# The Miraculous Universal Distribution

Walter Kirchherr* — San Jose State University

Ming Li† — University of Waterloo

Paul Vitányi‡ — CWI and University of Amsterdam

April 2, 1997

What is it, exactly, that scientists do? How, exactly, do they do it? How is a scientific hypothesis formulated? How does one choose one hypothesis over another?

It may be surprising that questions such as these are still discussed. Even more surprising, perhaps, is the fact that the discussion is still moving forward, that new ideas are still being added to the debate. Certainly most surprising of all is the fact that, over the last thirty years or so, the normally concrete field of computer science has provided fundamental new insights.

Scientists engage in what is usually called *inductive reasoning*. Inductive reasoning entails making predictions about future behavior based on past observations. But defining the proper method of formulating such predictions has occupied philosophers throughout the ages.

In fact, the British philosopher David Hume (1711 – 1776) has argued convincingly that in some sense proper induction is impossible, [3]. It is impossible because we can only reach conclusions by using known data and methods. Therefore, the conclusion is logically already contained in the start configuration. Consequently, the only form of induction possible is deduction. Philosophers have tried to find a way out of this conundrum. To see where the discussion stands today, let's put ourselves in the position of a budding young scientist with a specific prediction to make.

Let's follow the young Alice as she tries to win a bet.

*Address: Mathematics Department, San Jose State University, San Jose, California, USA

†Supported in part by the NSERC Operating Grant OGP0046506, ITRC, a CGAT grant, and the Steacie Fellowship. Current address: Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong. On sabbatical leave from: Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: mli@cs.cityu.edu.hk

‡Partially supported by the European Union through NeuroCOLT ESPRIT Working Group Nr. 8556, and by NWO through NFI Project ALADDIN under Contract number NF 62-376 and NSERC under International Scientific Exchange Award ISE0125663. Address: CWI, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands. Email: paulv@cwi.nl

## Alice Is Offered a Bet

Alice, walking down the street, comes across Bob, who is tossing a coin. He is offering odds to all passers-by on whether the next toss will be heads or tails. The pitch is this: he'll pay you two dollars if the next toss is heads; you pay him one dollar if the next toss is tails. Alice is intrigued. Should she take the bet? Certainly, if Bob is tossing a fair coin, it's a great bet. Probably she'll win money in the long run. After all, she would expect that half Bob's tosses would come up heads and half tails. Giving up only one dollar on each heads toss and getting two for each tails – why, in a while she'd be rich!

Of course, to assume that a street hustler is tossing a fair coin is a bit of a stretch, and Alice is no dummy. So she watches for a while, recording how the coin comes up for other betters, writing down a '1' for 'heads' and a '0' for 'tails'. After a while she has written 01010101010101010101. Perhaps Bob manipulates the outcomes. Common sense tells Alice that she can expect foul play when she plays with Bob.

What's her next move?

## Research

Alice is now equipped with data (her record of the tosses she has observed) and needs to formulate a hypothesis concerning the process producing her data – something like, 'The coin has a probability of 1/2 of coming up heads'. Or 'The coin alternates between heads and tails'. Which should it be? How does one formulate a hypothesis? As we said, Alice is no dummy. Like any good scientist, she first does her research. She first checks out what the great thinkers of the past had to say about it.

## Epicurus

The Greek philosopher Epicurus (342? B.C. – 270 B.C.) is mainly known to us through the writings of the Roman poet Titus Lucretius Carus (100 B.C.? – 55 B.C.?), who, in his long poem *On the Nature of the Universe*, popularized Epicurus's stoic philosophy among the Roman aristocracy. (What we moderns usually mean by "epicurean" has little to do with Epicurus, by the way.) Lucretius offers explanations for many natural phenomena and human practices (for example, he says, plausibly, that fire was delivered to humans by lightning), but he also admits to phenomena for which we know of no single explanation. To quote him:

> There are some phenomena to which it is not enough to assign one cause. We must enumerate several, though in fact there is only one. Just as if you were to see the lifeless corpse of a man lying far away, it would be fitting to state all the causes of death in order that the single cause of this death may be stated. For you would not be able to establish conclusively that he died by the sword or of cold or of illness

> or perhaps by poison, but we know that there is something of this kind that happened to him.[9]

This *multiple explanations* approach is sometimes called the *principle of indifference*. The British philosopher Bertrand Russell (1872 – 1970) summarizes it as follows: "When there are several possible naturalistic explanations ... there is no point in trying to decide between them."[10] In other words:

> **Principle of Indifference**: Keep all hypotheses that are consistent with the facts.

(To be fair, it should be pointed out that Epicurean philosophy is not concerned with scientific progress but rather with human happiness.)

