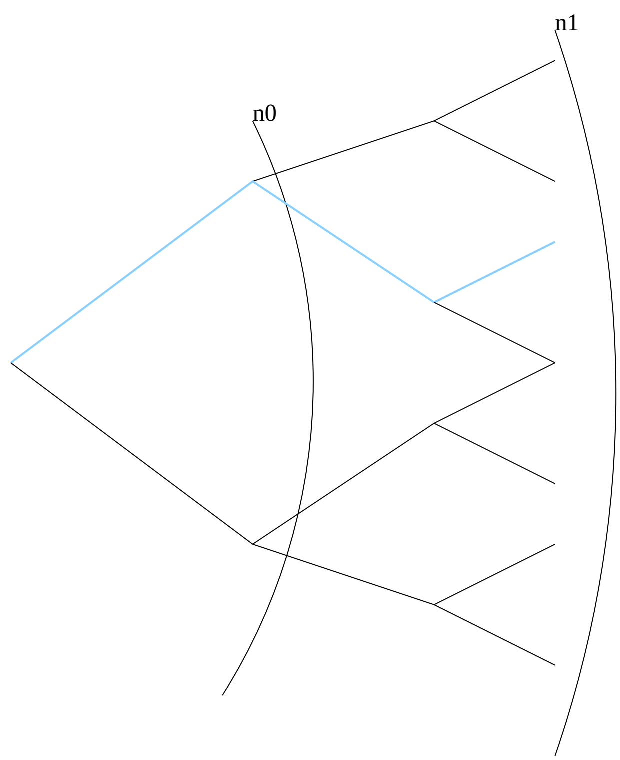
4 Randomness

Until now, we have theorems which estimate, for any given n, the number of strings of length n which are incompressible. They do not reveal how the incompressiblility of a given string evolves along its prefixes.

For example, consider two numbers n_0 and n_1 with n_1 being larger of the two. The estimate in Theorem 3 of section 3 gives us that most strings at lengths n_0 and n_1 are incompressible. But it does not mean that the n_1 long extension of an n_0 length incompressible string is incompressible. In the illustration (Figure 1), we cannot say that the blue path is incompressible at length n_0 as well as n_1.

In this section, we work towards results which say that almost all infinite binary sequences are incompressible at all large enough lengths. That is, in the infinite binary tree representing the sequences, almost all paths are incompressible at *all* large enough lengths. This would, of course imply Theorem 4 of section 3 as a corollary. Thus we are in search of a (much) stronger result.

Figure 1: Incompressibility at two different lengths.



4.1 Uniform Distribution

Most of this material is adapted from [1]. Even though our aim in this section is to characterize random infinite sequences, we will begin by studying finite strings. Since it is unnatural to fix a value k, and call an n length string random if the first k bits of the string are the same, but not if only $k-1$ bits are, we will introduce a notion of *deficiency of randomness* which will quantify the amount of non-randomness in the string.

Let us consider positive real-valued functions d defined on strings. Since we expect most strings to be nearly incompressible, one of the conditions for a deficiency function we impose is that at any length, the probability of the set of strings with deficiency at least k, falls off as 2^{-k}. That is, for strings of length n,

$$\sum_{\substack{x \in \Sigma^n \\ d(x) > k}} Pr[x] \leq \frac{1}{2^k}.$$

Here the probability we consider is the uniform probability. Hence the same condition may be rephrased as saying that there are at most 2^{n-k} strings of length n with deficiency at least k.

Also, we impose a computability restriction on a deficiency function. We do not require that d be computable: we only require that a Turing machine should be able to say yes and halt if a given number M is less than the deficiency of a string. We do not require that the machine say no and halt if M is at least as big as the deficiency of the string.

Definition 4.1.1. A real-valued function on strings f is said to be *lower semicomputable* if there is a total computable function

$$\hat{f} : \Sigma^* \times \mathbb{N} \to \mathbb{Q}$$

such that for any string x,

- For all numbers n, $\hat{f}(x, n) < \hat{f}(x, n+1) < f(x)$.
- $\lim_{n \to \infty} \hat{f}(x, n) = f(x)$.

