sequences in 2^n satisfy K(w) > n–m. □

Note the extreme simplicity of the argument: it can be formalized in any formal system capable of handling finite sets of integers. This is to be contrasted with the fact, proved in 5.3, that the set of irregular sequences contains no infinite recursively enumerable subsets.

5.1.2 Chaitin's modification While definition 5.1.1.1 captures the basic idea of a complexity measure for sequences, it is open to dispute whether it is really the most satisfactory definition. The intuition behind the definition is supposed to be that if p is a minimal program (on U) for w (i.e. a program of shortest length), then the *bits* of p contain all information necessary to reproduce w on U. But this might well be false: U might begin its operation by scanning all of p to determine its length, only then to read the contents of p bit for bit. In this way, the information p is really worth $|p| + \log_2|p|$ bits, so it's clear we have been cheating in calling |p| the complexity of p.

Chaitin [12–14] and Levin [55] independently observed that we may circumvent this problem

if we modify the construction of our Turing machines. We shall follow Chaitin's description. From now on, Turing machines are assumed to have worktapes, a read–only input tape and a write–only output tape. Furthermore, we constrain the reading head (operating on the input tape) to read the input in one direction only and we do not allow blanks as endmarkers. We say that a machine M (of this type) performs a *successful* computation on input p if M halts while the reading head is scanning the last bit of p. The fact that we defined a successful computation using the *last bit* of p and not the *first blank* following p means that p must itself indicate where it ends; in other words, p must be a *self delimiting* program. Formally, this means that the domain of M, that is, the set of p such that M performs a successful computation on p, is *prefixfree*: if p and q are both in the domain of M, then neither is an initial segment of the other. We may now introduce

5.1.2.1 Definition A *prefix algorithm* is a partial recursive function A: $2^{<\omega} \to 2^{<\omega}$ which has a prefixfree domain.

To define a reasonable complexity measure associated with prefix algorithms, we need a universal prefix algorithm. At first sight it might seem that no such algorithm exists, since the set of Gödelnumbers of prefix algorithms is $\prod_1$. But there exists nonetheless a recursive enumeration of the set of prefix algorithms, as follows. We construct an algorithm P which turns any number e into a Gödelnumber for a prefix algorithm P(e). Given e, generate the domain of the function ϕ_e with Gödelnumber e. A partial recursive function $\phi_{P(e)}$ with Gödelnumber P(e) is determined by the following prescription: $\phi_{P(e)}$ equals ϕ_e except for those $q \in \mathrm{dom}\phi_e$ which are initial segments or prolongations of previously generated $p \in \mathrm{dom}\phi_e$. If one of these cases occurs, $\phi_{P(e)}(q)$ is undefined. By construction, $\phi_{P(e)}$ is a prefix algorithm and all prefix algorithms have at least one Gödelnumber which occurs in the range of P. Hence the set of prefix algorithms, as opposed to the set of their Gödelnumbers, is recursively enumerable. (In other words, range(P) is not "extensional".)

We may now define a universal prefix algorithm as in definition 5.1.1.1: on inputs of the form $q = 0^{\lceil A \rceil}1p$, U simulates the action of A on p, where A is a prefix algorithm. We put

5.1.2.2 Definition Let A: $2^{<\omega} \to 2^{<\omega}$ be a prefix algorithm with Gödelnumber $\lceil A \rceil$. The *complexity* (also called *information*) $I_A(w)$ of w with respect to A is defined to be

$$I_A(w) = \begin{cases} \infty & \text{if there is no p such that } A(p) = w \\ |p| & \text{if p is a shortest input such that } A(p) = w. \end{cases}$$

If U is the universal prefix algorithm constructed above, we let $I(w) := \min \{|p| \mid U(p) = w\}$.

This definition is due to Chaitin [12;13]; the notation "I(w)" derives from the formal similarities of this complexity measure with Shannon's measure of information. Indeed, the complexity measure I is not only conceptually cleaner than K, it has also a number of technical advantages, as will become gradually clear in the sequel. We first state some fundamental properties, parallel to those of K.

5.1.2.3 Lemma For some constant c and for all w: $I(w) \leq |w| + I(|w|) + c$.

Proof Let A be the following algorithm: given input p, it simulates the action of the universal machine U on some initial segment q of p such that U(q) is defined; if m is the natural number determined by U(q), A reads the next m bits of the input tape and copies them on the output tape. By our conventions on a successful computation, A(p) is defined only if $|p| = |q| + m$; this turns A into a prefix algorithm. Now if q is a (minimal) program for |w|, then A(qw) = w and $I(w) \leq I_A(w) + \ulcorner A \urcorner + 1 \leq |w| + I(|w|) + \ulcorner A \urcorner + 1$. □

Here we see clearly the distinguishing feature of the new algorithms: acceptable inputs must themselves indicate where they end, hence the extra I(|w|)-term.

5.1.2.4 Lemma (a) for some constant c: $\#\{w \in 2^n | I(w) \leq n + I(n) - m\} \leq 2^{n-m} \cdot c$; (b) for some constant c: $\#\{w \in 2^n | I(w) > n + I(n) - m\} > 2^n \cdot (1 - 2^{-m} \cdot c)$.

A proof of this lemma may be found in Chaitin [12,337] (and in 5.1.3 we shall derive 5.1.2.4 from a property of conditional complexity). It should be noted that, whereas the corresponding result for K was trivial, the proof of 5.1.2.4 is rather involved. This fact may add fuel to a nagging suspicion on the reader's part, that Chaitin's definition introduces only gratuitious complications. This impression, however, is mistaken; although proofs are sometimes more difficult, theorems and formulae generally take on a pleasanter aspect. One example will be given below; we shall meet another instance of this phenomenon in 5.1.3, where we define *conditional* complexity.

5.1.2.5 Example The main technical advantage of I lies in the fact that desirable results which hold for K only with logarithmic error terms, are now true within O(1). E.g. for K we have only: $K(<v,w>) \leq K(v) + K(w) + \min[\log_2 K(v), \log_2 K(w)] + O(1)$, but the formula for I is more intuitive:

Claim For some constant c, for all v,w: $I(<v,w>) \leq I(v) + I(w) + c$.

Proof of claim Let A be the prefix algorithm which does the following. On input s, it sets U

reading s; if U performs a successful computation on s, it outputs U(s). If U halts while scanning the last bit of some proper initial segment s' of s, it stores U(s') on its worktape and continues reading s'', where s = s's''. If U halts again scanning the last bit of s'', A outputs <U(s'),U(s'')> and stops. Simulating A on U we get the desired result. □

The root of the superiority of I over K can thus be traced to the circumstance that we may concatenate self delimiting programs; we only have to add a couple of bits which tell the machine that it must expect two (or more) programs (this is what simulating A on U means).
One immediate application of the above formula for the complexity of a pair will illustrate its force: if T is the leftshift on 2^ω, we have for some constant c and all x in 2^ω,

$$I(x(n+m)) \leq I(x(n)) + I(T^n(x(n+m))) + c.$$

The sequence of functions $f_n(x) := I(x(n))$ thus forms a *subadditive* sequence and by the subadditive ergodic theorem[1], we have that for any ergodic measure μ there exists a constant H such that

$$\lim_{n\to\infty} \frac{I(x(n))}{n} = H \quad \mu\text{–a.e.}$$

(It is, however, notoriously difficult to identify the limit of a subadditive process; eventually, in 5.5.2, we shall show that H equals the metric entropy of μ, but via an entirely different route.) These considerations justify calling the property of I stated in claim 5.1.2.5 *subadditivity*. End of the example.

Parallel to definition 5.1.1.3 we have

5.1.2.6 Definition Fix a natural number m. A binary word w is *irregular* if $I(w) > |w| + I(|w|) - m$.

By lemma 5.1.2.4, the great majority of binary strings is irregular.

Before we turn to conditional complexity, we introduce an important technical tool. Since we defined I by restricting the class of admissible algorithms to those with a prefixfree domain, we need some criterion to decide whether a certain task can be performed by a prefix algorithm. Almost trivially, we have

5.1.2.7 Lemma (a) If A is a prefix algorithm, then $\sum_{A(p) \text{ defined}} 2^{-|p|} \leq 1$. (b) $\sum_{w \in 2^{<\omega}} 2^{-I(w)} \leq 1$.

Proof (a) The cylinders in {[p]| A(p) defined} are pairwise disjoint. (b) Apply (a) to the universal prefix algorithm. □

Part (a) of the following lemma, to be called the Chaitin–Kraft inequality[2] is a converse to lemma 5.1.2.7.

5.1.2.8 Lemma (a) Let S be an r.e. set of pairs <w,m> such that $\sum_{<w,m>\in S} 2^{-m} \leq 1$. Then there exists a prefix algorithm A with the property: <w,m> ∈ S iff ∃p (|p| = m & A(p) = w).
(b) Simulating A on the universal machine, we have for all <w,m>∈S: $I(w) \leq m + \lceil A \rceil + 1$.

For a proof, see Chaitin [12,333]. Part (b) will be our main tool in deriving upper bounds on I. Here is a useful consequence of lemma 5.1.2.8:

5.1.2.9 Lemma Let f: ω → ω be a total recursive function. (a) If $\sum_n 2^{-f(n)} = \infty$, then ∀m ∃n≥m (I(n) > f(n) + m). (b) If $\sum_n 2^{-f(n)} < \infty$, then ∃m ∀n (I(n) ≤ f(n) + m).

Proof (a) follows from part (b) of lemma 5.1.2.7. To prove (b), determine k such that $\sum_{n\geq k} 2^{-f(n)} \leq 1$. Lemma 5.1.2.8 (b), applied to the r.e. relation {<n,f(n)> | n ∈ ω} yields a constant m_0 such that for n ≥ k: I(n) ≤ f(n) + m_0. Put m_1:= max{I(n) | n≤k}. Then for all n: I(n) ≤ f(n) + m. □

In conclusion of this subsection, we mention a result on the relation between K and I due to Solovay [93]. Obviously, for all w: K(w) ≤ I(w).

5.1.2.10 Lemma For all w, $I(w) = K(w) + K[K(w)] + O(\log_2 K[K(w)])$.

The intuitive meaning of this expression is, that it takes $K[K(w)] + O(\log_2 K[K(w)])$ bits to turn a minimal program for w into a self delimiting program.

5.1.3 Conditional complexity In Chaitin's set–up, conditional complexity comes in two varieties. The most straightforward definition is the following. We consider algorithms B(p,q) in two arguments p and q, which can be thought of as being presented on the input tape and a work tape, respectively, of a Turing machine. Such an algorithm is called a *prefix algorithm* if for each q, the set {p| B(p,q) defined} is prefixfree. We shall use U interchangeably for both the one–argument and the two–argument universal prefix algorithm.

5.1.3.1 Definition $I_0(w|v) := \min\{|p|\ U(p,v) = w\}$.

For the second variant, denoted $I(w|v)$, we demand that U is presented, not with v itself, but rather with a minimal program for v.

5.1.3.2 Definition $I(w|v) := \min\{|p|\ U(p,v^*) = w\}$, where v^* is some minimal program for v.

It will be seen in the sequel that both notions are useful. Some easy facts:

5.1.3.3 Lemma For some constant c and all w: $I_0(w||w|) \leq |w| + c$.

Proof The algorithm B defined by $B(w,|w|) = w$ is a prefix algorithm in the new sense. □

5.1.3.4 Lemma For some constant c and for all w: $I(w||w|) \leq I_0(w||w|) + c$.

Proof Consider the following prefix algorithm B: on being presented with $<p,q>$, it calculates U(q); if and when this computation halts, it calculates U(p,U(q)). Hence if p is a program such that $U(p,|w|) = w$, then $B(p,|w|^*) = w$. □

The difference between the two notions of conditional complexity is brought out by the following lemma:

5.1.3.5 Lemma (a) $I_0(w||w|) - I(w||w|)$ is unbounded; (b) For some constant c and all w: $|I(w||w|) - I_0(w|<|w|,I(|w|)>)| \leq c$.

A proof may be found in Chaitin [12,338]. The main difference between I and I_0, however, is that the former satisfies

5.1.3.6 Lemma For some constant c, for all v,w: $|I(w|v) + I(v) - I(<w,v>)| \leq c$.

This formula is proved in Chaitin [12,336] and is desirable if we think of I as giving the *information* of a string. As an application of the preceding lemma, we may now prove lemma 5.1.2.4 (a): for some constant c, $\#\{w\in 2^n|\ I(w) \leq n + I(n) - m\} \leq 2^{n-m}\cdot c$.
Observe that for some constant d, all n and all w in 2^n: $|I(<w,n>) - I(w)| \leq d$.
This observation, taken in conjunction with the lemma 5.1.3.6, enables us to write (for some constant c): $\#\{w\in 2^n|\ I(w) \leq n + I(n) - m\} = \#\{w\in 2^n|\ I(w) - I(n) \leq n - m\} \leq \#\{w\in 2^n|\ I(w|n) \leq n - m - c\}$ (we apply 5.1.3.6 to the pair $<w,n>$).
But $\#\{p|\ |p| \leq n - m - c \ \&\ U(p,n) \text{ defined}\} \leq 2^{n-m-c+1}$.

5.1.4 Information, coding, relative frequency In the previous subsection, we studied the

effect of using the information contained in a word v upon the complexity of a word w. We now show how to take in account extraneous or global information, namely, knowledge of a recursively enumerable subset of $2^{<\omega}$ to which a given word belongs, or knowledge concerning the probability of a word, as given by some computable probability distribution. We first make explicit the relation between complexity and coding, which was used to motivate the definition of complexity in 5.1.1; the effect of the extra information may then be explained in terms of coding procedures.

5.1.4.1 Definition A *prefix code* is a prefix algorithm (in the sense introduced in 5.1.3) A: $2^{<\omega}\times\omega \rightarrow 2^{<\omega}$ such that for all n, $\{w| \exists p\ (A(p,n) = w)\} \subseteq 2^n$. Note that A is given n itself, not a minimal program for n.

A prefix code A provides for each n a coding scheme for the binary words of length n which is uniquely decipherable: the requirement that A be a prefix algorithm ensures that any sequence of length n·k can be coded into a uniquely decodable concatenation of k codewords. Observe that any prefix algorithm can be transformed into a prefix code by a suitable restriction of its domain. For instance, if U is the universal prefix algorithm, we may define a prefix code U* by setting U* equal to U on $\mathrm{dom}U^* = \{p| \exists w\ (U(p,|w|) = w)\}$. U* embodies many different coding schemes. The expression $I_0(w||w|) = \min\{|p|| U^*(p,|w|) = w\}$, where I_0 was defined in 5.1.3.1, gives the length of the shortest code for w with respect to U*. The expression $I_0(w||w|)/|w|$ might be called the *compression coefficient* of w; it measures how efficiently w can be coded, using the universal coding U*. In section 5.5 we shall derive various asymptotic estimates on the compression coefficient.

The fact that U* embodies many different coding schemes will now used to derive upper bounds on I in the presence of extraneous information. The following lemmas may be seen as elaborations of two aspects of the definition of irregularity (5.1.2.6). We motivated this definition as follows: a finite binary sequence w was judged to be irregular if its complexity is close to the theoretical upper bound |w| + I(|w|). But this upper bound can be interpreted in at least two ways: if |w| = n, then n is the logarithm of the cardinality of 2^n, or minus the logarithm of the probability of w on the uniform distribution. The first lemma elaborates the first interpretation.

5.1.4.2 Lemma Let $S \subseteq 2^{<\omega}$ be an r.e. set of words, $S_n := S \cap 2^n$, $\#S_n$ the cardinality of S_n. Then for some constant c, for all n and for all $w \in S_n$:

$$I_0(w|n) \leq [\log_2 \#S_n] + c \text{ and } I(w|n) \leq [\log_2 \#S_n] + c.$$

As a consequence, for some constant d and all $w \in S_n$:

$$I(w) \leq [\log_2 \#S_n] + I(|w|) + d.$$

Proof For each n, order the words in S_n lexicographically and enumerate them in this order. If p is the ordinal number of a word w in S_n, we may consider p to be a binary string of length $[\log_2 \#S_n] + 1$, by adding if necessary zeros to the left of the ordinal number p, written in binary notation. Now define an algorithm B as follows. If $|p| = [\log_2 \#S_n] + 1$, then B(p,n) is the p^{th} word in S_n. By construction, B is a prefix algorithm in the sense of 5.1.3. Hence for some c, for all n and w in S_n: $I_0(w|n) \leq [\log_2 \#S_n] + c$. To get $I(w|n) \leq [\log_2 \#S_n] + d$, replace B by B' defined as follows: B'(p,q):= B(p,U(q)), where U is the universal prefix algorithm. To get the upper bound on I(w), apply lemma 5.1.3.6. □

5.1.4.3 Lemma Let μ be a computable measure on 2^ω. Then for some c and all w:

$$I_0(w||w|) \leq [-\log_2 \mu[w]] + c \text{ and } I(w||w|) \leq [-\log_2 \mu[w]] + c.$$

As a consequence, for some c and all w:

$$I(w) \leq [-\log_2 \mu[w]] + I(|w|) + c.$$

Proof Since for each n

$$\sum_{w \in 2^n} 2^{-[-\log_2 \mu[w]] - 1} \leq 1,$$

we can, using the Chaitin–Kraft inequality, construct prefix algorithms A_n, uniformly in n, such that

$$\forall n \; \forall w \in 2^n \; \exists p \; (|p| = [-\log_2 \mu[w]] - 1 \;\; \& \;\; A_n(p) = w).$$

Defining B by $B(p,n) := A_n(p)$, we see that for some c and all w:

$$I_0(w||w|) \leq [-\log_2 \mu[w]] + c$$

and if we put B'(p,q) := B(p,U(q)), we get for some c,

$$I(w||w|) \leq [-\log_2 \mu[w]] + c.$$

The upper bound on I(w) follows again by applying lemma 5.1.3.6. □

As we said above, both lemmas can be seen as generalizations of lemma 5.1.2.3:

for some constant c *and for all* w: $I(w) \leq |w| + I(|w|) + c$,

corresponding to different interpretations of the expression "|w|". For $n = |w|$ denotes not only the length of $w \in 2^n$, but is also equal to $\#\log_2 S_n$ if $S = 2^{<\omega}$ (this observation leads to lemma 5.1.4.2) and to $[-\log_2 \lambda[w]]$ (which leads to lemma 5.1.4.3). The upper bound of lemma 5.1.2.3 is not always sharp; in particular, additional information on w may lead to a sharper estimate on I(w). The above two lemmas are cases in point.