## William of Ockham

The English cleric William of Ockham (1285 – 1349) is credited with formulating a different principle commonly called "Occam's Razor". He wrote, "Entities are not to be multiplied without necessity". Again, according to Bertrand Russell [10]: "That is to say if everything in science can be interpreted without assuming this or that hypothetical entity, there is no ground for assuming it." As popularly interpreted, we have:

> **Occam's Razor**: Among all hypotheses consistent with the facts, choose the simplest.

This is taken as given by most scientists and sometimes even explicitly stated. The great mathematician John von Neumann wrote,

> ...the sciences do not try to explain, they hardly try to interpret, they mainly make models. ...The justification (of a model) is solely and precisely that it is expected to work. ...Furthermore, it must satisfy certain aesthetic criteria – that is, in relation to how much it describes, it must be simple. [12]

Of course there are problems with this. Why should a scientist be governed by 'aesthetic' criteria? What is meant by 'simple'? Isn't such a concept hopelessly subjective?

We're wading in deep waters now. However, we are wading not alone but together with the greatest scientist of all time, "Fortunate Newton, happy childhood of science!" in Einstein's phrase. Isaac Newton (1642 – 1727) formulated in his *Principia* [8]:

> **Newton's Rule #1 for doing natural philosophy:** We are to admit no more causes of natural things than such as are both true and sufficient to explain the appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.

## Thomas Bayes

The English mathematician and cleric (clerics keep popping up in all this), Rev. Thomas Bayes (1702 – 1761) offered what is in essence a modified principle of indifference. Rather than accepting all hypotheses consistent with the facts *as equal* , he gave a method of assigning probabilities to hypotheses.

> **Bayes's Rule**: The probability that a hypothesis is true is proportional to the prior probability of the hypothesis multiplied by the probability that the observed data would have occurred assuming that the hypothesis is true.[2]

Suppose we have a priori a distribution of the probabilities $P(H)$ of the various possible hypotheses. We want the list of hypotheses to be exhaustive and mutually exclusive so that $\sum P(H) = 1$, the summation taken over all possible hypotheses $H$. Assume furthermore that for all such $H$ we can compute the probability $\Pr(D|H)$ [1] that sample $D$ arises if $H$ is the case. Then we can also compute the probability $\Pr(D)$ that sample $D$ arises at all:

$$\Pr(D) = \sum \Pr(D|H)P(H),$$

summed over all hypotheses. From the definition of conditional probability it is now easy to derive the familiar mathematical form of **Bayes's Rule**:

$$\Pr(H|D) = \frac{\Pr(D|H)P(H)}{\Pr(D)}.$$

Despite the fact that Bayes's Rule essentially rewrites the definition of conditional probability, and nothing more, it is its interpretation and application that are profound and controversial. The different $H$'s represent the possible alternative hypotheses concerning the phenomenon we wish to discover. The term $D$ represents the empirically or otherwise known data concerning this phenomenon. The factor $\Pr(D)$, the probability of data $D$, is considered as a normalizing factor so that $\sum \Pr(H|D) = 1$, the sum taken over all hypotheses.

The factor $P(H)$ is called the *a priori, initial*, or *prior* probability of hypothesis $H$. It represents the probability of $H$ being true before we have obtained any data. The prior probability $P(H)$ is often considered as the experimenter's *initial degree of belief* in hypothesis $H$.

The factor $\Pr(H|D)$ is called the *final*, *inferred*, or *posterior* probability, which represents the adapted probability of $H$ after seeing the data $D$. In essence Bayes's Rule is a mapping from prior probability $P(H)$ to posterior probability $\Pr(H|D)$ determined by data $D$.

Continuing to obtain more and more data, and repeatedly applying Bayes's Rule using the previously obtained inferred probability as the current prior, eventually the inferred probability will concentrate more and more on the "true" hypothesis. It is important to understand that one can find the true hypothesis also, using many

[1] We use notation "Pr()" to distinguish computed probabilities from prescribed probabilities like the *a priori* probability "$P()$."

examples, by the law of large numbers. In general, the problem is not so much that in the limit the inferred probability would not concentrate on the true hypothesis, but that the inferred probability give as much information as possible about the possible hypotheses from only a *limited number* of data. Given the prior probability of the hypotheses, it is easy to obtain the inferred probability, and therefore to make informed decisions.

Thus, with Bayes's rule, we keep all hypotheses that are consistent with the observations, but we consider some more likely than others. As the amount of data grows, we home in on the most likely ones.

But there is a nasty little phrase there: "the experimenter's initial belief in the hypothesis". How can a neutral observer have such an initial belief? How can the process of assigning probabilities get started? This is known as the problem of assigning *a priori* probabilities.

As a historical note: The memoir [2] was posthumously published in 1764 by Bayes's friend the Rev. Richard Price. Properly speaking, Bayes's Rule as given is not due to Bayes. Pierre-Simon, Marquis de Laplace, whom we will meet again later in this narrative, stated Bayes's Rule in its proper form and attached Bayes's name to it in [5].