A real-valued function g is said to be *upper semicomputable* if $-g$ is lower semicomputable.

Thus, a lower semicomputable function f has an approximation $\hat{f}$ which approaches f strictly from below. This makes it easy to construct another Turing machine which accepts a number M if $M < f(x)$ - it merely examines the values of $\hat{f}(x, n)$ for $n \in \mathbb{N}$. If ever $\hat{f}(x, n)$ exceeds M, then it is possible to conclude that $f(x) > M$, and the machine accepts. If $f(x) > M$, then eventually $\hat{f}(x, n)$ does exceed M since the limiting value of $\hat{f}(x, n)$ as n approaches ∞, is $f(x)$.

Such a deficiency function is called a Martin-Löf test.

Definition 4.1.2. A lower semicomputable function $d : \Sigma^* \to [0, \infty]$ is called a *Martin-Löf Test* if for any length n,

$$\sum_{\substack{x \in \Sigma^n \\ d(x) > k}} Pr[x] \leq \frac{1}{2^k}.$$

A Martin-Löf test d_0 is said to be *universal* if it additively dominates every other Martin-Löf test. That is, if d is a Martin-Löf test, then there is a non-negative constant c such that for all strings x, $d(x) < d_0(x) + c$. Thus, d_0 is nearly as good at computing the nonrandomness in a given string as any other specific d might be. The surprising fact is that there is a universal Martin-Löf test.

Theorem 4.1.3. *The function* $d_0 : \Sigma^* \to [0, \infty)$ *defined below is a universal Martin-Löf test.*

$$d_0(x) = |x| - K(x \mid |x|).$$

We will need several auxiliary results before we can prove this. We now work up to the proof, via the following results.

Lemma 4.1.4. *For any strings x and y and large enough i,*

$$K\,(\,x \mid y, i - K(x \mid y, i)\,) \leq K(x \mid y, i) + O(1).$$

Proof. Consider a shortest prefix program p which takes input y and i and outputs x. Equivalently, $U(\,\langle p, y, i\rangle\,) = p(\langle y, i\rangle) = x$. Then this program's length is $K(x \mid y, i)$.

Let A be a machine which on inputs y, n and p, outputs

$$A(\,\langle p, y, n\rangle\,) = U(\,\langle p, y, n + |p|\rangle\,).$$

That is, A computes $|p|$, and then simulates U with inputs p, y and $n + |p|$. Then $A(\,\langle p, y, i - |p|\rangle\,) = U(\,\langle p, y, i\rangle\,) = x$.

Thus by definition of conditional complexity,

$$K_A(x \mid y, i - |p|) \leq K(x \mid y, i) + O(1),$$

from which it follows that

$$K(x \mid y, i - K(x|y, i)) \leq K(x \mid y, i) + O(1).$$

□

Using this lemma, we prove a general characterization of upper bounds on Kolmogorov complexity.

Theorem 4.1.5. *Let $F : \Sigma^* \times \Sigma^* \to \mathbb{R}$ be an upper semicomputable function. Then, $K(x|y) < F(x, y) + O(1)$ if and only if for all y and m,*

$$\log |\{x \mid F(x, y) < m\}| < m + O(1), \tag{1}$$

Proof. We prove that if the inequality (1) is satisfied, then $K(x \mid y) < F(x, y) + O(1)$, and omit the proof of the converse.

Assume that F is a function as given. Then consider the set E defined as the set of triples (x, y, m) where $F(x, y) > m$. Since F is upper semicomputable, it follows that E is Turing acceptable [1]. By the property of F, if we fix y and m, there are at most $k \times 2^m$ many triples of the form (x, y, m). Thus, $K(x \mid y, m) < m + O(1)$.

By the above lemma, we can conclude that $K(x \mid y) = K(x \mid y, m - m) \leq K(\,x \mid y, m - K(x \mid y, m)\,) + O(1) < m + O(1)$. □

With this, we can now prove that the Martin-Löf test defined in Theorem 4.1.3 is a universal test.