Lemma 5.1.4.2 says roughly that if we know that w belongs to S, to specify w completely it suffices to give n (with cost I(n)) and then the ordinal number of w in S_n (with cost ≤ $[\log_2 \#S_n] + 1$). This might be called the *combinatorial* or *topological* aspect of I. The reason for this nomenclature will become clear in 5.5, when we discuss the relation of I to topological and metric entropy.

On the other hand, lemma 5.1.4.3 is based on the idea that words which have large probability (with respect to μ) can have short codes, at the expense of words with small probability, which must then receive long codes. This could be called the *metric* aspect of I. To give the reader an idea of the size of the upper bounds obtained in this way, we need the following corollary of the Shannon – McMillan – Breiman theorem. Unexplained concepts are defined in section 7.

5.1.4.4 Theorem (Petersen [82,263] Let μ be an ergodic measure on 2^ω with entropy H(μ). For all $\varepsilon > 0$ there exists $n_0(\varepsilon)$ such that for $n \geq n_0(\varepsilon)$, 2^n can be partitioned into two sets B_n (of "bad words") and G_n (of "good words") which satisfy
(1) $\mu[B_n] < \varepsilon$;
(2) for all $w \in G_n$, $2^{-n(H(\mu)+\varepsilon)} < \mu[w] < 2^{-n(H(\mu)-\varepsilon)}$.
In other words, if we know that w belongs to the "good" words of μ (for given ε), then the upper bound on I(w) is given by $I(w) \leq (H(\mu)+\varepsilon)\cdot|w| + I(|w|) + c$. For "bad" words the upper bound of lemma 5.1.4.3 may be much worse than that of lemma 5.1.2.3.

With these two interpretations on the upper bound of I at our disposal, we may develop the fundamental intuition that a string is irregular if its complexity is almost maximal, in two directions. We shall do so in section 5.5.

In conclusion, we note that lemma 5.1.4.2 can be used to derive an upper bound on I(x(n·k)) in terms of the relative frequencies of words of length k occurring *in* x(n·k). This upper bound is helpful when we study the relation between I and metric entropy.

5.1.4.5 Lemma (Kolmogorov [50]) Let $x \in 2^\omega$. Fix an integer k and denote by $q_i(n)$ the relative frequency of the i^{th} word of length k in x(n·k). Then

$$I(x(n\cdot k)) \leq -n\cdot \sum_{i=1}^{2^k} q_i(n)\log_2 q_i(n) + I(n\cdot k) + O(\log_2 n).$$

Proof By lemma 5.1.4.2, it suffices to show that the number N of words of length n·k which have the given set of frequencies $q_1(n),...,q_m(n)$, where $m = 2^k$, is less than

$$-n\cdot\sum_{i=1}^{2^k} q_i(n)\log_2 q_i(n) + O(\log_2 n).$$

For the verification that this is indeed so, the reader may consult Levin and Zvonkin [54]. (They prove the result for K, but the proof goes over unchanged.) □

It is instructive to compare the preceding lemma with lemma 5.1.4.3. Both determine an upper bound on I(w) in terms of probabilities; but in 5.1.4.5 these probabilities are the relative frequencies of small words in w, whereas in 5.1.4.3 the upper bound is derived using the frequency of w itself.

5.1.5 Discussion Obviously the definition of complexity is open to the charge of arbitrariness on various accounts. For one thing, we might have chosen a different Gödelnumbering or a different universal machine. The difference between the resulting complexity measures is then bounded by a constant. While this might impair the practical utility of complexity, it is quite harmless for theoretical purposes. In particular the asymptotic results derived later are not affected by such a change of scale.
More serious, perhaps, is the decision to restrict the concept of a rule to partial recursive functions. Here, we are confronted with the same problem as in Chapters 2 and 3: Why choose only *recursive* place selections, why choose only *recursive* sequential tests?
Complexity was invented to formalize an essentially negative concept, namely irregularity. This formalization can succeed only if we replace the implicit negation of *all* regularity by a negation of some particular form of regularity. The particular form of regularity we choose to reject depends upon our view of chance. If we regard it as something subjective, e.g. if we believe that the universe is really deterministic and that the appearance of chance is caused by our limited observational and computational abilities, then a definition of rule which reflects our mental powers is not unreasonable. But if we believe in objective chance, for instance because we believe in quantum mechanics and the no–hidden variable proofs, then there seems to be no reason at all why partial recursive rules should occupy a privileged position.

We have already seen, for example, that some Δ_2 definable sequences are random; but such sequences can with reason be regarded as far too regular, since they are produced by a Turing machine operating by trial and error. This fact prompted Müller [76] to define a complexity measure using $\sum_2$ instead of $\sum_1$ functions. The cynic might then ask: Why stop here? We would be surprised to find any arithmetical or analytical regularity in a sequence. On the positive side, we may remark that already the above complexity measures, which were defined using recursive functions only, reveal that Δ_2 definable sequences are really deterministic sequences: the asymptotic behaviour of K and I on a Δ_2 definable sequence is rather atypical

(see section 5.4).
On the whole, however, we must conclude that complexity as presented above fits the *subjective* aspect of irregularity and chance best. This is even more true of the resource–bounded complexity measures briefly discussed below.

One more source of arbitrariness might be given by the coexistence of different definitions of complexity for finite binary strings: for instance Kolmogorov–complexity, Chaitin–complexity and monotone complexity, of which more will be said in 5.4. Nor is this the end of the list. On this score, however, we are not so pessimistic: we believe that there are good arguments to show that Chaitin's definition is both conceptually and technically the most satisfactory.

5.1.6 Digression: Resource–bounded complexity In the definition of K and I one feature of computations has been left out of consideration: the amount of resources (time, space; in some cases the number of times an oracle is consulted) needed to compute a string from a given program. This is the motivation behind *resource–bounded complexity*. The gist of this concept can be gathered from the following definition:

5.1.6.1 Definition Let g be a total recursive function and U a universal Turing machine. Then $K^g(w) := \min\{|p|\ |\ U(p) = w \text{ and the computation takes} \leq g(|p|) \text{ steps}\}$.

Natural choices for g would be: polynomials, functions of order $f \cdot \log_2 f$, where f is a polynomial, or functions of order 2^{cn} etc. For information on the use of these complexity measures in computer science, the reader may consult the references [36], [59] and [90][2a].

5.2 Kolmogorov's program In [50,34], Kolmogorov writes

> The idea that "randomness" consists in a lack of "regularity" is thoroughly traditional. But apparently only now has it become possible to found directly on this simple idea precise formulations of conditions for the applicability of the mathematical probability theory to real phenomena.

In other words, irregularity leads to (stochastic) randomness and

> Practical deductions of probability theory can be justified as consequences of hypotheses about the *limiting* complexity, under given restrictions, of the phenomena in question [50,34]. The applications of probability theory can be put on a uniform basis. It is always a matter of consequences of hypotheses about the impossibility of reducing in one way or another the complexity of the description of the objects in question [50,39].

For later reference, we shall call this view *Kolmogorov's program*. Its most sophisticated presentation is [50], but some of the fundamental ideas are already present in [47]. We do not

give the formal details of the program, but limit ourselves to some philosophical comments. To give the reader an impression of the formal details, we state here a result for *infinite* sequences (proven in 5.4) which may be seen as an illustration (but *only* an illustration) of this program:

If μ *is a computable measure, then* $x \in R(\mu)$ *iff* (*) $\exists m\, \forall n\, I(x(n)) > [-\log_2\mu[x(n)]] - m$.

This theorem is an illustration of Kolmogorov's program in the following sense: it states that regular statistical behaviour, in this case the satisfaction of the effective probabilistic laws associated with the measure μ, is implied by the assumption of (almost) maximal complexity compatible with that measure. We saw in 5.1.4.3 that the upper bound on $I(x(n))$ is of the form $[-\log_2\mu[x(n))]] + I(n) + c$. Condition (*) indeed states that $I(x(n))$ is "sufficiently close to the upper bound": by lemma 5.1.2.9, if $a > 1$ (and computable), then for some c and all n, $I(n) \leq a\cdot\log_2 n + c$. Hence $I(n) \in o(n)$, whereas, at least for ergodic measures, $[-\log_2\mu[x(n)]]$ is of order n for almost all x. (Of course, (*) does not quite express that the complexity is maximal; although the term I(n) is of lower order, hence may be neglected for large n, it has to be explained why it doesn't occur in the right hand side of (*). This matter is taken up in the next section.)

One of the reasons why the theorem announced above cannot be taken as a literal fulfillment of Kolmogorov's program, is the fact that it is stated in terms of infinite sequences. Kolmogorov considered it to be a major advantage of complexity, that it allowed a smooth theory of randomness for *finite* sequences. Contra von Mises, he believed that infinite sequences could not serve as a foundation for probability theory.

> The set theoretic axioms of the calculus of probability [...] had solved the majority of formal difficulties in the construction of a mathematical apparatus [...] so successfully that the problem of finding the basis for real application of the results of the mathematical theory of probability became rather secondary to many investigators. I have already expressed the view that the basis for the applicability of the results of the mathematical theory of probability to real "random phenomena" must depend on some form of the *frequency concept of probability*, the unavoidable nature of which has been established by von Mises in a spirited manner. However, for a long time I had the following views.
> (1) The frequency concept based on the notion of limiting relative frequency as the number of trials increases to infinity, does not contribute anything to substantiate the applicability of the results of probability theory to real practical problems where we always have to deal with a finite number of trials.
> (2) The frequency concept applied to a large but finite number of trials does not admit a rigorous formal exposition within the framework of pure mathematics.
> Accordingly, I have sometimes put forward the frequency concept which involves the conscious use of certain not rigorously formal ideas about "practical reliability", "approximate stability of the frequency in a long series of trials", without the precise definition of the series which are "sufficiently large" etc.
> I still maintain the first of the two theses mentioned above. As regards the second, however, I have come to realise that the concept of random distribution of a property in a large finite

> population can have a strict formal mathematical exposition [47,369].

We do not think that the use of finite, instead of infinite Kollektivs connects probability theory closer with reality. Although it is theoretically possible to verify of a finite sequence of data that it is a finite Kollektiv[3], this is not the way probability theory is used in practice: one *assumes* that the data form a Kollektiv with respect to some distribution and one makes predictions on that hypothesis. If the predictions are wrong, then so is the hypothesis. Since the property of being a Kollektiv is thus never exhaustively verified, it does not seem mandatory to use finite Kollektivs only. In general, Kollektivs should be thought of as a vehicle for expressing the necessary presuppositions of sucessful applications of probability (when interpreted as relative frequency), not as an instrument yielding *immediately* verifiable or falsifiable predictions. In fact, on the frequency interpretation, in any of its versions, such *immediately* verifiable or falsifiable predictions are impossible. It then appears to be of secondary importance whether we express the necessary presuppositions in terms of a finite or an infinite model.

But even if we accept infinite sequences in the foundations of probability, the above theorem is still not quite what Kolmogorov has in mind. It is clear from the quotation just given, that Kolmogorov to a large extent subscribes to von Mises' version of the frequency interpretation. In particular, relative frequency is the primary concept, not measure, as in the propensity interpretation. But if that is so, (*) has to be replaced by a different condition; after explaining von Mises' definition of Kollektiv, Kolmogorov observes

> But it turns out that this requirement can be replaced by another one that can be stated much simpler. The complexity of a sequence of 0's and 1's [of length n and with frequency of 1 approximately equal to p] cannot be substantially larger than $nH(\mu_p) = n\cdot(-p\log_2 p - (1-p)\log_2(1-p))$ [cf. lemma 5.1.4.5]. It can be proved that *the stability of frequencies in the sense of von Mises is automatically ensured if the complexity of the sequence is sufficiently close to the upper bound indicated above* [50,35].

Clearly, Kolmogorov envisages a condition of randomness in which the complexity I(x(n)) is compared with an expression involving the (limiting) frequency p of 1; but in (*) I(x(n)) is compared with an expression which involves the (limiting) relative frequency of the word x(n) as given by μ_p (cf. the difference between lemmas 5.1.4.3 and 5.1.4.5). Hence (*) implicitly refers to coordinate-wise probabilities and not to the (limiting) relative frequency of 1. This is of course to be expected, given the material from section 4.6 and the fact that (*) is an equivalent condition for randomness. We have added these cautionary remarks to warn the reader that the characterization of (Martin-Löf) randomness in terms of complexity cannot be seen as an execution of Kolmogorov's program.

In our opinion, the most important feature of Kolmogorov's program is not so much its finitary character, but rather the explanation scheme that it offers. Von Mises based the applicability of probability theory on two (idealizations of) brute facts: existence of limiting relative frequencies and invariance under admissible place selections. Kolmogorov replaces admissibility by simplicity:

> In fact, we can show that in sufficiently large populations the distribution of the property may be such that the frequency of its occurrence will be almost the same for all subpopulations, when the *law of choosing these is sufficiently simple* [47,370].

In other words, a prediction is successful if the place selections which are involved in its derivation (in the sense of 2.4) have a simple description, while the phenomena are complex. This characterization of successful predictions seems correct for a number of cases, although it is not applicable to situations involving, for instance, two independent coins: the place selection determined by the second coin is, in an absolute sense, no less complex than the Kollektiv determined by the first coin. But a modification of Kolmogorov's program is able to handle this situation as well: what seems to be important is not so much that the selection is simple and the data complex, but rather that there exists an "information gap" between place selection and Kollektiv. The existence of such a gap can be stated precisely using some form of conditional complexity, and we shall do so in 5.6.

5.3 Metamathematical considerations on randomness The present section serves two purposes: we collect some recursion theoretic properties of the complexity functions K and I, and, more importantly, we investigate Chaitin's claim that the ideas of complexity theory may help to explain the incompleteness of (sufficiently rich) formal systems.

In [13,336] Chaitin reformulates Gödel's first incompleteness theorem as follows:

> Here is our incompleteness theorem for formal axiomatic theories whose arithmetical consequences are true. The set-up is as follows: the axioms are a finite string, the rules of inference are an algorithm for enumerating the theorems given the axioms and we fix the rules of inference and vary the axioms. Within such a formal system a specific string cannot be proven to be of entropy [=complexity] greater than the entropy of the axioms of the theory. Conversely, there are formal theories whose axioms have entropy n + O(1) in which it is possible to establish all true propositions of the form "I(specific string) > n".

In other words, Chaitin claims there exist constants c and d such that (i) an axiomatic theory with axiom p does not prove any statement of the form "$I(w) > I(p) + c$", and (ii) for any n, one may construct an axiomatic theory with axiom q_n which proves all statements of the form "$I(w) > n$" and for which $I(q_n) \leq n + d$. (i) implies that many assertions on the complexity of individual binary strings are undecidable in arithmetic or set theory and as such it can be

compared to the first incompleteness theorem. But (i) and (ii) go much further and assert that there exists a precise quantitative relationship between the information content of an axiom system (as measured by the complexity of the axioms) and the values of n such that $I(w) > n$ is not derivable in that system. Chaitin's ultimate aims are even more ambitious:

> I would like to be able to say that if one has ten pounds of axioms and a twenty-pound theorem, then the theorem cannot be derived from the axioms [14,942].

Hence not only the underivability of certain true *complexity* statements is to be explained by an appeal to the finite information content of a formal system, but *any* undecidability result is to be explained in this way. We must now investigate whether Chaitin's claim can be substantiated.

5.3.1 Complexity and incompleteness We first state precisely and prove Chaitin's version of the incompleteness theorem; a discussion follows in 5.3.2. We use Rogers' notation for partial recursive functions and recursively enumerable sets [86]: ϕ_n denotes the partial recursive function from $\mathbb{N}$ to $\mathbb{N}$ with Gödelnumber n and W_e denotes the r.e. subset of $\mathbb{N}$ with Gödelnumber e. As usual, we shall assume that sets such as $2^{<\omega}$ or $2^{<\omega}\times\omega$ etc. are coded into the natural numbers.