## Where Does This Leave Alice?

So now that Alice knows the thoughts of the ancients, what should she do. Should she take the bet? Her basic question is, "What process (that is, what kind of coin) caused the sequence 01010101010101010101?" That's a tough one, so like any good scientist, she first tries to answer a simpler question: Is Bob tossing a fair coin (one where heads and tails are equally likely to come up) or not? Since a fair coin *could* cause such a sequence, Epicurus says we can't reject that hypothesis. (But we can't reject a lot of other hypotheses, either.) Occam says: accept the fair coin hypothesis if it is simpler than any other. (But he offers no help in determining if 'the probability of heads is 1/2' is simpler than, say, 'the probability of heads is 1/3'.) Bayes's rule has some intuitive appeal here. The sequence doesn't seem *likely* to have resulted from tossing a fair coin. But why not? What does Alice expect a fair-coin sequence to look like?

## Randomness

What bothers Alice is that the sequence of coin tosses doesn't look 'random'. She expects that a fair coin produces a random sequence of heads and tails. But what is 'random'? She has intuition about the concept, to be sure – 00101110010010111110 looks more random than 01010101010101010101 – but, precisely, what is meant by 'random'?

More research is called for. Again, let's review the thoughts of the sages. Dr Samuel Johnson (1709 – 1784), the great eighteenth-century master of conversation, had something to say on just about all topics. His biographer James Boswell, describes him as quite adept at and amused by mathematics. Boswell wrote:

> Johnson was quite proficient in mathematics. ... Dr. Beattie observed, as something remarkable which had happened to him, that he chanced to see both the No. 1 and the No. 1000 of the hackney-coaches, the first and the last. "Why sir," said Johnson, "there is an equal chance for one's seeing those two numbers as any other two." He was clearly right; yet the seeing of two extremes, each of which is in some degree more conspicuous than the rest, could not but strike one in a stronger manner than the sight of any other two numbers.[1]

Many of us (including Alice) would agree with Boswell. (Most of us are not Samuel Johnson, of whom it was also said, "There's no arguing with Johnson; for when his pistol misses fire, he knocks you down with the butt end of it.") But why are the two numbers Dr. Beattie observed more 'conspicuous' than any other two? What does that mean? Aren't these two numbers just as likely as any other two numbers (all pairs with equal probability 1/1,000,000)?

The great French mathematician Pierre-Simon Laplace (1749 – 1827) addressed the question why our intuition tells us that a regular outcome of a random event is unlikely:

> We arrange in our thought all possible events in various classes; and we regard as *extraordinary* those classes which incluFor the different prediction question, while a very small number. In the game of heads and tails, if head comes up a hundred times in a row, then this appears to us extraordinary, because the almost infinite number of combinations that can arise in a hundred throws are divided into regular sequences, or those in which we observe a rule that is easy to grasp, and into irregular sequences, that are incomparably more numerous. [5]

But what is regular and what is irregular? If Alice could convince herself that the particular sequence she observed is random, she could reasonably assign a high probability to the hypothesis that Bob is tossing a fair coin, and she should take the bet he offered.

(In fact betting strategies were a basis for early definitions of randomness – in essence, a sequence of $n$ coin tosses is random if you can't predict the $n$th toss by looking at the first $n-1$ tosses. But such definitions of randomness ran into difficulties when attempts were made to make them mathematically precise.)

Yet, the classical calculus of probabilities tells us that 100 heads are just as probable as any other sequence of heads and tails, even though our intuition tells us that it is less "random" than some others. Laplace distinguishes between the object itself and a cause of the object:

> The regular combinations occur more rarely only because they are less numerous. If we seek a cause wherever we perceive symmetry, it is not that we regard the symmetrical event as less possible than the others, but, since this event ought to be the effect of a regular cause or that of chance, the first of these suppositions is more probable than the second. On a table we see letters arranged in this order `C o n s t a n t i`

> `n o p l e`, and we judge that this arrangement is not the result of chance, not because it is less possible than others, for if this word were not employed in any language we would not suspect it came from any particular cause, but this word being in use among us, it is incomparably more probable that some person has thus arranged the aforesaid letters than that this arrangement is due to chance."

Let us try to turn Laplace's argument into a formal one. Suppose we observe a binary string $s$ of length $n$ and want to know whether we must attribute the occurrence of $s$ to pure chance or to a cause. "Chance" means that the literal $s$ is produced by fair coin tosses. "Cause" means that there is a causal explanation for $s$ having happened—a causal explanation that takes $m$ bits to describe. The pure chance of generating $s$ *itself* literally is about $2^{-n}$. But the probability of generating a *cause* for $s$ is at least $2^{-m}$. In other words, if there is some simple cause for $s$ ($s$ is regular), then $m \ll n$, and it is about $2^{n-m}$ times more likely that $s$ arose as the result of some cause than literally by a random process. It now remains to make this intuition operational.