Proof of Theorem 4.1.3 We first show that d_0 is a Martin-Löf test. d_0 is a lower semicomputable function since length function over strings is computable and $K(x \mid |x|)$ is an upper semicomputable function. Also, we know that the number of strings for which $K(x \mid |x|)$ is at most $|x| - k$ is at most $2^{|x|-k}$. Thus d_0 is a Martin-Löf test.

[1] see the discussion in the paragraph following the definition of semicomputability

To show that it is universal, let d be an arbitrary Martin-Löf test. Then, consider the function $F(x,y)$ defined as

$$F(x,y) = \begin{cases} |x| - d(x) & \text{if } y = |x| \\ \infty & \text{otherwise.} \end{cases}$$

F is upper semicomputable. Also, $F(x,|x|) < k$ only if $|x| - d(x) < k$. There are at most 2^k strings with this property. Thus, F satisfies the conditions of (1) and we have

$$K(x \mid |x|) < F(x,|x|) + O(1) = |x| - d(x) + O(1).$$

Substituting for the value of $d_0(x)$, we get $d(x) < d_0(x) + O(1)$, which proves the claim. □

The above theorem is remarkable because d_0 is a Martin-Löf-test which is defined solely on the basis of the incompressibility of x. An arbitrary test d may measure how deficient x is from some desirable statistical property of a string drawn uniformly at random. The above theorem says that as long as d is lower semicomputable and has the exponential decrease in probability with increasing deficiency, the d_0 deficiency of x will be around the same as the d deficiency. Stated in another way, all incompressible strings will have very low d deficiency. This helps us to relate incompressibility to various statistical properties of the strings.

We will now prove that all Kolmogorov incompressible strings of length n will have roughly $n/2$ zeroes and $n/2$ ones.

Example 4.1.6. (Weak Law of Large Numbers) We will now show that Kolmogorov incompressible strings of length n obey the weak law of large numbers.

For any n-long binary string x and bit b, let $N(b \mid x)$ be the number of occurences of that bit in x. Let p_x denote the ratio

$$\frac{N(1 \mid x)}{|x|}.$$

We let the probability of a bit being 1 as p_x. For any string y, we define

$$P_x(y) = p_x^{N(1|y)} \times (1 - p_x)^{N(0|y)}.$$

This is a positive quantity, less than 1. [2]

We will consider the deficiency function

$$d(x) = \log P_x(x) + n - \log(n+1).$$

We will prove that this is a Martin-Löf test. In fact, it is more convenient to prove that d has a stronger property than the exponential fall-off in probability described above. We show that

$$\sum P(x) 2^{d(x)} \le 1.$$

If we show this, then we can conclude that

$$\sum_{\substack{x \in \Sigma^n \\ d(x) > k}} P(x) \le 2^{-k}$$

[2] However, it may not be a probability over all strings of length n, since the sum of $P_x(\cdot)$ over all such strings may exceed 1.

by Markov inequality. (HW 8 (a)).

We have

$$\begin{aligned}\sum P(x)2^{d(x)} &= \sum 2^{-n}2^n P_x(x)\frac{1}{n+1} \\ &= \frac{1}{n+1}\sum P_x(x)\end{aligned}$$

This sum runs over all strings of length n. Since P_x need not be a probability distribution over strings of length n, we cannot bound the sum from above by 1. We can sum over these strings in a different way - we will group strings together if they have equal number of 1s and 0s and then sum over all such groups. Thus, the sum above is

$$\begin{aligned}\frac{1}{n+1}\sum P_x(x) &= \frac{1}{n+1}\sum_{k=0}^{n} p_x^k(1-p_x)^{n-k} \\ &\leq \frac{1}{n+1}\sum_{k=0}^{n} 1 \\ &= \frac{1}{n+1}n+1,\end{aligned}$$

which is what we needed to show. It is also clear that d is computable, hence it is a Martin-Löf test.