5.3.1.1 Lemma $\{<w,m> \in 2^{<\omega}\times\omega \mid I(w) \leq m\}$ is recursively enumerable.

Proof If U is the universal machine defined in 5.1, we have, using the definition of I, $\{<w,m> \in 2^{<\omega}\times\omega \mid I(w) \leq m\} = \{<w,m> \in 2^{<\omega}\times\omega \mid \exists p\,(U(p) = w \;\&\; |p| \leq m)\}$; the condition on the right hand side is $\sum_1$. □

Hence $\{<w,m> \in 2^{<\omega}\times\omega \mid I(w) > m\}$ is $\prod_1$; but it also satisfies a stronger property:

5.3.1.2 Definition (a) A set A is *immune* if it is infinite but contains no infinite recursively enumarable subset; (b) a set A is *effectively immune* if for some total recursive function $g: \omega \to \omega$: $W_e \subseteq A$ implies $\#W_e \leq g(e)$; (c) a set B is *(effectively) simple* if B is r.e. and B^c is (effectively) immune.

5.3.1.3 Theorem There exists a constant c such that any r.e. subset W_e of $\{<w,m> \in 2^{<\omega}\times\omega \mid I(w) > m\}$ is bounded in the second coordinate by $I(e) + c$.

Proof Although the result is stated for I only, it holds for a wide variety of complexity measures. To bring this out, we give an abstract proof. Let U be the universal prefix algorithm and define a partial recursive function f as follows. f operates on inputs of the form 0^n1q.

Given this input, f first calculates U(q); if and when it has found e = U(q), it generates W_e until it has found a pair $\langle w,m\rangle \in W_e$ such that $m > |q| + n + 1$; it then outputs w. Now suppose $W_e \subseteq \{\langle w,m\rangle \in 2^{<\omega}\times\omega \mid I(w) > m\}$. Apply the recursion theorem to get an n such that for all q, $\phi_n(q) \cong f(0^n1q)$. (That is, the left hand side is defined iff the right hand side is and when defined the two sides are equal.) Since f first calculates U(q) it is a prefix algorithm, hence so is ϕ_n. Let q_0 be such that $e = U(q_0)$; we claim that $\phi_n(q_0)$ is undefined. For suppose that $\phi_n(q_0) = w$. Then on the one hand, by construction,

(1) $I(w) > m > |q_0| + n + 1$;

on the other hand, since ϕ_n is a prefix algorithm,

(2) $I(w) \leq I_{\phi_n}(w) + n + 1 \leq |q_0| + n + 1$.

Hence $\phi_n(q_0)$ is undefined. It follows that I(e) + n + 1 is an upper bound for the second coordinate of W_e. To obtain a *recursive* upper bound, we can take any recursive upper bound for I(e), e.g. $2\log_2 e$: observe that $\sum_e e^{-2} < \infty$ and apply lemma 5.1.2.9. □

If we had used K instead of I, we could have dispensed with the demand that f on input 0^n1q first compute U(q); this condition was introduced only to ensure that f be a prefix algorithm. We first apply the theorem to obtain some recursion theoretic information on I.

5.3.1.4 Corollary Let $g: \omega \to \omega$ be total recursive and suppose that $\lim_{n\to\infty} g(n) = \infty$. Then $\{w \mid I(w) > g(|w|)\}$ is immune. In addition, if $\lim_{n\to\infty} g(n) = \infty$ recursively, then $\{w \mid I(w) > g(|w|)\}$ is effectively immune. We obtain the same results if we replace I by K.

Proof Let $W_e \subseteq \{w \mid I(w) > g(|w|)\}$. Put $V_e := \{\langle w,g(|w|)\rangle \mid w \in W_e\}$; then for some total recursive f, $V_e = W_{f(e)}$. Since $W_{f(e)} \subseteq \{\langle w,m\rangle \mid I(w) > m\}$, $W_{f(e)}$ is bounded in the second coordinate, e.g. by $2\log_2 f(e)$. But then, if $\lim_{n\to\infty} g(n) = \infty$, W_e must be finite and if $\lim_{n\to\infty} g(n) = \infty$ recursively, we can choose effectively $n_0(e)$ such that for $n \geq n_0(e)$, $2\log_2 f(n)$. In the latter case we therefore have $\#W_e \leq 2^{n_0(e)+1}$. □

It follows from the corollary that the r.e. relation $\{\langle w,m\rangle \mid I(w) \leq m\}$ is not recursive and likewise that the function $I: 2^{<\omega} \to \omega$ is not recursive. We also have:

5.3.1.5 Example The set of irregular strings $\{w \mid K(w) > |w| - m\}$ is effectively immune. By a theorem of Martin (see Soare [92,87]) it follows that $\{w \mid K(w) \leq |w| - m\}$ is a *complete* recursively enumerable set[4]. On the other hand, the arithmetical complexity of the set $\{w \mid I(w)$

$\leq |w| + I(|w|) - m\}$is higher (namely $\sum_2$), due to the presence of the term "I(|w|)".

We now formulate the first half of Chaitin's incompleteness theorem. Recall that for any natural number m all except finitely many w satisfy I(w) > m. We proved this in 5.1 using only elementary properties of finite sets; the proof can be formalized in any theory which contains a modicum of arithmetic. Nevertheless, as the following theorem shows, it is well nigh impossible to verify that some *specific string* has high complexity.

5.3.1.6 Theorem Let S be a sound formal system, identified with its r.e. set of theorems. Delete from S all theorems not of the form "I(w) > m" and call the resulting sound formal system S'. Let p be an r.e. index for S'. Then for some constant c, independent of S', and for all w: $S \nvdash I(w) > I(p) + c$.

Proof S' may be identified with an r.e. subset of $\{<w,m> \in 2^{<\omega} \times \omega \mid I(w) > m\}$with Gödelnumber p. By Theorem 5.3.1.3, S' is bounded in the second coordinate by I(p) + c, for some constant c not depending on p. □

Let us call the constant I(p) + c, which depends on S, the *characteristic constant* of the formal system S. We shall denote the characteristic constant as c(S). If we compare the preceding theorem with Chaitin's formulation, we see that what matters is not the complexity or information content of the formal system S, but only that of its reduced version S'. Indeed, we shall see below, in 5.3.2, that it can't be otherwise. Before we discuss Chaitin's claims, however, we shall prove the second half of the theorem announced above.

5.3.1.7 Theorem The sets $\{w \mid I(w) > k\}$ are r.e. and have indices p_k such that for some constant d independent of k, $I(p_k) \leq k + d$.

Proof (Sketched in Chaitin [13]) Obviously the sets $\{w \mid I(w) > k\}$, being the complements of finite sets, are r.e.; but Theorem 5.3.1.3 tells us that their indices are not recursive in k. Let W be a listing of all pairs <w,m> for which $I(w) \leq m$. Let P be a set of programs for the <w,m> in W such that every pair <w,m> in W is produced by exactly one p in P. P can be chosen to be r.e. Let U be the universal prefix algorithm.
Consider $P' := \{<p,m> \mid p \in P \ \& \ (U(p) = <w,m> \rightarrow I(w) \leq m)\}$. P' is r.e. and

$$\sum_{<p,m> \in P'} 2^{-m} = \sum_{\{w \mid I(w) \leq m\}} 2^{-m} \leq \sum_{w} 2^{-I(w)} \leq 1,$$

hence there exists a constant c such that for all p in P, if U(p) = <w,m>, then $I(p) \leq m + d$, by lemma 5.1.2.8. Now fix k and let p_k be a program in P for the *last* pair <w,k> in W. (Such a

program exists, although it cannot be found effectively.) Using the program p_k, we can enumerate all of $\{w| I(w) > k\}$: enumerate W until we come to the last pair $<w,k>$ (given by p_k); all w not occurring in this finite list must satisfy $I(w) > k$. We have seen above that $I(p_k) \leq k + d$. □

Observe that if c is the constant determined in Theorem 5.3.1.6, then $I(p_k) + c \geq k$, so that the preceding theorem is more or less the best possible result.

5.3.2 Discussion Theorem 5.3.1.6 implies that any formal system can verify the irregularity of at most a finite number of words. Alternatively, one could say that a Turing machine can produce only a finite number of irregular sequences. This result may be seen as a modern version of von Mises' conviction [67,60] "das man die "Existenz" von Kollektivs nicht durch eine analytische Konstruktion nachweisen kann" and it justifies to some extent the misgivings of those who maintain that randomness or irregularity cannot be formalized. But Theorem 5.3.1.6 is really much more than a formal statement of these intuitions: it expresses a precise connection between the information content of some formal system (namely S') and its "degree of incompleteness". We now discuss the question whether this theorem supports Chaitin's philosophical claims.

1. Although Theorem 5.3.1.6 was hailed as a "dramatic extension of Gödel's theorem"[5], we should not forget that there is a big difference between the two results. Gödel's first incompleteness theorem is an *explicit* construction of an undecidable (hence true) $\prod_1$ formula: the fixed point lemma [91,827] associates with any formal system S in a primitive recursive way a formula ψ_S which says of itself "I am unprovable in S". But Theorem 5.3.1.6 provides no such explicit construction. First, its proof shows that the characteristic constant c(S) is not a recursive function of S. Second, suppose we take some recursive upper bound f(S) for c(S), then it is still not possible to determine recursively a word w(S) such that $I(w(S)) > f(S) \geq c(S)$. If this were so, we could define an infinite r.e. sequence of

formal systems S_n and words $w(S_n)$ such that $I(w(S_n)) > f(S_n)$ and $\lim_{n\to\infty} f(S_n) = \infty$ as

follows: $S_0 = PA$, $S_1 = S_0 \cup \{I(w(S_0)) > f(S_0)\}$ etc. An examination of the construction of $c(S_n)$ (cf. Theorem 5.3.1.6 and its proof) shows that $\lim_{n\to\infty} c(S_n) = \infty$, hence also

$\lim_{n\to\infty} f(S_n) = \infty$. But corollary 5.3.1.4 implies that we can construct only finitely many

$w(S_n)$. Hence it is impossible to determine effectively, given a formal system S, a word w(S) such that $I(w(S)) > c(S)$. In *this* sense, Theorem 5.3.1.6 is a weak form, rather than an extension, of the first incompleteness theorem.

2. Furthermore, there is nothing in theorem 5.3.1.6 which supports Chaitin's claim that the undecidability of a formula can be explained as the result of an excess of information content. Observe that we said nothing about the *information content* of the *formula* "I(w) > c(S)" (for some specific w); all that mattered was that the undecidable formula *asserts* that some specific string contains too much information, which is something entirely different.
This being said, it must be acknowledged that *some* true statements are undecidable in PA precisely because they contain too much information.The construction of such a statement utilizes the fixed point lemma:

5.3.2.1 Lemma [91,827] Let ϕ be an arithmetical formula in one free variable. Then, for infinitely many ψ, PA $\vdash (\psi \leftrightarrow \phi(\ulcorner \psi \urcorner))$.

We use the fixed point lemma to define a sentence ψ which says intuitively "I contain too much information for PA". Put $k_0 := \max \{k |\ I(k) \leq c(PA)\}$. Choose (non-effectively!) ψ such that $\ulcorner \psi \urcorner > k_0$ and PA $\vdash (\psi \leftrightarrow I(\ulcorner \psi \urcorner) > c(PA))$. Then PA $\nvdash \psi$, since otherwise PA $\vdash I(\ulcorner \psi \urcorner) > c(PA)$, which is impossible by theorem 5.3.1.6; but ψ is true, for if $\neg\psi$ were true then $I(\ulcorner \psi \urcorner) \leq c(PA)$, which implies $\ulcorner \psi \urcorner \leq k_0$. Since PA is sound, PA $\nvdash \neg\psi$. Hence ψ is true but undecidable in PA. The construction is somewhat trivial, however, since we essentially use the fact that there exist fixed points of "$I(\ulcorner \psi \urcorner) > c(PA)$" with arbitrarily large Gödelnumber.

3. The preceding discussion showed that Chaitin's explanation of the incompleteness of formal systems: " I would like to be able to say that if one has ten pounds of axioms and a twenty-pound theorem, then the theorem cannot be derived from the axioms", is at present only scantly supported by the facts. But also his more modest claim, " Within ... a formal system a specific string cannot be proven to be of entropy [=complexity] greater than the entropy of the axioms of the theory" is not borne out by theorem 5.3.1.6. Recall that what mattered was not so much the information content of the formal system S as a whole, but rather that of its intersection S' with the set of statements of the form "I(w) > m". Of course there exists a primitive recursive function which brings us from S to S', and this justifies the notation "c(S)" for the characteristic constant of S. But since the information content of S', and not that of S, determines the characteristic constant of S, we cannot say that stronger theories lead to larger characteristic constants. Indeed, this is false, as we now show.
By theorem 11 in Kreisel–Levy [53,121], the arithmetical fragment of ZF is not finitely axiomatisable over PA. Theorem 5.3.1.6 assigns finite constants c(PA) and c(ZF) such that no statement "I(w) > c(PA)" ("I(w) > c(ZF)") is provable in PA (ZF). (Note that we do not even know whether c(ZF) > c(PA)!) It follows that an infinity of ever stronger number theories S_n, which lie in between PA and (the arithmetical fragment of) ZF must have the *same*

characteristic constant c and they must prove the *same* (finite) set of statements of the form "I(w) > m". Since I is unbounded on axioms for the S_n, the information contents of these axioms are totally irrelevant for the determination of c.

These considerations do not completely rule out the possibility that some kind of information concept is useful in studying incompleteness. They do show, however, that the *complexity* of the axioms is not a good measure of information. Furthermore, if the information is an *integer–valued* function and obeys something like theorem 5.3.1.6, then we must accept the consequence that a theory S_1 may be stronger than S_2, while having the same information content as S_2. It is difficult to imagine a concept of information which allows this possibility. The most reasonable way-out appears to be, to define a *rational-valued* (or real-valued) measure of information[6].

Even if the information concept turns out to be useless for the study of formal systems, it may be worthwhile to investigate what *other* properties of formal systems are relevant for the values of their characteristic constants. This investigation, however, is seriously hampered by the extreme scarcity of concrete examples: as noted above, we do not even know whether c(PA) < c(ZF)!

5.4 Infinite sequences: randomness and oscillations Two themes will occupy us in the present section. First, we try to express randomness (in the sense of Martin-Löf) in terms of the notions of complexity developed in 5.1.1 and 5.1.2. Now one might conjecture that the following generalisation (to infinite binary sequences) of the definition of irregularity (5.1.1.3): $\exists m\, \forall n\, K(x(n)) > n - m$, is an equivalent condition for randomness with respect to Lebesgue measure; but Martin-Löf has shown that no sequence x satisfies this generalisation. Similarly, no x satisfies $\exists m\, \forall n\, I(x(n)) > n + I(n) - m$, the natural generalisation of definition 5.1.2.6. But it turns out that membership of $R(\mu)$ can be characterised in terms of I, if we choose a smaller lower bound instead of one of the form $n + I(n) - m$. This brings us to the second topic: the oscillatory behaviour of the complexity measures K and I. Although this oscillatory behaviour is usually considered to be a nasty feature, we believe that it illustrates one of the great advantages of complexity: the possibility to study degrees of randomness.

5.4.1 Randomness and complexity Early attempts to characterize randomness with respect to some computable measure μ of an infinite binary sequence, in terms of a condition on the complexity of the initial segments of the sequence, foundered upon the following obstacle:

5.4.1.1 Theorem (Martin-Löf [61]) For all x and for all m, there are infinitely many n such that $K(x(n)) \leq n - m$. More precisely, if $f: \omega \rightarrow \omega$ is a total recursive function such that $\sum_n 2^{-}$

$^{f(n)} = \infty$, then for all x there are infinitely many n such that $K(x(n)) \leq n - f(n)$.

A simple proof of a special case, namely $f(n) := [a \cdot \log_2 n]$, with $a \in (0,1)$ computable, is given in Schnorr [88,110]. His proof can easily be adapted to show:

5.4.1.2 Lemma Let $a \in (0,1)$ be computable and let μ be a computable measure. For all x, there are infinitely many n such that $I(x(n)) \leq [-\log_2 \mu[x(n)]] + I(n) - [a \cdot \log_2 n]$. In particular, no x satisfies $\exists m \, \forall n \, I(x(n)) > [-\log_2 \mu[x(n)]] + I(n) - m$.

Martin-Löf's theorem was considered to be a surprising result. To quote from Schnorr [89,377]: "This fact is hard to comprehend and is the main obstacle for a common theory of finite and infinite random sequences". In retrospect, it is somewhat difficult to understand why Martin-Löf's theorem should be surprising. After all, results indicating that total chaos in infinite binary sequences is impossible were known already. One example is van der Waerden's theorem (from 1928), which states that if the natural numbers are partitioned into two classes, then at least one of these classes contains arithmetic progressions of arbitrary lengths[7]. Another example is a theorem in Feller [25,210] (cf. theorem 5.4.2.5 below) which states that if $a \in (0,1)$, then for μ_p–a.a. x, for infinitely many n, x_n is followed by a run of $[a \cdot \log_q n]$ 1's, where $q = p^{-1}$.

More important, the association between the oscillatory behaviour of K (or I) and the difficulty of characterising randomness in terms of complexity appears to be unfortunate. Thus, although Chaitin's I also oscillates (and for at least three essentially different reasons), it *is* possible to characterise randomness using I.

5.4.1.3 Theorem[8] Let μ be a computable measure. Then $x \in R(\mu)$ if and only if $\exists m \, \forall n \, I(x(n)) > [-\log_2 \mu[x(n)]] - m$.