## Computer Science to the Rescue

In the mid-1960's three men – Ray Solomonoff, Andrei N. Kolmogorov, and Gregory Chaitin – independently invented the field now generally known as Kolmogorov complexity. (Actually, Solomonoff was the earliest by a couple of years, and Chaitin last, but Kolmogorov in the middle was already world famous and his mathematics impeccable, and his name got attached to the field. As Billie Holliday sang, "Them that's got shall get. Them that's not shall lose. So the Bible says, and, Lord, still it's true.") Solomonoff was addressing Alice's problem with Bayes's formula – how do we assign *a priori* probabilities to hypotheses when we begin an experiment? Kolmogorov and Chaitin were addressing Alice's problem of defining precisely what is meant by a random sequence. All three saw that the notion of 'computable' lay at the heart of their questions. They arrived at equivalent notions, showing that these two questions are fundamentally related, and made major strides toward answering the age-old questions described above. (An extensive history of the field can be found in [7].)

For those of us living in the computer age, the notion of 'computable' is pretty much intuitive. Many of us have written programs; most of us have run computers. We know what a computer program is. It's a finite list of instructions telling the computer how to calculate a particular function. And once you know that, the notion of a computable binary sequence or "string" is both natural and straightforward.

Consider a particular computer language like FORTRAN or C++. Once we fix the language we can look at the programs that take no input – you just start them running, and some of them print out a binary string and stop. There are infinitely many such programs. Now let $s$ be a particular binary string. Some of the programs print out $s$; in fact, infinitely many of them do. Let's list them: $P_1, P_2, P_3, \ldots$, and so on to infinity. Among these programs there is a *shortest* one. (Remember that programs themselves are binary strings – think of object code – so we can talk about the length of a program.)

Here's the key definition: The *complexity* of a binary string, $s$, is the length of a shortest program which, on no input, prints out $s$.

We'll use $C(s)$ to denote 'the complexity of $s$' (but we'll shortly replace it with a variant denoted as $K(s)$ that has the technical properties we need—just be prepared).

Once we have the definition down pat, we get a few easy facts. Suppose $s$ is a string $n$ bits long.

1. $C(s) \leq n$ (plus some constant).

   This says that there is a program not much longer than $s$ which will print out $s$. The one-line program, 'print(s)', will do the job.

2. There is a string, $s$, for which $C(s) \geq n$.

   This is true since there are only $2^n - 1$ programs of length less than $n$, but $2^n$ strings of length $n$. So at least one string must satisfy this inequality.

A string satisfying the second fact above may be called "random." Why? Such a string is its own shortest description. In other words, it contains no pattern. For example, with the phrase, 'a string of 10,000 zeros', one can describe a 10,000-bit string with just a few letters. But (plausibly) the shortest way to describe the string '00100011101011010010' is by writing it out.

So using this idea, Alice is justified in feeling that 01010101010101010101 is not random. She can describe it with a pattern: "ten alternating 0's and 1's". But how does that help her formulate a hypothesis about the nature of the coin Bob is flipping?

That's where Solomonoff's ideas come in. Remarkably, his initial idea gave rise to a way of assigning probabilities to binary strings. If we define $K(s)$ pretty much as $C(s)$ explained above (there are some technical details we're postponing until the next section), we can assign a probability to $s$ as follows:

$$P(s) = 2^{-K(s)}$$

$2^{-K(s)}$ is called the *a priori* probability of $s$. This assignment of probabilities is called the *semi-measure universal for the class of enumerable semi-measures.* We'll just call it the **universal distribution**.

But so what? What exactly does this mean? (And how does it relate to Alice's problem?) First recall the three ancient principles for formulating hypotheses: the principle of indifference; Occam's razor; and Bayes's rule. As pointed out earlier, Bayes's rule is in a sense a refinement of the principle of indifference. The importance of *a priori* probability is that it neatly combines all three principles.

Look at it this way: any sentence can be coded into a series of 0's and 1's. Hypotheses are sentences; so they can be coded as binary strings. Since *a priori* probability assigns a probability to every binary string, it assigns a probability to every hypothesis. But there's more. 'Simple' hypotheses – the ones you favor under Occam's razor – are precisely those with small complexity.

And if the complexity is small, the *a priori* probability is big. So with this method of assigning probabilities to hypotheses – as required by Bayes's rule – we make the simplest ones most probable – as William of Ockham said we should.

