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....".