Proof $\Rightarrow$ It suffices to show that $\{x \mid \forall m \, \exists n \, I(x(n)) \leq [-\log_2 \mu[x(n)]] - m\}$ is a recursive sequential test with respect to μ. By lemma 5.3.1.1, this set is $\prod_2$. We therefore have to show that $\mu\{x \mid \exists n \, I(x(n)) \leq [-\log_2 \mu[x(n)]] - m\} \leq 2^{-m}$ for each m. We may write

$$\mu\{x \mid \exists n \, I(x(n)) \leq [-\log_2 \mu[x(n)]] - m\} \leq \sum \{\mu[w] \mid w \in 2^{<\omega}, \, I(w) \leq [-\log_2 \mu[w]] - m\};$$

however, since $I(w) \leq [-\log_2 \mu[w]] - m$ iff $\mu[w] \leq 2^{-m} \cdot 2^{-I(w)}$, the right hand side of the above inequality is less than or equal to

$$\sum \{2^{-m} \cdot 2^{-I(w)} \mid w \in 2^{<\omega}, \, I(w) \leq [-\log_2 \mu[w]] - m\} \text{ and since } \sum_{w \in 2^{<\omega}} 2^{-I(w)} \leq 1, \text{ this is} \leq 2^{-m}.$$

$\Leftarrow$ Let $U = \bigcap_m U_m$ be the universal recursive sequential test with respect to μ. We may suppose $U_m = [T_m]$, with T_m prefixfree; hence $\mu U_m = \sum\{\mu[w] \mid w \in T_m\} \leq 2^{-m}$. Define $S := \{\langle w, [-\log_2\mu[w]] - \frac{1}{2}m\rangle \mid w \in T_m\}$. We show that $\sum\{2^{-k} \mid \exists w\, (\langle w,k\rangle \in S)\} < \infty$:

$$\sum_m \sum_{w\in T_m} 2^{[-\log_2\mu[w]]+\frac{1}{2}m} \leq \sum_m \sum_{w\in T_m} 2^{\frac{1}{2}m}\cdot\mu[w] = \sum_m 2^{\frac{1}{2}m}\cdot\mu U_m \leq \sum_m 2^{-\frac{1}{2}m} < \infty.$$

By lemma 5.1.2.8, we get for some constant c and all m and w: if $w \in T_m$, then I(w) is less than or equal to $[-\log_2\mu[w]] - \frac{1}{2}m + c$. In particular, if $x \in U$, then $\forall m\, \exists n\, (x(n) \in T_m)$, hence $\forall m\, \exists n\, (I(x(n)) \leq [-\log_2\mu[w]] - \frac{1}{2}m + c)$.
In other words, if $\exists m\, \forall n\, (I(x(n)) > [-\log_2\mu[w]] - \frac{1}{2}m + c)$, then $x \in R(\mu)$; but the antecedent is equivalent to $\exists m\, \forall n\, (I(x(n)) > [-\log_2\mu[w]] - m)$. □

The significance of this result has already been discussed in 5.2. The essence of the proof consists in the observation that randomness in the sense of Martin-Löf is a negative condition: x is random if it is not rejected at arbitrarily small levels of significance by the universal test U. Now U, conceived of as a r.e. set of finite sequences (namely $\bigcup_m T_m$), contains only elements of low complexity; hence for an infinite sequence to be random it is necessary and sufficient if it has no (except perhaps finitely many) initial segments of low complexity. In other words, *any* complexity measure C is able to characterise Martin-Löf randomness if the universal sequential test can be written in terms of C. Nothing more is necessary, but much more is possible. The *monotone complexity* of Schnorr [89] and Levin [54] developed in response to theorem 5.4.1.1 (see 5.4.4) also characterises randomness; but whereas I adds fine structure to the theory of random sequences (see 5.4.2–3), monotone complexity does not and we consider this to be a disadvantage.

5.4.2 Downward oscillations We now investigate more closely why the seemingly more reasonable condition of randomness $\exists m\, \forall n\, I(x(n)) > [-\log_2\mu[x(n)]] + I(n) - m$ is impossible. Not only doesn't this condition characterize randomness, it even cannot be satisfied by *any* sequence. Interestingly, this is true for several very different reasons and in this section we shall examine some of them. Martin-Löf's theorem 5.4.1.1 (and the simple version of it given as lemma 5.4.1.2) essentially use only the fact that $2^{<\omega}$ has a recursive enumeration. Below, we present two more derivations of Martin-Löf's theorem, the first based on the observation that Δ_2 definable sequences, even when random, have low complexity and the second elaborating the ancient idea that the existence of statistical regularities is incompatible with total chaos. For ease of notation, we consider Lebesgue measure only.

We first investigate the complexity of simply definable infinite binary sequences.

5.4.2.1 Lemma Let x be recursive, then for some c and all n, $I(x(n)) \leq I(n) + c$.

Proof Let A be an algorithm such that $A(n) = x(n)$ for all n. Define B as follows. On input q, it calculates U(q). If and when U halts on q, B computes $A(U(q)) = x(U(q))$ and outputs this sequence. B is a prefix algorithm, hence $I(x(n)) \leq I(n) + \lceil B \rceil + 1$. □

We now turn to Δ_2 definable sequences. The conditional complexity I_0 was defined in 5.1.3.

5.4.2.2 Theorem If x is Δ_2 definable, then $\lim_{n\to\infty} (n - I_0(x(n)|n)) = \infty$.

(As it stands the theorem is of course interesting only for $x \in R(\lambda)$.)

Proof By the modulus lemma (theorem 3.2.2.4), x can be written as: $x_n = \lim_{k\to\infty} \xi^k_n$, where $\xi^k \in 2^\omega$ such that $\{<k,n> | \xi^k_n = 1\}$ is recursive.

Define a prefix algorithm A as follows. Let A be given n on its worktape and q as input. On being presented with q, A first scans an initial segment s of q until it has determined an integer $i = U(s)$; it then calculates $n - i$, scans the remainder p of q, calculates $U(p,n-i)$ and outputs

$$A(q,n) = \xi^n(i)U(p,n-i).$$

For fixed i, if n is large enough, A(q,n) is of the form

$$A(q,n) = x(i)w.$$

Then there exist constants c,d such that $I_0(x(n)|n) \leq (I_A)_0(x(n)|n) + c \leq$
$\leq I(i) + I_0(x_{i+1}.....x_n|n-i) + d \leq I(i) + n - i + d$. Then $n - I_0(x(n)|n) \geq n - (n - i) - I(i) - d =$
$= i - I(i) + d$. In other words

$$\forall i\, \exists n_0(i)\, \forall n \geq n_0(i)\, (n - I_0(x(n)|n) \geq i - I(i) + d).$$

Because the right hand side is unbounded, $\lim_{n\to\infty} (n - I_0(x(n)|n)) = \infty$. □

5.4.2.3 Corollary If x is Δ_2 definable, then $\lim_{n\to\infty} (n + I(n) - I(x(n))) = \infty$.

Proof By lemmas 5.1.3.4/6, $I(x(n)) \leq I_0(x(n)|n) + I(n)$. □

The corollary is most likely not the best possible result; we used the estimate $I(x(n)) \leq I_0(x(n)|n) + I(n)$, which is far from being sharp (lemma 5.1.3.5). We conjecture that at least for *low* degrees x, i.e. x with $x' \equiv_T \emptyset'$, even $I(x(n)) \leq n + c$. Anyway, the result obtained just now will do for our purposes.

5.4.2.4 Theorem For all x: $\forall m\, \exists n \geq m\, (I(x(n)) < n + I(n) - m)$.

Proof We use the Basis Theorem (3.2.2.2). Suppose the theorem is false, then for some m, $\{x| \forall n \geq m\, (I(x(n)) \geq n + I(n) - m)\} \neq \emptyset$. This set is not itself $\prod_1$, but may be shown to be included in a set of the form $\{x| \forall n \geq m\, (I_0(x(n)) \geq n - c)\}$, which *is* $\prod_1$.
Indeed, by lemma 5.1.3.5, for some constant d, $I(x(n)) \leq I(x(n))|n) + I(n) + d$, hence the condition $I(x(n)) \geq n + I(n) - m$ can be rewritten as $I(x(n)|n) \geq n - c$. Now apply lemma 5.1.3.4, which says that $I_0(x(n)|n)$ is (much) larger than $I(x(n)|n)$.
It follows that the $\prod_1$ set $\{x| \forall n \geq m\, (I_0(x(n)) \geq n - c)\}$ has a Δ_2 definable element x. But this is impossible in view of theorem 5.4.2.2. □

We now give a second proof of the above theorem, based on a different idea: that statistical regularities must lead to a decrease in complexity. We use an exercise in Feller [25].

5.4.2.5 Theorem (after Feller [25,210]) Let $N_n(x)$ denote the length of the run of 1's beginning at x_n. Then for all $x \in R(\lambda)$:

$$\limsup_{n\to\infty} \frac{N_n(x)}{\log_2 n} = 1.$$

Proof (1) Let $a > 1$ be computable. We have to show that $\{x| \forall m\, \exists n\, N_n(x) > a\cdot\log_2 n\}$ is a recursive sequential test with respect to λ. We use the first effective Borel–Cantelli lemma (3.3.1). Define $A_n := \{x| x_n$ is followed by $[a\cdot\log_2 n] + 1$ 1's$\}$. It suffices to show that $\sum_n \lambda A_n$ converges constructively. But this is so, since $\sum_n \lambda A_n \leq \sum_n n^{-a}$.
(2) Let $a < 1$ be computable. Since the set $\{x| \exists m\, \forall n\, N_n(x) < a\cdot\log_2 n\}$ is Σ_2, it suffices to show that it has Lebesgue measure 0. Define a total recursive function f by $f(n) := n + (n-1)\cdot[a\cdot\log_2 n]$. Then we have $f(n+1) - f(n) > [a\cdot\log_2 n]$.
Define $A_n := \{x| x_n$ is followed by $[a\cdot\log_2 n]$ 1's$\}$, then the $A_{f(n)}$ are independent. Because $\sum_n \lambda A_{f(n)} \geq \sum_n n^{-a}$ diverges for $a < 1$, the second Borel–Cantelli lemma (3.3.2) gives the desired result. □

5.4.2.6 Corollary Let $a \in (0,1)$ be computable. Define $b_n := n + [a\cdot\log_2 n]$. Then for some constant c, for all $x \in R(\lambda)$: for infinitely many n, $I(x(b_n)) \leq b_n + I(b_n) - [a\cdot\log_2 n] + c$.

Proof Define a prefix algorithm A(s,k) as follows. A first solves the equation $k = b_n$ for n. If it has succeeded in doing so, it computes U(s) and when this computation terminates, it outputs

$$A(s,k) = U(s)1^{[a\cdot\log_2 n]}.$$

It follows that, for $x(b_n) = x(n)1^{[a\cdot\log_2 n]}$, $I(x(b_n)) \leq I(x(n)) + \ulcorner A \urcorner + 1 \leq n + I(n) + d =$

$= b_n + I(b_n) - [a\cdot\log_2 n] + c$, for some constants c and d. Now apply theorem 5.4.2.5. □

5.4.2.7 Corollary For all x and for all m there are infinitely many n such that $I(x(n)) \leq \leq n + I(n) - m$.

Proof If $x \notin R(\lambda)$, the result follows from theorem 5.4.1.3. If $x \in R(\lambda)$, apply corollary 5.4.2.6. □

With corollary 5.4.2.6 at our disposal, we may understand the often repeated query: "How can a random sequence exhibit statistical regularities, since randomness entails the *absence* of regularities?" In a sense, the implied objection is right; we might even say that it is illustrated by the failure of the putative definition of irregularity $\exists m \forall n\, I(x(n)) > n + I(n) - m$.
This definition turned out to be impossible because a statistical regularity brought about a decrease of I (although this is not the *only* source of downward oscillations of I). We see, however, that some regularities are more regular than others; in particular, statistical regularities are not simple, that is, they do not lead to a *significant* decrease in complexity.
We may also observe that there are essentially different reasons why total chaos in infinite binary sequences is impossible: Martin-Löf's 5.4.1.1 (or Schnorr's 5.4.1.2) uses in essence only the fact that $2^{<\omega}$ is recursive, whereas our theorem 5.4.2.4, although also of a recursion–theoretic character, uses some less trivial facts about the arithmetical hierarchy. Corollary 5.4.2.6 is of a different nature altogether and depends on statistical properties of product measures.

5.4.3 Upward oscillations We now prove some results which show that the behaviour of I on Δ_2 definable sequences is rather atypical: for most sequences x, I(x(n)) comes close to its theoretical upper bound infinitely often. Our method of proof again involves Turing degrees. In 5.4.2 we derived the existence of downward oscillations from the fact that the degrees between Ø and Ø' have low information content; we derive the existence of upward oscillations from the fact that the degrees above (and including) Ø', the so called *complete* Turing degrees, have high information content.
We use "high information content" in the following sense. Let y be an infinite binary sequence and let I^y be defined as I, except that we allow functions *partial recursive in* y, instead of partial recursive functions only. Clearly, for all w: $I^y(w) \leq I(w)$.The following theorem shows that if y is a complete Turing degree, then for most x, the difference $I(x(n)) - I^y(x(n))$ is large infinitely often, indicating that y contains some information about most x. We use $\equiv_T$ to denote Turing equivalence and $\leq_T$, $\geq_T$ to denote Turing reducibility.

5.4.3.1 Theorem Let $y \geq_T \emptyset'$ and let $g: \omega \to \omega$ be a total recursive function such that $\sum_n 2^{-g(n)}$ diverges. Then (*) $\lambda\{x \mid \forall m\, \exists n \geq m\, (I^y(x(n)) < I(x(n)) - g(n))\} = 1$.

Proof Since for some c and all w, $I^y(w) \leq |w| + I^y(|w|) + c$, it suffices to prove that

$$\lambda\{x \mid \forall m\, \exists n \geq m\, (n + I^y(n) + c < I(x(n)) - g(n))\} = 1.$$

We absorb c into g. We show that for each m, $\lambda\{x \mid \forall n \geq m\, (n + I^y(n) + g(n) < I(x(n))\} = 0$.

Observe that for each n, the measure of this set is smaller than

$$\sum_n \{2^{-|w|} \mid w \in 2^n,\ I(w) \leq n + g(n) + I^y(n)\}.$$

Now the number of $w \in 2^n$ satisfying $I(w) \leq n + g(n) + I^y(n) = n + I(n) - (I(n) - g(n) - I^y(n))$

is less than or equal to $2^{n-(I(n)-g(n)-I^y(n))} \cdot d$, for some constant d (lemma 5.1.2.4).

It follows that for each n, the required measure is smaller than $2^{-(I(n)-g(n)-I^y(n))} \cdot d$;

and we have to show that $\forall k\, \exists n \geq k\, (I(n) - g(n) - I^y(n) > k)$.

Now the function f_k defined by $f_k(n) := I(n) - g(n) - k$ is recursive in y since $\emptyset' \leq_T y$ and unbounded since $\sum_n 2^{-g(n)}$ diverges (by lemma 5.1.2.9). Relativizing the definition of *immunity* (5.3.1.2) to y, we see that the set $\{n \mid I^y(n) \geq f_k(n)\}$ must be y–immune for each k. Hence for each k, $\{n \mid n \geq k\} \not\subset \{n \mid I(n) - g(n) - I^y(n) \leq k\}$; in other words, for all k there is some n larger than k for which $I(n) - g(n) - I^y(n) > k$. □

The assumption that $\emptyset' \leq_T y$ is essential for the proof, since for some f_k, we may have $f_k \equiv_T \emptyset'$. We conjecture that condition (*) in fact characterizes the complete Turing degrees. In any case the results of 5.4.2 and 5.4.3 indicate that it may be profitable to study the Turing degrees using complexity measures.

In conjunction with theorem 5.4.1.3 (with "I^y" replacing "I")), the preceding theorem immediately implies:

let $g: \omega \to \omega$ be a total recursive function such that $\sum_n 2^{-g(n)}$ diverges; then $\lambda\{x \mid \exists k\, \forall m\, \exists n \geq m\, (I(x(n)) > n + g(n) - k)\} = 1$.

Using the following lemma due to Chaitin, we can do slightly better:

5.4.3.2 Lemma (Chaitin [12,337]) $\lambda\{x \mid \exists m\, \forall n \geq m\, I(x(n)) > n\} = 1$.

Proof By the first Borel–Cantelli lemma, it suffices to show that $\sum_n \lambda\{x \mid I(x(n)) > n\} < \infty$. But this is so, since $\lambda\{x \mid I(x(n)) > n\} \leq 2^{-I(n)} \cdot c$ by lemma 5.1.2.4. □

5.4.3.3 Corollary (Solovay) Let g: $\omega \to \omega$ be a total recursive function such that $\sum_n 2^{-g(n)}$ diverges; then $\lambda\{x| \forall m\, \exists n \geq m\, (\, I(x(n)) > n + g(n))\} = 1$.

The following observation is also due to Solovay (both results are announced, without proof, in Chaitin [13]).