Let's look at an example. Suppose we have two working hypotheses, $H_1$ and $H_2$. Occam's razor says we should favor the simpler one. In this approach, that means that we should favor the one with the lower complexity (the one with the shorter description). Bayes's formula (as with the principle of indifference) says we should keep them both, but assign probabilities to each one. The universal distribution satisfies both Occam and Bayes. And Epicurus, too! Suppose the shortest description of $H_1$ is 100 bits long, and the shortest description of $H_2$ is 200 bits long. Then we conclude that the probability of $H_1$ being the correct explanation is $1/2^{100}$ or about $8 \times 10^{-31}$, and that the probability of $H_2$ being the correct explanation is $1/2^{200}$, or about $6 \times 10^{-61}$.

These numbers also determine their *relative* probabilities, so that we can choose the most likely one: $H_1$ is about $10^{30}$ times more likely than $H_2$.

We keep both hypotheses (satisfying Epicurus), assign probabilities to our 'initial beliefs' in them (as Bayes suggested we do), and favor the simpler one with a higher probability (so William of Ockham won't feel left out).

One simple theory ties up a couple of millennia of philosophy!

## Some Details

This section is for those readers who would like a few more details.

The first thing we want to do is to justify our calling a string which is its own shortest description "random." Why should this definition be preferable to any other we might come up with? The answer to that was provided by the Swedish mathematician Per Martin-Löf (who was a postdoc of Kolmogorov). Roughly, he demonstrated that the following definition "an $n$-bit string, $s$, is *random* iff $C(s) \geq n$" ensures that every such individual random string possesses *with certainty* all effectively testable properties of randomness that hold for strings produced by random sources *on the average*. To see where this goes think about the pre-Kolmogorov complexity traditional problems of whether or not the infinite sequence of decimal digits in $\pi = 3.1415\ldots$ can be distinguished from typical outcomes of a random source. To determine this, the sequence has been submitted to several statistical tests for randomness called, appropriately "pseudo-randomness tests." (For example, whether each digit occurs with frequency $1/10$ within certain fluctuations.) If $\pi$ would have failed *any* of these tests, then we would have said that the sequence is not random. Luckily, it satisfies all of them.

What Martin-Löf proved was this: suppose you come up with your own definition of such a 'statistical test exposing non-randomness.' If your definition is at all reasonable, then any string which meets Martin-Löf's definition also meets yours. Now, what is 'reasonable'? Here we have to examine our intuition. First of all, we feel that most strings are random, so we demand that of your definition. (Specifically, we demand that, of all the $2^n$ strings of length $n$, at least $2^n(1-1/n^2)$ of them do not fail your randomness test.) Second, we demand that there be a computer program to execute your statistical test — it must be effective. Technically, the set of all strings that *don't* meet your definition should be what mathematicians call *recursively enumerable*, which means that there is a computer program that determines if a string *fails* the randomness test — that is, is not random.

For instance, suppose you define random as "passing statistical test A." Now if an $n$-bit string, $s$, meets Martin-Löf's definition of randomness, we want to prove that it will pass statistical test A. Well, suppose it doesn't, in other words, suppose it's one of the at most $2^n/n^2$ strings that fail test A. Then here is a description of $s$:

The $m$th string of length $n$ which fails test A.

We know that $m$ is a number between 1 and $2^n/n^2$. We may not know what number $m$ *is*, but we know it's in that range. The length of that description (if we code it in binary) involves coding both $n$ (in $\log n$ bits) and $m$ (in $n - 2\log n$ bits). This comes to at most $n - \log n$ bits (plus some negligible terms which we ignore here); hence, we can conclude that $C(s) \leq n - \log n$. But then $s$ does not meet Martin-Löf's definition.

To see that Martin-Löf's definition actually is itself such a randomness test: In the first place we can approximate $C(s)$ by running all programs of length at most $s$ (plus some constant) for as long as it takes, in rounds of one step of each program. As soon as a program halts we check whether its output is $s$, and if so whether this program beats the current champion that output $s$ by being shorter. This way we approximate the shortest program better and better as time goes by. This process will for each $s$ eventually determine if it is *not* random. This shows also the second property, namely that less than $2^n/n^2$ strings fail Martin-Löf's test. For instance, let A test $C(s) \leq n$.

That's the justification for calling a string which is its own shortest description 'random'. It gets a bit stickier when you go into the details of the universal distribution. You have to be a bit careful when you talk about 'shortest description'. Of course, when we talk about the description of a string we mean, as mentioned above, a program which on no input will print out that string. But if we want to get probabilities out of all this, we are subject to a certain key restriction: the probabilities must add up to (no more than) 1. So we are faced with the task of establishing that the sum of all our *a priori* probabilities add up to no more than 1.

This almost killed Solomonoff's original idea. It was soon shown that, if we use the simple definition of shortest description, we get that for every $n$ there is an $n$-bit string, $s$, where the value of $C(s)$ is at most $\log n$. This means that for every $n$ there is a string $s$ with $P(s)$ at least $2^{-\log n}$ or $1/n$. And, of course, the infinite sum $1/2 + 1/3 + 1/4 + \ldots$ diverges – it's certainly not one or less!