It follows that d_0 additively dominates d. Thus all Kolmogorov incompressible strings will have low value of d.

Now, why does d measure the closeness of the ratio of 1s in x to 1/2?

We will rewrite d as

$$d(x) = n(1-h(p_x)) - \log(n+1),$$

where $h(p)$ is the two-dimensional entropy function $-p\log p - (1-p)\log(1-p)$.

The entropy function $h(p)$ is maximal at $p = 1/2$, when $h(p) = 1$. (see Figure 2) So test d tells us that the probability of strings x with $p_x < p < 1/2$ for some constant p is at most

$$(n+1)2^{-n(1-h(p))}.$$

We know that $(1-h(p)) < c(p-\frac{1}{2})^2$ for some positive constant c. (see Figure 2) [3] So, if $\left|p-\frac{1}{2}\right|$ is greater than

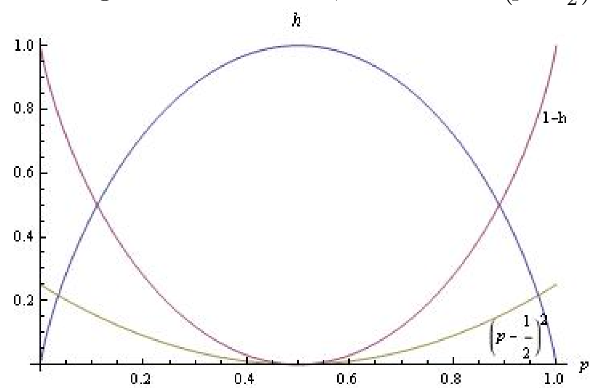
$$\sqrt{\frac{\log n+1}{cn}+k},$$

then $d(x) > k$, and since d_0 additively dominates d, no such x can be incompressible. That is, all incompressible strings have approximately half zeroes and half ones in them.

(End of example)

[3] We want to lower bound $(1-h(p))$ by a small-degree polynomial. It is impossible to do this with positive linear functions, and a quadratic polynomial is enough. A quadratic polynomial thus gives the greatest estimate for the distance $\left|p-\frac{1}{2}\right|$ with $d(x) \leq k$.

Figure 2: Plots of h, $1-h$ and $\left(p-\frac{1}{2}\right)^2$



4.2 Computable Distributions

Most of the material in this subsection is adapted from [1]. We now investigate randomness with respect to non-uniform probability distributions.

Definition 4.2.1. A probability distribution over strings is said to be *computable* if it is both lower and upper semicomputable.

In the following discussion, we will consider *positive* probability measures - that is, the probability of every string is strictly greater than 0. For an arbitrary computable probability distribution, the notion of a deficiency test will be more stringent than that for the uniform distribution. In particular, we consider tests of the following form.

Definition 4.2.2. A lower semicomputable real-valued function d over strings is said to be an *integrable* test if

$$\sum P(x)2^{d(x)} \leq 1.$$

The probability of a string having d at least k, is at most 2^{-k}, as can be seen using the Markov inequality.

We could contrast these tests with Martin-Löf tests. The uniform distribution over n-length strings is a computable distribution. However, since we are considering a restricted class of tests in this section, the deficiency functions for uniform distribution in this section may attain lesser values than the ones in the previous section.

An integrable test d_0 is *universal* if it additively dominates every other integrable test. That is, for every integrable test d,

$$d(x) < d_0(x) + O(1),$$

for every string x.

To prove that there is a universal integrable test for a computable positive probability P, we define a universal semimeasure $\mathbf{M}$ - that is, for every lowersemicomputable semimeasures p, we have $p(x) < O(1)\mathbf{M}(x)$ for all strings x. The universal integrable test will depend on the ratio of probability that $\mathbf{M}$ places on a given string to that placed by P on the same string.

Theorem 4.2.3 (Levin)**.** *There is a universal lowersemicomputable semimeasure.*

The proof can be found in the additional notes.