5.4.3.4 Theorem $\lambda\{x| \exists m\, \forall k\, \exists n \geq k\, (\, I(x(n)) > n + I(n) - m)\} = 1$.

Proof It obviously suffices to show that for some c and all m,

$$\lambda\{x| \exists k\, \forall n \geq k\, (\, I(x(n)) \leq n + I(n) - m)\} \leq 2^{-m} \cdot c.$$

But the collection $\{\{x| \forall n \geq k\, (\, I(x(n)) \leq n + I(n) - m)\}| k \in \omega\}$ is increasing in k and, for all n, $\lambda\{x| \forall n \geq k\, (\, I(x(n)) \leq n + I(n) - m)\} \leq \sum_n \{2^{-|w|} | w \in 2^n, I(w) \leq n + I(n) - m\} \leq 2^{-m} \cdot c$ by lemma 5.1.2.4. □

It follows from this theorem that the behaviour of Δ_2 definable sequences, for which we could show $\lim_{n\to\infty} (n + I(n) - I(x(n))) = \infty$, is not typical of arbitrary random sequences.

5.4.4 Digression: monotone complexity We saw in 5.4.1 that, according to Schnorr [89], the difficulties encountered in characterising randomness in terms of K, were due to K's oscillatory behaviour. In response to Martin-Löf's theorem 5.4.1.1, he (and independently Levin [54]) developed a notion of complexity which does not oscillate on random sequences. The new notion, so called *monotone complexity*, is again obtained by restricting the class of algorithms. Schnorr considers *monotone* algorithms, i.e. those partial recursive functions A such that $v \subseteq w$ implies $A(v) \subseteq A(w)$. The set of monotone algorithms is recursively enumerable[9], so we may define a universal monotone algorithm U by $U(0^{\lceil A \rceil}1p) = A(p)$. Let KM denote the resulting concept of complexity. Schnorr [89,380] proves

$$x \in R(\lambda) \text{ if and only if } \exists c\, \forall n\, |KM(x(n)) - n| \leq c;$$

and generally (see Gacs [32])

$$x \in R(\mu) \text{ if and only if } \exists c\, \forall n\, |KM(x(n)) - [-\log_2 \mu[x(n)]]| \leq c.$$

This is obviously in sharp contrast with the behaviour of I. The lower bound is the same (and the proof follows very much the same lines), but the upper bound is not, and this is due to the fact that the identical function F(w) = w is a monotone algorithm, but not a prefix algorithm: since F is monotone, we have $KM(w) \leq |w| + \lceil F \rceil + 1$. (In general, every prefix algorithm is a

monotone algorithm, but not conversely.) However, the only effect of lowering the upper bound is, that KM obliterates distinctions which I is able to make. For instance, consider the algorithm A defined in the proof of corollary 5.4.2.6; define B similarly but with the universal monotone algorithm replacing the universal prefix algorithm. B *is not a monotone algorithm*, whereas A is. The operation of suffixing words with strings of 1's is not monotone, except when the domain of the suffixing algorithm is prefixfree; in other words, when the suffixing algorithm is like A. But $KM << KM_A$, so KM doesn't see these regularities.

Thus, although a characterisation of randomness in terms of KM can be given, this is where its utility stops. Using I, we can learn something about random sequences over and above the fact that they satisfy Martin-Löf's definition; it suggests questions such as "Does the complexity of easily definable random sequences differ from the complexity of those which are not?", a question which has only a trivial answer for KM. Historically, complexity oscillations have earned their bad repute from the apparent impossibility of characterising randomness in terms of complexity. Now that such a characterisation has been given, we see that oscillations need not be feared. In fact, if a (downward) oscillation occurs, then, in accordance with the motivation given in 5.1, we must accept the presence of a temporary regularity. These regularities do not vanish the moment we decide to adopt a different complexity measure, to wit, monotone complexity.

5.5 Complexity and entropy Two problems will occupy us in this section. The first is to explain the meaning of the phrases "topological aspect of I" and "metric aspect of I", used in 5.1.4. The second is to link I, which is a measure of disorder for *sequences*, with more traditional measures of chaotic behaviour, defined for *dynamical systems*, such as (metric or topological) entropy. This problem has received some attention in the physics literature (see Ford 27], Lichtenberg and Lieberman [58], Alekseev and Yakobson [2] and Brudno [10]), in connection with research on chaotic dynamical systems. It is shown here (theorem 5.5.2.5) that if μ is an ergodic measure, then μ-a.a. x satisfy

$$\lim_{n\to\infty} \frac{I(x(n))}{n} = H(\mu),$$

where $H(\mu)$ is the metric entropy of μ. We use theorem 5.5.2.5 to elucidate the metric aspect of I in terms of (un)predictability.

We then proceed to an investigation of the relation between $\mathbf{E}(A)$, the topological entropy of a $\prod_1$ set A, and the behaviour of I on sequences x in A. It is shown that A must satisfy special conditions (A must be "homogeneous") if there are to be many sequences in A with

$$\lim_{n\to\infty} \frac{I(x(n))}{n} = \mathbf{E}(A).$$

Lastly, we compare I with another measure of randomness for sequences, viz. Kamae–

entropy.

5.5.1 Dynamical systems Our set–up is as follows. A *symbolic dynamical system* on a set of symbols $n = \{0,...,n-1\}$ is a set $X \subseteq n^{\omega}$ (or $n^{\mathbb{Z}}$, as the case may be), together with the left–shift (or two–sided shift) T. We assume that X is closed under the action of T. Symbolic dynamical systems arise naturally in the study of general dynamical systems, in the following way.

Suppose (Γ,S) is a dynamical system, where Γ can be thought of as a phase space, equipped with a σ–algebra of measurable sets, and S is a measurable transformation on Γ, which represents the evolution of the system, considered in discrete time. A measurement with finite accuracy on (Γ,S) is represented (ideally) by a measurable partition $A_0,...A_{n-1}$ of Γ, corresponding to "pointer readings" 0,...,n–1.
Define a mapping $\psi: \Gamma \to n^{\omega}$ by $\psi(\gamma)_k = i$ iff $S^k(\gamma) \in A_i$; then $\psi(\gamma)$ represents the sequence of pointer readings obtained upon repeatedly measuring $\{A_0,...A_{n-1}\}$ on a system which is in state γ at time t = 0.
If the system (Γ,S) is also equipped with a probability distribution P, this distribution generates a measure μ on n^{ω} by $\mu A := P\psi^{-1}A$.
One may now study the dynamical system (Γ,S,P) by means of its symbolic representative $(\psi[\Gamma],T,\mu)$. In particular, the question whether, and to what extent, (Γ,S,P) displays chaotic behaviour can be investigated in this way. Below, we introduce various measures of disorder directly for *symbolic* dynamical systems, where for notational convenience we assume that the alphabet consists of just two symbols, 0 and 1. For an overview of the theory of dynamical systems, the reader may consult Petersen [82].

5.5.2 Metric entropy Let μ be a *stationary* measure on 2^{ω}; that is, for all Borel sets A, μ satisfies $\mu T^{-1}A = \mu A$. In other words, T conserves μ. For such measures, we may define the *metric entropy* $H(\mu)$ as follows:

5.5.2.1 Definition Let μ be a *stationary* measure on 2^{ω}. The metric entropy $H(\mu)$ of μ is defined to be $H(\mu) := \lim_{n\to\infty} -\frac{1}{n}\sum_{w\in 2^n} \mu[w]\log_2\mu[w]$. (Petersen [82,240])

5.5.2.2 Example It is easy to verify that $H(\mu_p)$ equals $-p\log_2 p - (1-p)\log_2(1-p)$.

The interpretation of $H(\mu)$ is roughly as follows. $w \in 2^n$ is a possible series of outcomes if we perform n experiments upon the system under consideration. The probabilistic information present in w is (by definition) $-\log_2\mu[w]$; then

$$-\frac{1}{n}\sum_{w\in 2^n}\mu[w]\log_2\mu[w]$$

is the average amount of information gained per experiment if we perform n experiments. $H(\mu)$ is obtained if we let n go to infinity. A positive value of $H(\mu)$ indicates that each repetition of the experiment provides a non–negligable amount of information; systems with this property may be called random. Obviously, $H(\mu)$ is a global characteristic of the system $(2^\omega,T,\mu)$; it depends only on μ and T and reflects the the randomness of the system as a whole. We must now investigate how this global characteristic is related to randomness properties of individual sequences.

The measures occurring in 5.5.2 will be assumed to be *ergodic*; that is, if $T^{-1}A = A$, μA is either 0 or 1. If μ is ergodic, then $\mu[w]$ can be interpreted as the limiting relative frequency of w in a typical sequence x:

5.5.2.3 Ergodic theorem (see Petersen [82,30]) Let μ be a stationary measure on 2^ω, $f: 2^\omega \to \mathbb{R}$ integrable. Then

$$f^*(x) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(T^k x)$$

exists μ-a.e., f^* is T-invariant and $\int f d\mu = \int f^* d\mu$. In addition, if μ is ergodic then f^* is constant μ-a.e. As a consequence, if μ is ergodic, then for any $w \in 2^{<\omega}$:

$$\mu\{x \mid \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} 1_{[w]}(T^k x) = \mu[w]\} = 1.$$

Below, we use not only the ergodic theorem, but also one of its consequences, the Shannon–McMillan–Breiman theorem:

5.5.2.4 Theorem (see Petersen [82,261]) Let μ be an ergodic measure on 2^ω, $H(\mu)$ its entropy. Then for μ–a.a. x: $\lim_{n\to\infty} -\frac{\log_2\mu[x(n)]}{n} = H(\mu)$.

One immediate application of the Shannon–McMillan–Breiman theorem in this context is the computation of the constant H such that

$$\lim_{n\to\infty}\frac{I(x(n))}{n} = H \quad \mu\text{–a.e.}$$

We saw in 5.1.2 that this constant exists, due to the subadditivity of I; but we couldn't compute

it. However, at least for computable μ it is easy to see that H must equal $H(\mu)$. Combining lemma 5.1.4.3 and theorem 5.4.1.3, we get: $x \in R(\mu)$ if and only if $\exists m\, \forall n\, (m + I(n) + [-\log_2\mu[x(n)]] \geq I(x(n)) > [-\log_2\mu[x(n)]] - m)$. Since $\mu R(\mu) = 1$, the preceding theorem implies for μ–a.a.x: $\lim_{n\to\infty} \frac{I(x(n))}{n} = H(\mu)$ [10].

Hence for computable ergodic μ, the statement that $I(x(n))/n$ converges to $H(\mu)$ μ almost everywhere, is a trivial (and less informative) consequence of the characterization of randomness. For arbitrary ergodic μ, we must do some more work.

5.5.2.5 Theorem Let μ be an ergodic measure, $H(\mu)$ its entropy. Then for μ–a.a. x: $\lim_{n\to\infty} \frac{I(x(n))}{n} = H(\mu)$ [11].

Proof Stripped of its recursive content, the "$\Rightarrow$" half of theorem 5.4.2.3 shows that $\mu\{x \mid \forall m\, \exists n\ I(x(n)) > [-\log_2\mu[x(n)]] - m\} = 0$. Using theorem 5.5.2.4 it follows that $\liminf_{n\to\infty} \frac{I(x(n))}{n} \geq H(\mu)$, for μ–a.a. x. To get $\limsup_{n\to\infty} \frac{I(x(n))}{n} \leq H(\mu)$ for μ–a.a. x, we remark first that, for each x and for each k, $\limsup_{n\to\infty} \frac{I(x(n))}{n} = \limsup_{n\to\infty} \frac{I(x(n\cdot k))}{n\cdot k}$.

Indeed, by the subadditivity of I, there exists a constant c such that for all k: $I(x(n)) = I(x(n_0\cdot k + r)) \leq I(x(n_0\cdot k)) + I(x_{n_0\cdot k+1},\ldots,x_{n_0\cdot k+r}) + c$.

Clearly, then, $\limsup_{n\to\infty} \frac{I(x(n))}{n} \leq \limsup_{n\to\infty} \frac{I(x(n\cdot k))}{n\cdot k}$; the converse inequality istrivial.

We now use lemma 5.1.4.5, slightly rephrased:

$$I(x(n\cdot k)) \leq n\cdot\Big[-\sum_{w\in 2^k}\Big(\frac{1}{n}\sum_{j=1}^{n} 1_{[w]}(T^{j\cdot k}x)\log_2\Big(\frac{1}{n}\sum_{j=1}^{n} 1_{[w]}(T^{j\cdot k}x)\Big)\Big) + \frac{O(\log_2 n)}{n}\Big],$$

which implies

$$(*)\quad \frac{I(x(n\cdot k))}{n\cdot k} \leq -\frac{1}{k}\sum_{w\in 2^k}\Big(\frac{1}{n}\sum_{j=1}^{n} 1_{[w]}(T^{j\cdot k}x)\log_2\Big(\frac{1}{n}\sum_{j=1}^{n} 1_{[w]}(T^{j\cdot k}x)\Big)\Big) + \frac{O(\log_2 n)}{n\cdot k}.$$

Since μ is stationary (although not necessarily ergodic) with respect to the T^k, the ergodic theorem implies that $f_w(x) := \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} 1_{[w]}(T^{j\cdot k}x)$ exists μ–a.e. and that $\int f_w d\mu = \mu[w]$.

Taking limsups (with respect to n) and integrals (with respect to μ) on the left hand side and right hand side of (*), we get, for all k:

$\int \limsup_{n\to\infty} \frac{I(x(n\cdot k))}{n\cdot k}\, d\mu \le -\frac{1}{k}\sum_{w\in 2^k} \int f_w \log_2 f_w d\mu$; hence by Jensen's inequality

$$\int \limsup_{n\to\infty} \frac{I(x(n\cdot k))}{n\cdot k} d\mu \le -\frac{1}{k}\sum_{w\in 2^k} \int f_w d\mu \log_2 \int f_w d\mu = -\frac{1}{k}\sum_{w\in 2^k} \mu[w]\log_2 \mu[w].$$

Since $\limsup_{n\to\infty} \frac{I(x(n\cdot k))}{n\cdot k} = \limsup_{n\to\infty} \frac{I(x(n))}{n}$, we have, for each k: $\int \limsup_{n\to\infty} \frac{I(x(n))}{n} d\mu \le$

$-\frac{1}{k}\sum_{w\in 2^k} \mu[w]\log_2 \mu[w]$. Letting k go to infinity, we see that $\int \limsup_{n\to\infty} \frac{I(x(n))}{n} \le H(\mu)$ and

the desired result follows since $\limsup_{n\to\infty} \frac{I(x(n))}{n}$ is T–invariant, hence constant μ–a.e. □

5.5.2.6 Remark Use of Solovay's formula (5.1.2.10) immediately gives $\lim_{n\to\infty} \frac{K(x(n))}{n} =$

$= H(\mu)$ μ–a.e., but employing I instead of K reduces one half of the proof to a triviality.

We now interpret the preceding theorem as a result on the amount of computer power necessary to predict the outcome sequence x(n), given x(m), where m < n. This problem arises for instance in the study of dynamical systems (Γ,S) on which we perform a measurement given by the partition $A_0,...A_{k-1}$: we have observed the state of the system (i.e. one of the numbers 0,...,k–1) at instants t = 1,...,m and we wish to predict the state at instants t = m+1,...,n.

To calculate x(n) from x(m) we may use the evolution S, but other algorithms are also allowed. We impose but one restriction: the algorithm should not be too large. So we fix some constant c (representing the size of a program too large for practical purposes) and we call x(n) *unpredictable* given x(m) if $I(x(n)|x(m)) > c + I(n)$, or, what comes down to the same thing (by lemma 5.1.3.6), if $I(x(n)|<n,x(m)>) > c$. (We use as conditions both x(m) and n, since the instant n chosen in advance also belongs to the data.) The term *unpredictable* is used here in the sense of *not potentially predictable*.

We now show that there exists a close connection between entropy and unpredictability. Since c has been chosen so large, we may write the following chain of equivalent inequalities:

$I(x(n)|x(m)) > c + I(n) \Leftrightarrow$

$I(x(n)|x(m)) + I(x(m)) > c + I(n) + I(x(m)) \Leftrightarrow$ (by lemma 5.1.3.6)

$I(<x(n),x(m)>) > c + I(n) + I(x(m)) \Leftrightarrow$ (since m and x(n) determine x(m))

$$I(x(n)) + I(m) > c + I(n) + I(x(m)) \Leftrightarrow$$

$$(*) \quad I(x(n)) > c + I(n) + I(x(m)) - I(m).$$

Since $I(x(m)) \leq m + I(m) + d$, with $d \ll c$, (*) surely holds if $I(x(n)) > c + m + I(n)$.

Now let μ be an ergodic measure with entropy $H(\mu)$ and suppose $\lim_{n\to\infty} \frac{I(x(n))}{n} = H(\mu)$.

Assume $H(\mu) > 0$, choose $\varepsilon > 0$ small compared to $H(\mu)$ and let n_0 be so large that $I(x(n)) > n(H(\mu) - \varepsilon)$ for $n \geq n_0$.

Then (*) is surely satisfied if $n > \frac{c+m+I(n)}{H(\mu)-\varepsilon}$, an inequality which can thus be taken as a sufficient condition for unpredictability.

Note that this condition can be significantly improved if we assume in addition that μ is computable. In this case we may replace the upper bound $I(x(m)) \leq m + I(m) + d$ by $I(x(m)) \leq [-\log_2\mu[x(m)]] + I(m) + d$. By the Shannon–McMillan–Breiman theorem (5.5.2.4), there is $m_0(\varepsilon)$ such that for $m \geq m_0(\varepsilon)$: $[-\log_2\mu[x(m)]] \leq m(H(\mu) + \varepsilon)$. For suitable choices of n and m the above sufficient condition for unpredictability can thus be sharpened to:

$$n > \frac{c + m(H(\mu) + \varepsilon) + I(n)}{H(\mu) - \varepsilon}.$$

If $\varepsilon \ll H(\mu)$, then this boils down to: $n > m + \frac{c + I(n)}{H(\mu)}$.