It was about a decade before Solomonoff's idea was rescued—by Leonid A. Levin, another student of Kolmogorov. Chaitin had the same idea, but again later. The key was this: instead of considering the length of computer programs *in general*, only certain computer programs would be considered. Specifically, we would restrict our attention to *prefix-free* computer programs, that is, a set of programs, no one of which is a prefix of any other. (This is not too hard to imagine. For instance, if we design a computer language in which every program ends with the word 'stop', (and 'stop' may not appear anywhere else), the programs written in our language form a prefix-free set.)

The reason this approach saved the day is due to a key theorem proved in 1949 by L.G. Kraft (in his master's thesis at MIT [4]). For our purposes, it says in essence that if we restrict our attention to prefix-free sets, then the resulting *a priori* probabilities will sum to no more than one.

From now on we'll use this slightly different definition of $C(s)$ which we denote by $K(s)$. So $K(s)$ is the length of the shortest program for $s$ among all prefix-free syntactically correct programs in our fixed programming language.

So the universal distribution meets the requirements of probability theory. Now, what is our justification for calling it 'universal'? Briefly, it's this: suppose you have defined a probability distribution on strings. As long as it meets a reasonable criterion (namely, that it be *enumerable*, which is weaker than requiring that there is a computer program which, given $s$ as input, will print out the probability you assign to $s$), then the universal distribution dominates yours in the sense that there is some constant, $k$ which depends on your probability but not on $s$, for which $k\,\Pr(s)$ is as least at big as the probability you assigned to $s$. This is called "multiplicative domination" and was proved by Levin in the early 70's.

In a way, this is similar to the idea of a "universal" Turing machine which is universal in the sense that it can simulate any other Turing machine when provided with an appropriate description of that machine. It is universally accepted that the Turing machine is a mathematical precise version of what our intuition tells us is "computable," and therefore the universal Turing machine can compute all intuitively computable functions [11]. Since the latter statement is not a mathematical one, it cannot be proved: it is known as *Turing's Thesis*; in related form known as "Church's Thesis." Just like the Kolmogorov complexity is minimal (up to an additive constant) among all description lengths that can be approximated from above by a computational process, so does the universal distribution multiplicatively dominate (and is in a particular sence close to) each and every enumerable distribution—distributions that can be approximated from below by a computational process. Hence there are a lot of universalities here, and the Turing Thesis spawns:

> **Kolmogorov's Thesis:** "the Kolmogorov complexity gives the shortest description length among all description lengths that can be effectively approximated from above according to intuition. "

> **Levin's Thesis:** "The universal distribution gives the largest probability among all distributions that can be effectively approximated from below according to intuition."

## Repaying the Source

So the normally concrete field of computer science has contributed to an abstract philosophical debate which has occupied the ages. One can, in turn, use the philosophical ideas presented to contribute to the field of computer science. Several areas of computer science have benefited from the basic concept of *universal distribution*. We will look at one of them here – namely, the area called *algorithm analysis*. Algorithm analysis is concerned with determining the amount of time it takes to run a particular program. Of course, the amount of time a program takes depends on the size of the input to the program. For example, the number of steps required to sort $n$ numbers using the sorting technique computer scientists call *quicksort* [2] will, *in the worst case*, be proportional to $n^2$.

The italics are needed. Another way to approach algorithm analysis is to determine how fast a program runs *on the average*. This needs to be made precise. Let's look closer at the problem of sorting $n$ numbers. We may as well assume our input is the $n$ numbers $i_1, i_2, \ldots, i_n$ mixed up somehow, and we want to output them in order. The time required by quicksort is proportional to $n^2$ when the input is the numbers already sorted. But interestingly, for most inputs (where the numbers are not even close to being in order) the time required will be proportional to $n \log n$. In other words, there is an input which will force quicksort to use $n^2$ (or so) steps, but for most inputs, quicksort will actually run much faster than this. Now, *if all inputs are equally likely* the average running time for quicksort is much less than its running time in the worst case.

Again, the italics are needed. Are all inputs equally likely in 'real' life? Actually, it doesn't seem that way. In 'real' life it seems that much computer time is spent sorting lists which are nearly sorted to begin with. In any case, what interests us here is the remarkable relation between worst case running times and average case running times revealed by the universal distribution.

To reiterate, the term *average* implies the uniform distribution. The term *worst case* is independent of any distribution. It is somehow natural to assume that in many cases the average is better (in this case lower) than the worst case.

It came as a surprise when it was shown by two of us (ML and PV) [6] that if we assume the *universal distribution*, that is, if we assume that inputs with low complexity (ones with 'short descriptions') are more likely than inputs with high complexity, then the running time we expect under this distribution is (essentially) the worst-case running time of the algorithm.