Let us denote this semimeasure by $\mathbf{M}$. Then, for any lowersemicomputable semimeaure μ, let us define

$$\mathbf{M}(\mu) = \sum\{\mathbf{M}(p) \mid p \text{ computes } \mu\}.$$

Then we have the following theorem.

Theorem 4.2.4 (Levin)**.** *If μ is a lowersemicomputable semimeasure, then for any string x, we have*

$$\mu(x)M(\mu) < O(1).\mathbf{M}(x).$$

The proof can be found in the additional notes.

With this, we can now define the universal integrable test for a computable probability measure.

Theorem 4.2.5. *Let P be a computable positive probability. The integrable test $d_0 \colon \Sigma^* \to [0, \infty)$ defined by*

$$d_0(x) = \log \frac{\mathbf{M}(x)}{P(x)} = -\log \mathbf{M}(x) - K(x)$$

is universal over all integrable tests for P.

Proof. □

4.3 Randomness for Infinite Sequences via Martingales

We introduce the notion of martingales which is strongly analogous to the concept of integrable tests, and show that we can characterize randomness for infinite sequences using martingales.

Martingales evolved out of a betting strategy in gambling. Suppose there are two outcomes of an an experiment, heads and tails, which occur with equal probability. It is common sense that it is impossible to win betting on such a system. But consider the following scheme. A player bets a rupee on heads. If the outcome is heads, the player wins 2 Rupees. (This is how a fair betting system works - if you bet n on an outcome which has probability p, you get n/p if the outcome happens.) So his/her net gain is 1 Rupee and (s)he leaves the game.

If the player loses, (s)he bets 2 Rupees that the second outcome is a head. If it turns out a head, the player gets 4 Rupees - (s)he has lost 3 Rupees in betting so far, so his/her net gain is 1 Rupee and (s)he pull outs of the game. If the outcome is a tail, the player bets 4 Rupees that the third outcome is a head, and so on. Every time the player loses, (s)he doubles the bet for the next head. The moment (s)he wins, (s)he pull out of the game. It is easy to analyze that this scheme assures me of a win of 1 Rupee. This scheme can be generalized. If the player start with n Rupees, by this scheme, the player will always be assured of a gain of n Rupees when the player leave the game.

What is a flaw in the above scheme? How was the player able to win betting against such a random outcome? One of the flaws is that the player is losing money exponentially fast. In order to gain by a round m, the player should have started the game with a bank account of $2^{m+1} - 1$. Also, there is a positive probability that no heads occurs in m rounds - in this event, we lose $2^{m+1}-1$ rupees, a catastrophic loss. [4] Thus this scheme performs poorly in practice.

The characteristic of this scheme can be succinctly expressed as: the *expected* amount of money after a bet is the same as the amount before the bet. Compare this with the definition of an independent identically distributed process, where the *probabilities* are the same from round to round. This scheme led to a mathematical abstraction which forms a very powerful tool in probability theory, called the *martingale process*. We will study a very restricted version of this tool, specialized to the study of infinite binary sequences.

First, we have to define a probability measure on Σ^∞, the space of infinite binary sequences. It is customary to consider a particular kind of probability defined on finite strings and extend the definition to infinite sequences. The idea behind the extension is that

$$Pr\{\omega \mid x \text{ is a prefix of } \omega\} = Pr[x],$$

that the probability of a set of infinite sequences with the common prefix x is the same as $Pr[x]$.

Definition 4.3.1. A function $P\colon \Sigma^* \to [0,1]$ is a *probability measure on* Σ^∞ if P obeys the following conditions.

1. $P(\lambda) = 1$.
2. For any string x, $P(x) = P(x0) + P(x1)$.

The first condition expresses the fact that every infinite sequence has λ as a prefix, and the space Σ^∞ has probability 1. The second condition expresses the fact that the set of sequences starting with x is the disjoint union of the set of sequences starting with $x0$ and the set starting with $x1$, and probability has to be additive over a disjoint union.

Example 4.3.2. The uniform probability distribution over Σ^∞ is the following distribution $\mu\colon \Sigma^* \to [0,1]$: For any string x,

$$\mu(x) = 2^{-|x|}.$$

It is easy to verify that μ satisfies the condition for a probability measure over Σ^∞.