In other words, the complexity theoretic characterisation of randomness shows that random sequences have a definite "predictability horizon", which is approximately (modulo the term I(n), which is small compared to n) linear in the data x(m).

5.5.3 Topological entropy Like metric entropy, topological entropy is a global measure of disorder, pertaining to the dynamical system as a whole, not to individual trajectories. Again our main interest concerns the relation between this global measure and the behaviour of I.

5.5.3.1 Definition Let $A \subseteq 2^\omega$ be closed. Call $w \in 2^n$ *admissible for* A if $A \cap [w] \neq \emptyset$. Put $A_n := \{w \in 2^n \mid w \text{ admissible for } A\}$. $\#A_n$ denotes the cardinality of A_n.

5.5.3.2 Definition Let $A \subseteq 2^\omega$ be closed. **E**(A), the *topological entropy* of A is defined to be $\mathbf{E}(A) := \limsup_{n\to\infty} \frac{\log_2 \#A_n}{n}$.

5.5.3.3 Remark If A is shift–invariant, i.e. if $T^{-1}A = A$, where T is the left–shift, we have in fact $\mathbf{E}(A) = \lim_{n\to\infty} \frac{\log_2 \#A_n}{n}$. This is so, for instance, if A is of the form $\psi[\Gamma]$, where ψ and Γ are as in 5.5.1. In this case, $\mathbf{E}(A)$ measures the extent to which the transformation S on Γ scatters points around Γ. It may be of interest to note that for shift–invariant A, $\mathbf{E}(A)$ equals the Hausdorff dimension of A.

5.5.3.4 Example Let A consist of all those infinite binary sequences in which maximal blocks of 0's and of 1's have even length. Clearly $\#A_{2n} = 2^n$, hence $\mathbf{E}(A) = \frac{1}{2}$.

The calculation of topological entropy is sometimes made difficult by the circumstance that the set of admissible words for a $\prod_1$ set A need not be recursive, as it was in the example just given. For instance, if A is a $\prod_1$ set without recursive elements (one may think of the set of complete consistent extensions of Peano arithmetic; or the set $A = \{x | \forall n\ V(x(n)) \leq m\}$ where V is a universal subcomputable Martingale (cf. 3.4)), then its set of admissible words cannot be recursive, for if it were, the leftmost infinite branch would also be recursive. (We conjecture that in fact the following holds: if A is $\prod_1$ without recursive elements, then $\mathbf{E}(A) =$ 0,1 or non–computable.)

5.5.3.5 Lemma Let $A \subseteq 2^\omega$ be $\prod_1$. The set of admissible words for A is $\prod_1$.

Proof By König's lemma, w is admissible for A iff $\forall n \geq |w|\ \exists v \in 2^n\ (v \in T\ \&\ w \subseteq v)$, where T is the recursive binary tree associated with A. □

The relation between topological and metric entropy is given by

5.5.3.6 Variational principle (Petersen [82,269]) Let $A \subseteq 2^\omega$ be shift–invariant and closed. Then $\mathbf{E}(A) = \sup\{H(\mu) | \mu \text{ stationary measure on } A\}$.

A measure μ on A for which in fact $\mathbf{E}(A) = H(\mu)$ is called a *maximum entropy* measure (e.g. λ is the maximum entropy measure on 2^ω).

At last, we may now discuss the relation between complexity and topological entropy. In order to see what kind of relation can be expected, let us first derive some simple consequences of the material presented so far.

5.5.3.7 Lemma Let $A \subseteq 2^\omega$ be $\prod_1$ with a recursive set of admissible words. Then for all

x in A: $\limsup_{n\to\infty} \frac{I(x(n))}{n} \leq \mathbf{E}(A)$.

Proof Since the set of admissible words is Δ_1, we have by lemma 5.1.4.2, for $w \in A_n$, $I(w) \leq [\log_2 \#A_n] + I(|w|) + d$. Hence also for all n, $x \in A_n$: $I(x(n)) \leq [\log_2 \#A_n] + I(n) + d$, and the result follows since $\lim_{n\to\infty} \frac{I(n)}{n} = 0$. □

5.5.3.8 Lemma Let $A \subseteq 2^\omega$ be shift–invariant, μ a stationary measure on A. Then $\mu\{x \in A \mid \lim_{n\to\infty} \frac{I(x(n))}{n} \leq \mathbf{E}(A)\} = 1$.

Proof By theorem 5.5.2.5, the limit equals $H(\mu)$ μ–a.e. By the variational principle, $H(\mu) \leq \mathbf{E}(A)$. □

These results show that $\mathbf{E}(A)$ is in some interesting cases an upper bound for $\limsup_{n\to\infty} \frac{I(x(n))}{n}$.

Now obviously, if μ is a maximum entropy measure for (A,T), then "≤" can be replaced by "=" in 5.5.3.8.

But one would like to know whether,without special assumptions (such as shift–invariance) about A, there exist x in A for which $\lim_{n\to\infty} (\sup) \frac{I(x(n))}{n} = \mathbf{E}(A)$, and if so, how many.

A little reflection shows, that the condition "$\lim_{n\to\infty} (\sup) \frac{I(x(n))}{n} = \mathbf{E}(A)$" implies something about the structure of A; and this becomes particularly clear when we consider the slightly stronger form "$\exists m\ \forall n\ I(x(n)) > [-\log_2 \#A_n] - m$", the topological analogue of the criterion for randomness. In fact, this topological analogue seems to embody the pure form of irregularity or lawlessness; irregularity which does not necessarily imply statistical regularity. The condition roughly means the following (cf. 5.1.4). We are given a $\prod_1$ set A, which determines a priori restrictions on our freedom to choose x(n). For each n, we may choose among $\#A_n$ possibilities to determine x(n). Obviously, once x(n) has been chosen, there is not much freedom to choose x(n+1); but we are entirely free in choosing a *program* for x(n+1). Bearing in mind that, at least when A has a recursive set of admissible words, the upper bound for I(x(n)) is of the form $[-\log_2 \#A_n] + I(n) + d$, the condition for topological irregularity means by and large (modulo the unavoidable oscillations) (1) that a program for x(n) is of the form "program for n plus ordinal number of x(n) in A_n" and (2) that we need the *full range of*

possibilities in the A_n in order to determine x, so that we have not restricted our freedom of choice more than demanded by the a priori restrictions imposed by A. This seems to be a pleasant way of saying what irregularity or lawlessness means in a classical setting.

But we only *need* the full range of possibilities in the A_n if it is not possible to restrict the freedom of choice significantly (as measured on the logarithmic scale) by specifying, say, a finite number of bits in advance. These considerations suggest that it may not be possible to find many elements of A satisfying the topological irregularity condition if A can be (effectively) resolved into components with properties very different from those of A itself[12]. We attempt to formalize this idea in the following definition.

5.5.3.9 Definition Let $A \subseteq 2^\omega$ be $\prod_1$. A is called *homogeneous* if there exists a constant c such that for every $\prod_1$ subset B of A: $\forall n\ \forall k \geq n\ \frac{\#B_k}{\#B_n} \leq c \cdot \frac{\#A_k}{\#A_n}$ (where B_n is the set of words of length n admissible for B).

For homogeneous $\prod_1$ sets there is indeed a connection between complexity and topological entropy.

5.5.3.10 Theorem Let $A \subseteq 2^\omega$ be a homogeneous $\prod_1$ set. Then for some x in A: $\exists m\ \forall n\ I(x(n)) > [\log_2 \#A_n] - m$.

Proof Put $C(m,k) := \{w \in A_k \mid \forall n \leq k\ I(w(n)) > [\log_2 \#A_n] - m\}$. By compactness, it suffices to show that there exists m such that for all k: $C(m,k) \neq \emptyset$.

Now $\#C(m,k) \geq \#A_k - \bigcup_{n \leq k} \{w \in 2^k \mid I(w(n)) \leq [\log_2 \#A_n] - m\}$. To calculate $\#\{w \in 2^k \mid I(w(n)) \leq [\log_2 \#A_n] - m\}$, note that $\#\{v \in 2^n \mid I(v) \leq [\log_2 \#A_n] - m\} \leq \#A_n \cdot 2^{-I(n)-m} \cdot d$ by lemma 5.1.2.4. Hence by homogeneity, $\#\{w \in 2^k \mid I(w(n)) \leq [\log_2 \#A_n] - m\} \leq$

$\leq c \cdot \frac{\#A_k}{\#A_n} \cdot \#A_n \cdot 2^{-I(n)-m} \cdot d = \#A_k \cdot 2^{-I(n)-m} \cdot c \cdot d$. Take m so large that $c \cdot d$ is dwarfed. We may then write: $\#C(m,k) \geq \#A_k - \sum_{n \leq k} \#A_k \cdot 2^{-I(n)-m} \geq \#A_k(1 - 2^{-m}) > 0$. Hence there exists m such that for all k, $C(m,k) \neq \emptyset$.

□

This is not quite the optimal result. The topological analogue of theorem 5.4.2.3, $x \in R(\mu)$ if and only if $\exists m\ \forall n\ I(x(n)) > [-\log_2 \mu[x(n)]] - m$,would be: under suitable restrictions on A, for sufficiently large m, $\mathbf{E}(A) = \mathbf{E}\{x \in A \mid \forall n\ I(x(n)) > [\log_2 \#A_n] - m\}$[13]. By putting a condition on A which is an elaboration of the considerations which lead up to the definition of

homogeneity, we can indeed achieve this.

Observe that, if A is homogeneous, for all w, n and k≥n: $\frac{\#(A\cap[w])_k}{\#(A\cap[w])_n} \leq c\cdot\frac{\#A_k}{\#A_n}$.

However, this fact does not exclude the possibility that $\frac{\#(A\cap[w])_k}{\#(A\cap[w])_n}$ is of lower order than $\frac{\#A_k}{\#A_n}$. This happens for instance if $A\cap[w] = \{x\}$, whereas $\#A_n$ is unbounded.

Hence, even if A is homogeneous in the sense of definition 5.5.3.9, it may still be possible to resolve A effectively into components which do not resemble A in the least. We therefore put

5.5.3.11 Definition A is *strongly homogeneous* if A is homogeneous and if for some constant e, for all w such that $A\cap[w] \neq \emptyset$, for all n and k≥n: $\frac{\#A_k}{\#A_n} \leq e\cdot\frac{\#(A\cap[w])_k}{\#(A\cap[w])_n}$.

We then have

5.5.3.12 Corollary Let $A \subseteq 2^\omega$ be a strongly homogeneous $\prod_1$ set. Then for sufficiently large m, $\mathbf{E}(A) = \mathbf{E}\{x\in A \mid \forall n\, I(x(n)) > [\log_2\#A_n] - m\}$.

Proof If A is strongly homogeneous, then for all w such that $A\cap[w] \neq \emptyset$ and for all $\prod_1$ subsets B of $A\cap[w]$:

$$\frac{\#B_k}{\#B_n} \leq \frac{c}{e}\cdot\frac{\#(A\cap[w])_k}{\#(A\cap[w])_n}.$$

For each w such that $A\cap[w] \neq \emptyset$ we may therefore repeat the argument of theorem 5.5.3.10. Since c/e is independent of w, we get m such that for all w such that $A\cap[w] \neq \emptyset$, there is x in $A\cap[w]$ satisfying $\forall n\, I(x(n)) > [\log_2\#A_n] - m$. Hence w is admissible for A iff it is admissible for $\{x\in A \mid \forall n\, I(x(n)) > [\log_2\#A_n] - m\}$, which shows that the topological entropies must be equal. □

5.5.3.13 Remark If A is a strongly homogeneous $\prod_1$ set, and if $\#A_n$ is unbounded, A must be perfect. It follows that $\{x\in A \mid \forall n\, I(x(n)) > [\log_2\#A_n] - m\}$ must have the cardinality of the continuum, e.g. by observing that a non–empty $\prod_1$ set without recursive elements has the cardinality of the continuum (cf. lemma 26 in Jockusch and Soare [38,38]).

5.5.3.14 Corollary Let $A \subseteq 2^\omega$ be a strongly homogeneous $\prod_1$ set with a recursive set of admissible words and such that $\#A_n$ is unbounded.

Then $\mathbf{E}(A) = \mathbf{E}(\{x \in A \mid \lim_{n\to\infty} \frac{I(x(n))}{n} = \mathbf{E}(A)\}$ and $\{x \in A \mid \lim_{n\to\infty} \frac{I(x(n))}{n} = \mathbf{E}(A)\}$ has the cardinality of the continuum.

Digression: oscillations We investigate briefly the oscillations of complexity of sequences x in a $\prod_1$ set A. The material in 5.4.2 leads one to conjecture that there is no x in A which satisfies $\exists m\, \forall n\, I(x(n)) > [\log_2 \#A_n] + I(n) - m$. That this is indeed so, at least for A such that $\#A_n$ does not grow too slowly, is the content of the following theorem. To state the condition of growth in a simple form, we assume that A is shift–invariant.

5.5.3.14 Theorem Let A be a shift–invariant $\prod_1$ subset of 2^ω with a recursive set of admissible words. Suppose there exists a total recursive $f: \omega \to \omega$ with $\lim_{i\to\infty} f(i) = \infty$,

such that for all n and i: $\frac{\#A_n}{\#A_{n-i}} \geq f(i)$. Then no sequence x in A satisfies
$\exists m\, \forall n\, I(x(n)) > [\log_2 \#A_n] + I(n) - m$.

Proof The proof is modelled upon that of theorem 5.4.2.4. It suffices to show that for every Δ_2 definable sequence x in A:
$\lim_{n\to\infty} ([\log_2 \#A_n] - I_0(x(n)|n)) = \infty$.

To this end, we may copy the proof of theorem 5.4.2.2 until we come to the inequality: $I_0(x(n)|n) \leq I(i) + I_0(x_{i+1}.....x_n|n-i) + d$. By shift invariance, $T^i x \in A$, hence (forgetting about the constants) $I_0(x_{i+1}.....x_n|n-i) \leq [\log_2 \#A_n]$. We then have $[\log_2 \#A_n] - I_0(x(n)|n) \geq$

$$\geq [\log_2 \#A_n] - [\log_2 \#A_{n-i}] - I(i) \geq \log_2 \frac{\#A_n}{\#A_{n-i}} - I(i) \geq f(i) - I(i).$$

Since f is total recursive and $\lim_{i\to\infty} f(i) = \infty$, $\forall m\, \exists i\, (f(i) > I(i) + m)$, which proves the theorem. □

Although natural examples from probability theory (such as example 5.5.3.4) satisfy the hypothesis of the theorem, equally natural examples from the logic (such as the set of complete consistent extensions of Peano arithmetic) do not. It is conceivable that in those cases the complexity is considerably higher.

5.5.4 Kamae–entropy This measure of disorder is local, i.e. pertains to individual trajectories and as such can be compared directly to the quantity $\limsup_{n\to\infty} \frac{I(x(n))}{n}$.

5.5.4.1 Definition Given $x \in 2^\omega$, define measures μ_n on 2^ω by: $\mu_n[w] = \frac{1}{n}\sum_{k=1}^{n} 1_{[w]}(T^k x)$.

Let V(x) denote the set of limit points of the μ_n (with respect to the topology of weak convergence). Each limitpoint μ is stationary, so we may associate to each $\mu \in V(x)$ its metric entropy $H(\mu)$. Put $h(x) := \sup\{H(\mu) | \mu \in V(x)\}$. h(x) is called the *Kamae–entropy* of x (Kamae [40]).

5.5.4.2 Example Let μ be a stationary measure and x an *ergodic point* with respect to μ, i.e. for all w, $\mu[w] = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{[w]}(T^k x)$. Then $V(x) = \{\mu\}$ and $h(x) = H(\mu)$.

5.5.4.3 Example (Sturmian trajectories) Let C be the unit circle, parametrized as $C = \{e^{ia} | a \in [0,2\pi)\}$. Let $\alpha \in [0,2\pi)$ be irrational and let S be the transformation $S(e^{ia}) = e^{i(a+\alpha)}$. S represents an irrational rotation of the circle around angle α. Put $C_0 := \{e^{ia} | a \in [0,\pi)\}$, $C_1 := \{e^{ia} | a \in [\pi,2\pi)\}$. C_0 and C_1, together with the excluded points $e^{i\pi} = -1$ and $e^{2\pi i} = 1$ represent a partition (or "measurement") of the "phase space" C. As in 5.5.1, we may define a mapping $\psi: C \to 2^\omega$ by $\psi(\gamma)_k = j$ iff $S^k(\gamma) \in C_j$. Let $A := \psi[C]$, then A is an uncountable closed shift–invariant set. Elements of A are called *Sturmian trajectories*. It can be shown that there exists only one stationary measure μ on A, and that this measure has zero entropy. As a consequence, the Kamae–entropy of all x in A equals zero. Kamae calls sequences x with h(x) = 0, *deterministic*. An examination of the definition of entropy shows that such sequences are in a sense asymptotically predictable. It will be seen in 5.6 that deterministic sequences have some of the properties postulated of admissible place selections.
The relation between Kamae–entropy and I is given by

5.5.4.4 Theorem (Brudno [10, 145]) For all x, $\limsup_{n\to\infty} \frac{I(x(n))}{n} \leq h(x)$.

In this case, use of I does not seem to have technical advantages, so we refer the reader to Brudno's proof (l.c.). Note that the inequality is strict for recursive points which are ergodic for a measure with positive entropy. Examples are recursive Bernoulli sequences; for instance, the sequence constructed by Champernowne: 0100011011000001...