This is not too hard to see in the case of quicksort. As we said, the worst case for quicksort is when the input is already sorted. In this case, the input has a short description (namely, "the numbers 1 through $n$ in order"), whereas if the input is all mixed up, it is 'random' and is its own shortest description. Under the universal distribution, the already sorted input is far more likely than the unsorted input, so we expect that quicksort will require $n^2$ steps.

But why is this true for algorithms in general? Again, it's not too hard to

[2] Briefly, the sorting algorithm known as quicksort is this:

- Rearrange the list of numbers into two smaller lists, the 'left half' and the 'right half' in such a way that every member of the left half is less than every member of the right half.
- Sort the left half, and sort the right half.
- Merge the two sorted halves into one sorted list.

see why, but the explanation is a bit more abstract. Suppose algorithm **A** runs fast on some inputs and slowly on others. Then the particular input which causes **A** to run slowest has a short description, namely, "that input of size $n$ which causes algorithm **A** to run slowest". This is a description of length $\log n$ (plus some constant) which describes a string of length $n$. So the length-$n$ string described has low complexity and is assigned a high probability under the universal distribution. This is intuitively the reason why the universal distribution assigns high enough probability to such simple strings to slow the average running time of **A** to its worst-case running time.

The universal distribution is a great benefactor for learning and induction; but it is so bad for average running time and all other reasonable computing resources like computer memory use, that such distributions are now called 'malignant' by computer scientists.

## The Universal Bet

We left Alice back there aways. Have we really helped her? After all, Bob is about to flip the coin again, and it's time for Alice to put up or shut up. Well, the fact is, we are forced into the theoretician's defense here. Yes, we have helped Alice. We have provided her with a solid theoretical framing of the problem she confronts.

And if she's clever, she can make a safe bet with Bob. In fact, Alice plays the stock market because just like Bob's offer the profits go up all the time. However, in the stock market investment companies tell you that "past performance is no guarantee for future performance." So Alice is used to cover her position with side bets called "puts" and "calls." So let's see how Alice can cover her position with Bob.

What she can propose is this: Bob flips his coin 1000 times and one part of the scheme is his original offer of two dollars for one dollar pay-out on "heads." Alice is a scientist, remember. This is lab work, and long lab hours are no deterrent to her. The results of these 1000 flips are recorded, yielding a string – let's call it $s$ – of 1000 1's and 0's representing heads and tails. With the second part of the scheme Alice covers her position: Alice pays Bob one dollar and Bob pays Alice

$$2^{1000-K(s)}$$

dollars. Now, if Bob's on the square, like the stock market, he has to take this side-bet, since his expected pay-out is less than \$1. This follows from Kraft's work mentioned above which sets the expected pay-out at

$$\sum 2^{-1000} 2^{1000-K(s)} \leq 1,$$

where the sum is taken over all binary strings $s$ of length 1000. The expected pay-out is actually smaller than 1 because there are programs that have length $\neq 1000$ and there are other programs than shortest programs.)

If Bob is as honest as Alice's stock broker (who accepts Alice's buying and selling orders including her side bet orders for puts and calls) Bob should be happy to accept Alice's proposal. In fact, he can expect to earn a little on the side-bet. But if Bob's crooked, and his flips do not result in a random string, but something like 000000...00000000000000 then he'll receive 1000 dollars from Alice on the main bet but he'll pay out something like $2^{1000-\log 1000}$ dollars on the side bet. That's about – well, who cares? Bob doesn't have that much money.

If Bob's honest, this is no worse than his original proposal. But if Bob has any brains, he'll pack up and move to another corner where Alice can't bother him, because in this game Alice wins big if Bob is honest — about five hundred bucks — and even bigger if he cheats!

Using the universal distribution, we have constructed the perfect *universal bet* that protects against all fraud!

But there's a catch – none of these schemes can actually be carried out. The complexity of a string is *non-computable*. Given a string $s$, there is *no way* to determine what $K(s)$ is. So Alice can't determine which of her hypotheses have low complexity and which do not. And the pay-off schemes she proposes to Bob can't be calculated. So it appears she's on her own.

But all is not lost. An idea as elegant as the universal distribution cannot be just tossed out. To make the universal side bet feasible, Alice can pay Bob one dollar and Bob pays Alice

$$2^{1000-\text{length of } p}$$

dollars for *any* prefix-free program $p$ that Alice can exhibit after the fact and that computes $s$. This involves a computable approximation "length of some program $p$ to compute $s$" of $K(s)$ = "length of the shortest program to compute $s$." Consequently, Alice may not win as much in case of fraud since length of $p \geq K(s)$ by definition (and Kolmogorov's thesis). But for the particular bet proposed by Bob, Alice *only cares* about compressibility based on deviating frequency of 1's since she can just bet that each bit will be 1.