Once we have a probability distribution over Σ^∞, we can define a martingale over infinite sequences. Again, our strategy is to define the functions for finite strings, obeying a certain additive property.

Definition 4.3.3. Let $P\colon \Sigma^* \to [0,1]$ be a computable probability measure on the sample space Σ^∞. A P*-martingale* is a function $\partial\colon \Sigma^* \to [0,1]$ with the properties:

1. $\partial(\lambda) \leq 1$.
2. For any string x,

$$\partial(x)P(x) = P(x0)\partial(x0) + P(x1)\partial(x1).$$

[4] The book “The Black Swan” by Nassim Nicholas Taleb talks more about how such events deemed to have low probability pulled the economy down during the downturn of 2008-09.

The first condition says that the martingale starts with a finite amount of money. The second condition says that expected value after betting on xb where b is a bit, is the same as the expected value at x. This models a fair betting game with the probability of an outcome x being $P(x)$.

To proceed further, we will generalize the concept of a martingale slightly to that of *submartingales* - where the equality in condition 2 is replaced by an inequality - the intuition being that we allow some money to be taken away after the next bet, modelling an unfair game where the house (betting agency) might take away some money after a bet.

Definition 4.3.4. Let $P\colon \Sigma^* \to [0,1]$ be a computable positive probability measure on the sample space Σ^∞. A P-*submartingale* is a function $\partial\colon \Sigma^* \to [0,1]$ with the properties:

1. $\partial(\lambda) \leq 1$.
2. For any string x,

$$\partial(x)P(x) \geq P(x0)\partial(x0) + P(x1)\partial(x1).$$

Thus the expected value before a bet is at least that after the bet.

A lower semicomputable P-submartingale d_0 is said to be *universal* if for any lower semicomputable P submartingale d,

$$d(x) < O(1).d_0(x)$$

for any string x.

Theorem 4.3.5. *Let P be a computable positive probability measure. The lower semicomputable P-submartingale ∂_0 defined on strings by*

$$\partial_0(x) = \frac{\mathbf{M}(x)}{P(x)}$$

is universal.

Proof. We can show that ∂_0 is lower semicomputable since $\mathbf{M}$ is lower semicomputable and P is computable. ∂_0 is a submartingale because $\partial_0(\lambda)$ is $\mathbf{M}(\lambda)/P(\lambda)$, which is at most 1, and for any string x, the following holds.

$$\begin{aligned}\partial_0(x)P(x) = \mathbf{M}(x) &\geq \mathbf{M}(x0) + \mathbf{M}(x1)\\ &= \frac{\mathbf{M}(x0)}{P(x0)}P(x0) + \frac{\mathbf{M}(x0)}{P(x1)}P(x1)\\ &= d(x0)P(x0) + d(x1)P(x1).\end{aligned}$$

The inequality in the first line is due to the fact that $\mathbf{M}$ is a semimeasure.

To show that ∂_0 is universal, consider an arbitrary lower semicomputable P-martingale d. Then we can verify that $\nu(x) = d(x)P(x)$ is a lower semicomputable semimeasure. We know that there is a program of length $K(d) + K(P) + O(1)$ which computes ν. Thus

$$\nu(x)2^{-[K(d)+K(p)]} < O(1)\mathbf{M}(x),$$

implying

$$d(x)P(x)2^{-[K(d)+K(p)]} < O(1)\mathbf{M}(x).$$

Rearranging and substituting for ∂_0, we get

$$d(x) < \partial_0(x)2^{K(d)+K(p)}.O(1),$$

which completes the proof. $\square$

In fact, it is easy to see the following fact.