5.6 Admissible place selections In conclusion of this chapter, we come back to one of the issues raised in Chapter 2, namely, the intensional character of admissible place selection. We observed in 2.3.3 that, in general, admissibility is not a property of the graph of a place selection, but, as indicated by the phrase *ohne Benützung der Merkmalunterschiede*, a relation between the process generating the Kollektiv and the process determining the place selection.

In some degenerate cases, namely, when the admissibility of a place selection is assumed for a priori reasons, one may predicate admissibility of a place selection itself. This is so, for instance, if the selection is lawlike. But we noted in 2.5.1 that it is doubtful whether a priori admissibility and lawlikeness really coincide. To substantiate this claim, we present in 5.6.1 a theorem due to Kamae, which states that the *deterministic* sequences introduced in 5.5.4 have many of the virtues of admissible place selections. In 5.6.2 we widen the framework and attempt to capture the intensional aspect of admissible place selection.

5.6.1 Deterministic sequences A deterministic sequence, as introduced in 5.5.4, is one which is asymptotically predictable. A nice way to see this, is to apply Brudno's theorem 5.5.4.4, which implies that if $h(x) = 0$, then $I(x(n))/n$ converges to 0. Using a computation similar to the one given in 5.5.2, we see that the predictability horizon, which is approximately linear in the data for positive entropy, must recede in this case. In this sense, deterministic sequences are generalisations of recursive sequences. (In another sense, they are not: it is easy to show that each Turing degree contains, e.g., a Sturmian trajectory (5.5.4.3).) It stands to reason that two sequences, one of which is asymptotically predictable and the other having a predictability horizon linear in the data, are independent. The following theorem bears this out. Recall that B(p) is the set of Bernoulli sequences with parameter p (definition 2.5.1.3).

5.6.1.1 Theorem (Kamae [40]) Under the hypothesis $\liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} y_k > 0$, the following are equivalent for all $p \in (0,1)$:

(1) $h(y) = 0$

(2) for all $x \in B(p)$: ${}^{x}/_{y} \in B(p)$.

The hypothesis of the theorem is necessary, since given $x \in B(p)$ it is easy to construct a y in which 1 occurs with limiting relative frequency 0, such that ${}^{x}/_{y} \notin LLN(p)$.

It is out of the question to prove Kamae's theorem here. To give the reader nevertheless an inkling of the fundamental idea involved, we have decided to include a quick calculation, which illustrates the direction (2) $\Rightarrow$ (1) of the theorem.

5.6.1.2 Proposition Let $p \in (0,1)$ and let μ be a stationary measure on 2^{ω} such that $\mu\{y \mid \forall x \in LLN(p): {}^{x}/_{y} \in LLN(p)\}=1$. Then for μ–a.a. y: $h(y) = 0$.

Proof By the ergodic decomposition theorem, it suffices to prove the proposition for ergodic μ. By the ergodic theorem (5.5.2.3), μ–a.a. y are ergodic points with respect to μ. Hence (cf. example 5.5.4.2) the conclusion holds if we can show that, under the hypothesis of the theorem, $H(\mu) = 0$. Suppose $H(\mu) > 0$. By a result of Furstenberg (lemma 3.1 in Kamae [40]),

in this case there exists a stationary measure ν on $2^\omega \times 2^\omega$ which has μ and μ_p as marginals, but for which $\nu([0]\times[1]) \neq \mu_p[0]\cdot\mu[1]$. By the ergodic theorem

$$\nu\left\{<x,y>\mid \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{[0]\times[1]}(T^k<x,y>) \neq \mu_p[0]\cdot\mu[1]\right\} > 0.$$

But then, by the properties of /,

$$\nu\left\{<x,y>\mid x \in LLN(p), \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} y_k = \mu[1], {}^{x}/_{y} \notin LLN(p)\right\} > 0.$$

Disintegrating ν, i.e. constructing a family of measures $\{\nu_y\}_{y\in 2^\omega}$ such that for all $E \subseteq 2^\omega\times 2^\omega$,

$$\nu E = \int_{2^\omega} \nu_y E_y d\mu(y),$$

we see that for some $A \subseteq 2^\omega$ with $\mu A > 0$, and all y in A: $\nu_y(LLN(p)\cap(/y)^{-1}LLN(p)^c) > 0$, whence $\mu\{y\mid LLN(p)\cap(/y)^{-1}LLN(p)^c \neq \emptyset\} < 1$, a contradiction. □

The key ingredient of the proofs, both of Kamae's theorem and the above proposition, is provided by Furstenberg's theorem which states, very loosely speaking, that two processes of positive entropy cannot be entirely independent. One may now wonder whether Kamae's theorem has an analogue for random sequences. In particular, do we have, under suitable restrictions on y:

for all computable $p \in (0,1)$, the following are equivalent

(1) $\lim_{n\to\infty} \frac{I(y(n))}{n} = 0$

(2) for all $x \in R(\mu_p)$: ${}^{x}/_{y} \in R(\mu_p)$?

5.6.2 Admissibility and complexity We now turn to the intensional aspect of admissibility. One way to explain admissibility is as follows: we might say that a sequence y is an admissible place selection for a Kollektiv x if y contains no information about x. In other words, y cannot use the *Merkmalunterschiede* of x since it knows too little about x. There are various ways to formalize this idea. One might use conditional complexity I(x(n)|y(m)), or the relative complexity I^y, which was defined in 5.4.3. We choose the latter possibility.

5.6.2.1 Definition Let $p \in (0,1)$ be computable. If $x \in R(\mu_p)$, then y is an *admissible place selection* with respect to x if $\exists m\ \forall n\ I^y(x(n)) > [-\log_2\mu_p[x(n)]] - m$.

5.6.2.2 Remark This definition may seem surprising, in view of the preceding motivation. In fact, a definition of the form: "y is an admissible place selection with respect to x if $\exists m\ \forall n$

$I^y(x(n)) > I(x(n)) - m$" would be rather more elegant. But then it is not clear that there exist *non–recursive* y which are admissible (in this sense) with respect to a non–negligible set of x's. We have already seen (in 5.4.3.1) that if y is a complete Turing degree, i.e. if $Ø' \leq_T y$, then $\lambda\{x| \forall m \exists n \geq m\ (I^y(x(n)) \leq I(x(n)) - m)\} = 1$. On the other hand, with the definition of admissibility we have chosen, it is immediately clear that for all computable μ: $\mu\{x| \exists m \forall n\ I^y(x(n)) > [-\log_2\mu[x(n)]] - m\} = 1$: just relativize theorem 5.4.1.3 to y.

We now put definition 5.6.2.1 to work.

5.6.2.3 Theorem (a) If $x \in R(\mu_p)$ and y is admissible with respect to x, then $^x/_y \in R(\mu_p)$. (b) If $x \in R(\mu_p)$, then the set of y not admissible with respect to x is recursively small (cf. 4.5).

Proof (a) follows by relativizing theorem 5.4.1.3 to y. For (b), we have to show that for any computable measure ν: $\nu\{y| \forall m \exists n\ I^y(x(n)) > [-\log_2\mu_p[x(n)]] - m\} = 0$.
By the Fubini theorem for recursive sequential tests (4.4.4), it suffices to show that $\{<x,y>| \forall m \exists n\ I^y(x(n)) \leq [-\log_2\mu_p[x(n)]] - m\}$ is a recursive sequential test with respect to $\mu_p \times \nu$. Now this set is obviously $\prod_2$; moreover, we have

$$\mu_p \times \nu\{<x,y>| \exists n\ I^y(x(n)) \leq [-\log_2\mu_p[x(n)]] - m\} =$$

$$\int \mu_p\{x| \exists n\ I^y(x(n)) \leq [-\log_2\mu_p[x(n)]] - m\} d\nu(y) \leq \int 2^{-m} d\nu(y) = 2^{-m},$$

the inequality following from the relativized version of theorem 5.4.1.3. □

A trivial combination of the Fubini theorem and theorem 5.4.1.3 thus allows us to capture at least some of the content of the randomness axiom.

Notes to Chapter 5

1. For the subadditive ergodic theorem, see e.g. Y. Katznelson, B. Weiss, A simple proof of some ergodic theorems, *Isr. J. Math* **42** (1982) 291–300.
2. It is a generalisation of the Kraft–inequality from coding theory.
2a. See also Ker-I Ko, On the definition of infinite pseudo–random sequences, *Theor. Comp. Sc.* **48** (1986), 9-34.
3. But with the condition of randomness proposed by Kolmogorov, this verification cannot be effective. A finite sequence w may be called *random* with respect to the distribution $(\frac{1}{2},\frac{1}{2})$ if for some m, $I(w) > |w| - m$. It can be shown that finite random sequences have many of the desired statistical properties, such as (approximate) stability of relative frequency etc.; but, as will be shown in 5.3, there exists no infinite r.e. set of finite random sequences, so that

randomness for finite sequences is in a very strong sense not effectively verifiable. In this respect, Kolmogorov's proposal substitutes one kind of unverifiability for another.

4. The argument used to prove corollary 5.3.1.4 also proves that the graph of the complexity measure I, $\{\langle w,m\rangle \mid I(w) = m\}$ has degree $\emptyset'$.

5. Martin Davis, What is a Computation? in L.A. Steen (ed.), Mathematics Today, Springer Verlag (1978).

6. One might try to define a real–valued measure of the information content of a formal system S along the following lines. Let A(S) be the set of complete consistent extensions of S, then A(S) may be identified with a $\prod_1$ subset of 2^ω. If S_1 is stronger than S_2, then $A(S_1)$ is contained in $A(S_2)$. One may now define the information content of S as the inverse of the *topological entropy* (see section 5.5.3) of A(S). Of course, this measure is interesting only if it can be shown that it is independent of the Gödelnumbering adopted.

7. Although perhaps the usual proofs of van der Waerden's theorem are too ineffective to bring about a decrease in complexity.

8. It is not clear to whom to attribute this result. Chaitin credits Schnorr in [12] and Solovay in [13]. The first published proof appears to be Dies [19].

9. This should be understood (and is proved) in the same way as the corresponding result for prefix algorithms.

10. The proof of the Shannon–McMillan–Breiman theorem does not yield: $x \in R(\mu)$ implies $\lim_{n\to\infty} \frac{I(x(n))}{n} = H(\mu)$. For certain special μ, e.g. those of the form μ_p, this can be proved.

11. Brudno [10,132] proves: if μ is an ergodic measure, then for μ–a.a. x:

$$\limsup_{n\to\infty} \frac{K(x(n))}{n} = H(\mu).$$

12. A simple example of a $\prod_1$ set which can be so resolved is the set A consisting of sequences of the form 1^n0^ω for $n \geq 0$. Any element of A is determined by finitely many bits. Having specified these bits, there is no more need to choose in A_n.

13. We cannot define topological entropy for the set $\{x \in A \mid \exists m\, \forall n\, I(x(n)) > [\log_2 \#A_n] - m\}$, since this set need not be compact. We therefore choose the formulation "for m sufficiently large....".

6 Appendix:
Notation and definitions

6.1 Notations for sequences. 2^ω is the set of infinite binary sequences. If $x \in 2^\omega$, then x(n) is the intial segment of x of length n, and x_n is the n^{th} term (also called coordinate) of x. The mapping T: $2^\omega \to 2^\omega$ (called the left shift) is defined by $(Tx)_n = x_{n+1}$. x is used consistently as a variable over 2^ω; ξ always denotes a variable over $(2^\omega)^\omega$.
$2^{<\omega}$ is the set of all finite binary sequences. An finite binary sequence is alternatively called a *word* or a *string*. The length of a word w is denoted |w|. 2^n is the set of all strings w such that |w| = n. If $m \le |w|$, then w(m) is the initial segment of w of length m, and w_m is the m^{th} term of w. If v is an initial segment of w, we write $v \subseteq w$; if $v \subseteq w$ and $v \neq w$, we write $v \subset w$. The empry string is denoted <>.

6.2 Topology on 2^ω. If B is a set, 1_B denotes the characteristic function of B. Let 2 = {0,1} have the discrete topology and form the product topology on 2^ω. The open sets in this topology are then unions of *cylinders* [w] defined by $[w] := \{x \in 2^\omega | \; x(|w|) = w\}$. If $S \subseteq 2^{<\omega}$, then the open set generated by S, namely $\{x \in 2^\omega | \; \exists w \in S \; (x(|w|) = w)\}$, is denoted [S]. The topology on spaces of the form $(2^\omega)^m$ is constructed analogously.
For any subset A contained in 2^ω, Cl(A) denotes the closure of A, and Int(A) the interior of A. The *boundary* of A, denoted ∂A, is defined to be $\partial A := Cl(A) - Int(A)$.
The Borel σ-algebra on 2^ω is the smallest σ-algebra containing the open sets in 2^ω. Elements of this algebra are called *Borel sets.*

6.3 Measures on 2^ω. A measure on the Borel σ-algebra is completely determined by its values on the cylinders. We shall consider *probability measures* only, i.e. measures μ for which $\mu(2^\omega) = 1$. Now let $(p_n)_n$, where $p_n \in [0.1]$, be a sequence of reals. This sequence determines a product measure on 2^ω, denoted $\prod_n (1 - p_n, p_n)$ and defined as

$$\prod_n (1-p_n, p_n)\,[w] = \prod_{k=1}^{|w|} \bar{p}_k, \text{ where } \bar{p}_k := p_k \text{ if } w_k = 1 \text{ and } \bar{p}_k := 1-p_k \text{ otherwise.}$$

One product measure on 2^ω occurs so often that it isgiven a special name: $\lambda = (\frac{1}{2}, \frac{1}{2})^\omega$.

λ is the image of the Lebesgue measure on the unit interval under the natural map and will also be called Lebesgue measure.

The following relationships among probability measures μ and ν are of special importance.
- μ is *singular* with respect to ν (denoted: $\mu \perp \nu$) if there exists a Borel set A such that $\mu A = 1$ and $\nu A = 0$.
- μ is *absolutely continuous* with respect to ν (denoted: $\mu << \nu$) if for all Borel sets A such that $\nu A = 0$, also $\mu A = 0$.
- μ and ν are *equivalent* (denoted: $\mu \approx \nu$) if $\mu << \nu$ and $\nu << \mu$.

Let $(\mu_n)_n$ be a sequence of measures. We say that μ_n *converges weakly* to ν if for all Borel sets A such that $\nu \partial A = 0$, $\mu_n A$ converges to νA. The Portmanteau theorem [4] states (among else) that weak convergence is equivalent to convergence on the cylinders. We say that μ_n *converges strongly* to ν if for all Borel sets A, $\mu_n A$ converges to νA.

6.4 Computability We shall take as primitive the notion of an algorithm operating on natural numbers, which yields as output natural numbers. It is understood that an algorithm need not terminate on every input. A *partial recursive function* $f: \omega \to \omega$ is a function which can be computed by an algorithm. With this intuitive description it is more or less clear that there exists an effective procedure which associates to each partial recursive function a natural number, its *Gödelnumber*. A *recursive function* is a partial recursive function which is in fact total. More formal definitions of (partial) recursive function and Gödelnumber are possible; see Rogers [86] and Soare [92]. The connection between the informal concept of an algorithm and the formal definition of a partial recursive function is provided by *Church's Thesis*, which states that every algorithm computes a partial recursive function.
Usually one does not formally verify that an apparently recursive function is indeed recursive; one exhibits an algorithm which computes the function and Church's Thesis is invoked to guarantee that the function is in fact recursive. We shall do likewise. We must, however, warn the reader that in constructing algorithms we freely use classical logic; as a consequence, proving the existence of a recursive function need not mean that we can lay our hands on it.
Although we defined partial recursive functions to have the natural numbers as domain and range, this restriction is not as severe as may seem, since many objects can be coded into the natural numbers. In particular, this is true for $\mathbb{Q}$ and $2^{<\omega}$. The following concepts thus make sense. A function $f: \omega \to \mathbb{R}$ is called *computable* if there exists a recursive function $g: \omega \times \omega \to \mathbb{Q}$ such that for all n,k: $|f(n) - g(n,k)| < 2^{-k}$. A measure μ on is computable if there exists a recursive function $g: 2^{<\omega} \times \omega \to \mathbb{Q}$ such that for all w, n: $|\mu[w] - g(w,n)| < 2^{-k}$
.
We shall often use the *arithmetical hierarchy* for subsets of ω and of 2^{ω}. We say that $A \subseteq \omega^k$ is *recursive* if its characteristic function is a recursive function. Starting from the recursive sets, we can define increasingly complex subsets of ω^k using quantification over ω. A is *recursively enumerable or* $\sum_1$ if there exists a recursive $B \subseteq \omega^{k+1}$ such that

$$A = \{u \in \omega^k \mid \exists n\ (<n,u> \in B)\}.$$

A is $\prod_1$ if A^c is $\sum_1$. In general, A is $\sum_n$ if there exists a $B \subseteq \omega^{k+1}$ such that B is $\prod_{n-1}$ and

$$A = \{u \in \omega^k \mid \exists n\ (<n,u> \in B)\};$$

A is $\prod_n$ if A^c is $\sum_n$. Note that $\prod_n$ sets A can be written as

$$A = \{u \in \omega^k \mid \forall n\ (<n,u> \in B)\},$$

for some $\sum_{n-1}$ set B. A is called Δ_n if it is both $\sum_n$ and $\prod_n$. This is the arithmetical hierarchy for subsets of ω^k. (In the textbooks the $\sum$, $\prod$ and Δ usually have superscripts "0", to indicate quantification over natural numbers. Since we shall never quantify over sequences, we have dropped the superscripts.)