## Wonderful Universal Induction

We started out by asking how learning and induction can take place at all and we have followed Alice in her quest to the universal distribution. Now for the full problem: from universal gambling to universal induction.

It is a miracle that this ages-old problem can be satisfactorily resolved by using the universal distribution as a "universal a priori probability" in Bayes's rule. Ray Solomonoff invented a perfect theory of induction. Under the relatively mild restriction that the true *a priori* distribution to be used in Bayes's rule is computable, it turns out that one can *mathematically prove* that using the single fixed universal distribution instead of the actually valid distribution (which may be different for each problem we want to apply Bayes's rule to) is almost as good as using the actually true distribution itself! In fact, this is the case both when we want to determine the most likely hypothesis and when we want to determine the best prediction — which are two different things.

They are two different things because the best single hypothesis does not necessarily give the best prediction. For example, consider a situation where we are given a coin of unknown bias $p$ of coming up "heads" which is either $p_1 = \frac{1}{3}$ or $p_2 = \frac{2}{3}$. Suppose we have determined that there is probability $\frac{2}{3}$ that $p = p_1$ and probability $\frac{1}{3}$ that $p = p_2$. Then the "best" hypothesis is the most likely one: $p = p_1$ which predicts a next outcome "heads" as having probability $\frac{1}{3}$. Yet the best prediction is that this probability is the expectation of throwing "heads" which is

$$\frac{2}{3}p_1 + \frac{1}{3}p_2 = \frac{4}{9}.$$

To take the prediction case: Solomonoff has shown that using the universal distribution the total expected prediction error over infinitely many predictions is less than a fixed constant (depending on the complexity of the true *a priori* distribution). This means that the expected error in the $n$th prediction goes down faster than $1/n$. This is good news, and in fact it is better news than any other inference method can offer us. It turns out that these ideas can be used to prove Occam's razor itself.

Traditional wisdom has it that the better a theory compresses the learning data concerning some phenomenon under investigation, the better we learn, generalize, and the better the theory predicts unknown data. This belief is vindicated in practice but apparently has not been rigorously proved in a general setting. Two of us [PV and ML] have recently shown that indeed optimal compression is almost always a best strategy in hypotheses identification. For the different prediction question, while the single best hypothesis does not necessarily give the best prediction, we demonstrated that nonetheless compression is almost always the best strategy in prediction methods.

Statisticians like Jorma Rissanen and Chris Wallace know about Alice's problem. They have translated Solomonoff's ideas into workable (that is, easily computable) forms called 'minimum description length' algorithms, [7]. Such algorithms help many Alice's get on with practical problems nowadays, such as video scene analysis, risk minimization, and even playing the stock market.[3] [4]

# 1 Acknowledgments

We thank Harry Buhrman, Richard Cleve, Lance Fortnow, and John Tromp for comments.

# References

[1] J. Boswell, *The Life of Samuel Johnson*, Doubleday and Co., 1945.

[3] The authors have been informed of a scientific stock investment company using software based on approximations of the universal distribution hitting the stock market in the near future. The authors decline *a priori* any responsibility for any happenings resulting from any methods disclosed in this article.

[4] All these things and the accomplishments of the heros in this story, Ray Solomonoff, Andrei N. Kolmogorov, Gregory Chaitin, Per Martin-Löf, Claus Schnorr, Leonid A. Levin, Péter Gács, Tom Cover (who introduced universal gambling and portfolio management), and many others, are explained in great detail in [7].

[2] T. Bayes, An essay towards solving a problem in the doctrine of chances, *Philos. Trans. Roy. Soc.*, 53:376-398 and 54:298-31, 1764.

[3] D. Hume, *Treatise of Human Nature, Book I*, 1739.

[4] L. G. Kraft, A device for quantizing, grouping, and coding amplitude modulated pulses, Master's thesis, Department of Electrical Engineering, M.I.T., Cambridge, Mass., 1949.

[5] P.S. Laplace, *A philosophical essay on probabilities*, 1819. English translation, Dover, 1951.

[6] M. Li and P.M.B. Vitányi, Worst case somplexity is equal to average case complexity under the universal distribution, *Information Processing Letters*, 42:145-149, 1992.

[7] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications, 2nd edition*, Springer-Verlag, 1997.

[8] I. Newton, *Philosophiae Naturalis Principia Mathematica* also known as the *Principia*, 1687.

[9] Titus Lucretius Carus, *The Nature of the Universe* (Ronald Latham, translator), Penguin Books, 1965.

[10] B. Russel, *A History of Western Philosophy*, Simon and Schuster, 1945.

[11] A.M. Turing, On computable numbers with an application to the Entscheidungsproblem, *Proc. London Math. Soc., Ser. 2*, 42(1936), 230-265; Correction, Ibid, 43(1937), 544-546.

[12] J. von Neumann, *Collected Works, Volume V*, Pergamon Press, 1963.