Lemma 4.3.6. *Let P be a computable probability measure. If d is a lower semicomputable P-submartingale, then $\log d$ is an integrable test for the probability measure P.*

We now define random sequences in terms of the money that the martingale makes on prefixes of the infinite sequence. Since we are dealing with martingales which are sort of computable, it would be reasonable to call an infinite sequence *nonrandom* if such a martingale can make a lot of money on it. For example, we could try defining a sequence to be nonrandom if the limiting value of the money that a lowersemicomputable semimartingale attains on it, is infinity. However, the sequence of capital of a martingale need not have a limiting value - it could infinitely often grow to an unprecedented level, and between periods of growth, reach arbitrarily low positive values. For a sequence to be *nonrandom*, we require only that the martingale increases its capital infinitely often betting on its prefixes.

Definition 4.3.7. Let $P : \Sigma^* \to [0,1]$ be a computable positive probability distribution on Σ^∞. An infinite binary sequence ω is *constructively nonrandom* with respect to P if there is a lowersemicomputable P-martingale ∂ such that

$$\forall N \in \mathbb{N}\, \exists n \quad \partial(\,\omega[0 \dots n-1]\,) > N.$$

In this case, ∂ is said to *succeed* on ω. In other words, if we use a standard limiting notion from analysis, we have that ω is constructively nonrandom if

$$\limsup_{n\to\infty} \partial(\,\omega[0 \dots n-1]\,) = \infty.$$

By the universality property, it is clear that if any lowersemicomputable P-submartingale succeeds on ω, then so does ∂_0. Thus, ω is *random* if ∂_0 can win at most a finite amount on prefixes of ω. In the subsequent discussion, we will limit to the behavior of ∂_0.

It is clear by Markov inequality that the probability of the set of strings on which ∂_0 attains k is at most $1/k$. Let us denote this set by A_k. In other words,

$$P(A_k) = P\{\omega \mid \exists n\; \partial_0(\,\omega[0 \dots n-1]\,) > k\} < \frac{1}{k}.$$

Since a nonrandom sequence is one where a martingale should increase its money beyond any given number, every such sequence is a member of the set

$$\bigcap_{k=1}^{\infty} A_k = \bigcap_{k=1}^{\infty} \{\omega \mid \exists n\; \partial(\,\omega[0 \dots n-1]\,) > k\}.$$

Since for every positive number k, A_{k+1} is a subset of A_k, the intersection above is the intersection of a nested sequence of sets, and hence has probability equal to

$$\lim_{k\to\infty} P(A_k) = \lim_{k\to\infty} \frac{1}{k} = 0.$$

Thus, we have the following theorem.

Theorem 4.3.8 (Martin-Löf)**.** *Let P be a computable positive probability measure. Then the set of constructively random sequences has probability 1.*

Thus, if we pick a number in the unit interval "at random" according to distribution P, the probability that the sequence representing the number we choose will be a P-random, is 1. With probability 1, all long enough prefixes of a sequence ω will have a small ratio for

$$\frac{\mathbf{M}(\omega[0 \ldots n-1])}{P(\omega[0 \ldots n-1])}.$$

Taking logarithms, we can say a random sequence will have small values for $-\log P(\omega[0 \ldots n-1]) + \log \mathbf{M}(\omega[0 \ldots n-1])$ and by a version of the coding theorem, the second term is approximately $-K(\omega[0 \ldots n-1])$, we have that random sequences have very low values for

$$-\log P(\,\omega[0 \ldots n-1]\,) - K(\,\omega[0 \ldots n-1]\,)$$

for all large enough n. If we are given just $P(x)$ for some x, then our best guess for the cardinality of the sample space would be $1/P(x)$, since we would do no worse than thinking that P was the uniform distribution on some unknown sample space. In this case, it would take $-\log P(x)$ bits to describe a point in the sample space. Thus, $\log \partial_0(x)$ measures the deficiency of $K(x)$ from this estimate.

This is the sense in which we say that all large enough prefixes of a P-random sequence are incompressible.

4.4 Bibliographic Notes

Most of the material in section 4.1 and 4.2 is based on [1].

References

[1] Peter Gács. Lecture notes on descriptional complexity and randomness. Available at `http://www.cs.bu.edu/faculty/gacs/papers/ait-notes.pdf`.