We now generalize the concept of recursiveness to spaces of the form $\omega^k \times (2^\omega)^m$. Roughly, a relation $R \subseteq \omega \times 2^\omega$ is *recursive* if for each natural number n and each x, the truth value of R(n,x) can be computed using only a finite piece of x; similarly for relations in $\omega^k \times (2^\omega)^m$.

A subset A of $\omega^k \times (2^\omega)^m$ is $\sum_1$ if there exists a recursive relation B in $\omega^{k+1} \times (2^\omega)^m$ such that

$$A = \{<\bar{n},\bar{x}> \in \omega^k \times (2^\omega)^m \mid \exists j\ B(j,\bar{n},\bar{x})\}.$$

A $\prod_1$ set is the complement of a $\sum_1$ set. The reader can now copy the defintions of $\sum_n$, $\prod_n$ and Δ_n from the corresponding definitions for subsets of ω^k.

We now specialize the preceding definition to the case that subsets of $(2^\omega)^m$ are defined using recursive relations and quantification over natural numbers. Let A be of the form

$$A = \{\bar{x} \in (2^\omega)^m \mid R(\bar{n},\bar{x})\},$$

for some recursive relation R. It follows from the intuitive explanation of recursiveness and the compactness of $(2^\omega)^m$ that A is of this form is A is clopen. The clopen sets will also be called $\sum_0$ sets. It is easily verified that $\sum_1$ sets are open and that $\prod_1$ sets are closed. The converse is of course false, as a cardinality argument shows.

6.5 Ergodic Theory A measure μ on 2^ω is called *stationary* if for all Borel sets A, $\mu T^{-1}A = \mu A$, where T is the left shift defined in 7.1. A measure μ is *ergodic* if for all Borel sets A: $T^{-1}A = A$ implies that μA is either 0 or 1. The single most important fact about stationary measures is the

Ergodic theorem (see [82]) Let μ be a stationary measure on 2^ω, $f: 2^\omega \to \mathbb{R}$ integrable. Then

$$f^*(x) = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(T^k x)$$

exists μ-a.e., f^* is T-invariant and $\int f d\mu = \int f^* d\mu$. In addition, if μ is ergodic then f^* is constant μ-a.e.

We say that a measure μ on 2^{ω} is *strongly mixing* if for all Borel sets A, B: $\mu(T^{-n}A \cap B)$ converges to $\mu A \cdot \mu B$.

References

Asterisks indicate items which contain extensive bibliographies on random sequences.

1. L. Accardi, The probabilistic roots of the quantum mechanical paradoxes, in: S. Diner et al. (eds), The Wave-Particle Dualism, Reidel (1984), 297-330.
2. V.M. Alekseev, M.V. Yakobson, Symbolic dynamics and hyperbolic dynamical systems, *Physics Reports*, **75** (1981), 287-325.
3. V. Barnett, Comparative statistical inference, 2nd ed., Wiley (1982).
4. P. Billingsley, Convergence of probability measures, Wiley (1983).
5. E. Bishop, Foundations of constructive analysis, McGraw-Hill (1967).
6. E. Bishop, D.S. Bridges, Constructive analysis, Springer-Verlag (1985).
7. E. Bishop, H. Cheng, Constructive measure theory, *Mem. AMS* **116**.
8. E. Borel, Les probabilités dénombrables et leurs applications arithmétiques, Note V in: E. Borel, Leçons sur la théorie des fonctions, Gauthiers-Villars (1914), 182-216.
9. D.S. Bridges, Constructive functional analysis, *Research notes in mathematics* **28**, Pitman (1979).
10. A.A. Brudno, Entropy and the complexity of the trajectories of a dynamical system, *Trans. Mosc. Math. Soc.* **44** (1983), 127-151.
11.* G.J. Chaitin, Information theoretic limitations on formal systems, *J. Ass. Comp. Mach.* **21** (1974), 403-424.
12.* G.J. Chaitin, A theory of program size formally identical to information theory, *J. Ass. Comp. Mach.* **22** (1975), 329-340.
13.* G.J. Chaitin, Algorithmic information theory, *IBM J. Res. Dev.* **21** (1977), 350-359;496.
14.* G.J. Chaitin, Gödel's theorem and information, *Int. J. Theor. Phys.* **21** (1982), 941-954.
15. D.G. Champernowne, The construction of a decimal normal in the scale of ten, *J. Lond. Math. Soc.* **8** (1932),254-260.
16. A. Church, On the concept of a random sequence, *Bull. AMS* **46** (1940), 130-135.
17. A.H. Copeland, The theory of probability from the point of view of admissible numbers, *Ann. Math. Stat.* **3** (1932), 143-156.
18. D. Cox, The role of significance tests, *Scand. J. Stat.* **4** (1977), 49-70.
19. J.-E. Dies, Information et complexité, *Ann. Inst. Henri Poincaré B* **12** (1976), 365-390; ibidem **14** (1978), 113-118.
20. J.L. Doob, A note on probability, *Ann. Math.* **37** (1936), 363-367.
21. J.L. Doob, Probability as measure, *Ann. Math. Stat.* **12** (1941), 206-214; 216-217.
22. K. Dörge, Eine Axiomatisierung der v. Misessche Wahrscheinlichkeitsrechnung,

Jahresber. DMV **43** (1934), 39-47.

23. W. Feller, Sur les axiomatiques du calcul des probabilités et leurs rélations avec les expériences, in [35], 7-21.
24. W. Feller, Über die Existenz sogenannter Kollektive, *Fund. Math.* **32** (1939), 87-96. See also [25,203].
25. W. Feller, Introduction to probability theory and its applications, Volume 1, 3rd ed. Wiley (1968).
26. W. Feller, Introduction to probability theory and its applications, Volume 2, Wiley (1971).
27. J. Ford, How random is a coin toss?, *Physics Today*, April 1983.
28. M. Fréchet, Exposé et discussion de quelques recherches récentes sur les fondéments du calcul des probabilités, in [35], 22-55.
29. H. Freudenthal, Realistic models in probability, in I. Lakatos (ed.), Problems in inductive logic, North-Holland (1969).
30. H. Furstenberg, Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation, *Math. Syst. Theory* **1** (1967), 1-49.
31. P. Gacs, Exact expressions for some randomness tests, *Z. Math. Logik. Grundl. Math.* **26** (1980).
32.* P. Gacs, On the relation between descriptional complexity and algorithmic probability, *Theor. Comp. Sc.* **22** (1983), 71-93.
33. P. Gacs, Every sequence is reducible to a random one, *Inf. and Control* **70** (1986), 186-192.
34. H. Gaifman, M. Snir, Probabilities over rich languages, randomness and testing, *J. Symb. Log.* **47** (1982), 495-548.
35. Colloque consacré au calcul des probabilités, Proceedings of a conference held at the Université de Genève in 1937. The papers concerning the foundations of probability were published in the series *Actualités Scientifiques et Industrielles*, **735**, Hermann (1938).
36.* J. Hartmanis, Generalized Kolmogorov complexity and the structure of feasible computations, Proc. 24th FOCS (1983), 439-445.
37. F. Hausdorff, Grundzüge der Mengenlehre, von Veit (1914).
38. C.G. Jockusch, R.I. Soare, $\prod_1$ classes and degrees of theories, *Trans. AMS* **173** (1972), 33-56.
39. S. Kakutani, On the equivalence of infinite product measures, *Ann. Math.* **49** (1948), 214-224.
40. T. Kamae, Subsequences of normal sequences, *Isr. J. Math.* **16** (1973), 121-149. See also: T. Kamae, B. Weiss, Normal numbers and selection rules, *Isr. J. Math.* **21** (1975), 101-110.

41. E. Kamke, Über neuere Begründungen der Wahrscheinlichkeitsrechnung, *Jahresber. DMV* **42** (1932), 14-27.
42. A.S. Kechris, Measure and category in effective descriptive set theory, *Ann. Math. Logic* **5** (1973), 337-384.
43. A. Khintchin, Review of J. Blume, Zur axiomatische Grundlegung der Wahrscheinlichkeitsrechnung, *Zentralblatt* **10** (1935), 69.
44. A.N. Kolmogorov, Grundbegriffe der Wahrscheinlichkeitsrechnung, *Ergebnisse der Mathematik und ihrer Grenzgebiete*, J. Springer (1933).
45. A.N. Kolmogorov, Das Gesetz der geiterierten Logarithmus, *Math. Ann.* **101** (1929), 126-135.
46. A.N. Kolmogorov, The theory of probability, Chapter IX of A.D. Aleksandrov, A.N. Kolmogorov, M.A. Lavrent'ev (eds.): Mathematics, its contents, methods and meaning, *Translations of mathematical monographs*, Volume 1, AMS, Providence, R.I. (1963).
47. A.N. Kolmogorov, On tables of random numbers, *Sankhya* **25** (1963), 369-376.
48. A.N. Kolmogorov, Three approaches to the definition of the concept of "amount of information", *Sel. Transl. Math. Stat. and Prob.* **7**, AMS, Providence, R.I. (1968).
49. A.N. Kolmogorov, The logical basis for information theory and probability theory, *IEEE Trans. Inf. Theory* **IT-14** (1968), 662-664.
50. A.N. Kolmogorov, Combinatorial basis of information theory and probability theory, *Russ. Math. Surveys* **38** (1983), 29-40.
51. A.N. Kolmogorov, On logical foundations of probability theory, in: K. Itô, J.V. Prokhorov (eds), Probability theory and mathematical statistics, *Lecture Notes in Mathematics*, **1021**, Springer-Verlag (1984).
52. G. Kreisel, Informal rigour and completeness proofs, in: I.Lakatos (ed.), Problems in the philosophy of mathematics, North-Holland (1969), 138-157.
53. G. Kreisel, A. Levy, Reflection principles and their use for establishing the complexity of axiomatic systems, *Z. Math. Logik Grundl. Math.* **14** (1968), 97-142.
53. M. van Lambalgen, Von Mises' definition of random sequences reconsidered, to appear, *J. Symb. Logic* **57** (1987).
54.* L.A. Levin, A.K. Zvonkin, The complexity of finite objects and the algorithmic foundations of the notions of information and randomness, *Russ. Math. Surveys* **25** (1970).
55. L.A. Levin, Measures of complexity for finite objects (axiomatic description), *Sov. Math. Dokl.* **17** (1976), 552-526.
56. L.A. Levin, Randomness conservation inequalities: information and independence in mathematical theories, *Inf. and Control* **61** (1984), 15-37.
57. P. Lévy, Calcul des probabilités, Gauthiers-Villars (1925).

58. A.J. Lichtenberg, M.A. Lieberman, Regular and stochastic motion, Springer-Verlag (1983).
59. W. Maass, Are recursion theoretic arguments useful in complexity theory?, in: Barcan Marcus et al. (eds.), Logic, methodology and philosophy of science VII, North-Holland (1986), 141-158.
60. D. Malament, S. Zabell, Why Gibbs averages work – the role of ergodic theory, *Phil. of Science* **47** (1980), 339-349.
61. P. Martin-Löf, Algorithmen und zufällige Folgen, lecture notes, Universität Erlangen (1966). See also: Complexity oscillations in infinite binary sequences, *Z.Wahrsch. verw. Geb.* **19** (1971), 225-230.
62. P. Martin-Löf, The definition of random sequences, *Inf. and Control* **9** (1966), 602-619.
63.* P. Martin-Löf, The literature on von Mises' Kollektivs revisited, *Theoria* (1969).
64. P. Martin-Löf, On the notion of randomness, in: A. Kino, J. Myhill, R.E. Vesley (eds.), Intuitionism and proof theory, North-Holland (1970), 73-78.
65. P. Martin-Löf, The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data, *Scand. J. Stat* **1** (1974), 3-18.
66. P. Martin-Löf, Reply to Sverdrup's polemical article "Tests without power", *Scand. J. Stat.* **2** (1975), 161-165.
67. R. von Mises, Grundlagen der Wahrscheinlichkeitsrechnung, *Math. Z.* **5** (1919), 52-99.
68. R. von Mises, Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und theoretische Physik, Deuticke (1931).
69. R. von Mises, Über Zahlenfolgen die ein Kollektiv-ähnliches Verhalten zeigen, *Math. Ann.* **108** (1933), 757-772.
70.* R. von Mises, Wahrscheinlichkeit, Statistik und Wahrheit, 2nd ed., J. Springer (1936).
71. R. von Mises, Kleines Lehrbuch des Positivismus, Library of Unified Science Book Series, Volume I, W.P. van Stockum en Zoon (1939). English translation: Postivism, a study in human understanding, Harvard University Press (1951).
72. R. von Mises, Quelques rémarques sur les fondements du calcul des probabilités, in [35], 57-66.
73. R. von Mises, On the foundations of probability and statistics, *Ann. Math. Stat.* **12** (1941), 191-205; 215-216.
74. R. von Mises, H. Geiringer, Mathematical theory of probability and statistics, Academic Press (1964).
75*. R. von Mises, Probability, Statistics and Truth (English translation of 3rd ed. of Wahrscheinlichkeit, Statistik und Wahrheit; cf. [70]), Dover (1981).

76. D.W. Müller, Randomness and extrapolation, in J. Neyman, L.M. Lecam, E.L. Scott (eds.), Sixth Berkeley Symposium on Probability and Statistics (1970), 1-31.
77. J. Neveu, Discrete parameter Martingales, North-Holland (1975).
78. J. Neyman, E.S. Pearson, Joint statistical papers, Cambridge University Press (1967)
79. A. Novikoff, J. Barone, History of axiomatic probability theory from Borel to Kolmogorov Part I, *Arch. Hist. Exaxt Sciences* **18** (1978), 123-190.
80. J.C. Oxtoby, Measure and category, 2nd ed., Springer-Verlag (1980).
81. O. Penrose, J. Lebowitz, Modern ergodic theory, *Physics Today*, February 1973.
82. K. Petersen, Ergodic theory, Cambridge University Press (1983).
83. K.R. Popper, Logik der Forschung, J. Springer (1935). English translation: Logic of scientific discovery, Hutchinson & Co. (1975).
84. K.R. Popper, Realism and the aim of science, Rowman and Littlefield (1983).
85. H. Reichenbach, Axiomatik der Wahrscheinlichkeitsrechnung, *Math. Z.* **34** (1932), 568-619.
86. H. Rogers, Theory of recursive functions and effective computability, McGraw Hill (1967).
87. G. Sacks, Measure-theoretic uniformity in recursion theory and set theory, *Trans. AMS* **142** (1969), 381-424.
88.* C.P. Schnorr, Zufälligkeit und Wahrscheinlichkeit, *Lecture Notes in Mathematics* **218**, Springer-Verlag (1971).
89. C.P. Schnorr, Process complexity and effective random tests, *J. Comp. Syst. Sc.* **7** (1973), 376-388.
90.* M. Sipser, A complexity theoretic approach to randomness, Proc. 15th ACM Symp. Th. Comp. (1983), 330-335.
91. C. Smorynski, The incompleteness theorems, in: J. Barwise (ed.), Handbook of mathematical logic, North-Holland (1977), 821-865.
92. R.I. Soare, Recursively enumerable sets and their degrees, to appear in the series *Perspectives in Mathematical Logic*, Springer-Verlag.
93. R.M. Solovay, On random r.e. sets, in: A. Arruda, N. da Costa, R. Chuaqui (eds.), Non-classical logics, model theory and computability, North-Holland (1977).
94. H. Steinhaus, Les probabilités dénombrables et leur rapport à la théorie de la mésure, *Fund. Math.* **4** (1922), 286-310.
95. E. Sverdrup, Tests without power, *Scand. J. Stat.* **2** (1975), 158-160.
96. E. Tornier, Comment, *Math. Ann.* **108** (1933), 320.
97. A.S. Troelstra, On the origin and development of Brouwer's concept of a choice sequence, Report 82-07, Dept. of Mathematics, University of Amsterdam.
98. A.S. Troelstra, Analysing choice sequences, *J. Phil. Logic* **12** (1983), 197-260.
99. J. Ville, Étude critique de la notion de collectif, Gauthiers-Villars (1939).

100. A. Wald, Die Widerspruchsfreiheit des Kollektivbegriffes der Wahrscheinlichkeitsrechnung, *Ergebnisse eines math. Koll.* **8** (1936), 38-72.
101. A. Wald, Die Widerspruchsfreiheit des Kollektivbegriffes, in [35], 79-99.
102. A. Wald, Sequential analysis, Wiley (1947). Reprint Dover (1973).
103. M. Fréchet, The diverse definitions of probability, lecture at the fourth International Congress for the Unity of Science,*Erkenntnis* (1938).
104. A.R. Bernstein, F. Wattenberg, Non-standard measure theory, in: W.A.J. Luxemburg (ed.), Applications of model theory to algebra, analysis and probability, Holt, Rhinehart and Winston (1969), 171-185.