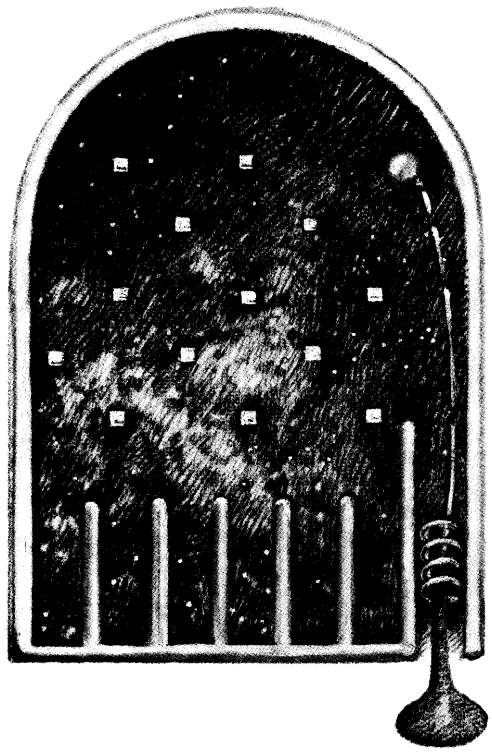


PROBABILITY and NONLINEAR SYSTEMS

by R. Daniel Mauldin

Part 1
An Introduction

Stan Ulam, at sixty-five, was vigorous, handsome, full of ideas. It was the spring of 1974, and he had come to lecture at the University of Florida, where I was a young assistant professor. I had known Stan by reputation for several years. In fact, the very first paper that I read as a part of my German language requirement in graduate school was his landmark 1930 paper on measurable cardinals, "Zur Masstheorie in der allgemeinen Mengenlehre." But listening to him in person was quite an inspiration. He did not lecture in the usual sense but presented snapshots of mathematical ideas, a style reminiscent of Steinhaus, one of Stan's teachers in Poland. Afterwards, several of us talked with him for a remarkably long time. I was immediately impressed with his ability to take up a mathematical topic and breathe new life into the subject.

The following year Stan took a position at Florida. His weekly seminar was similar to his book *A Collection of Mathematical Problems.* A topic would be brought up for discussion, and if it appeared to intrigue someone, we would return to it at a slightly deeper level. Stan soon became a stimulating source of encouragement to the younger mathematicians, and to me he became a mentor. As always, he was very generous in sharing his ideas. Throughout his life Stan nourished mathematics in that manner.

At first he would listen to us for a very short time—and then expound his own ideas. Eventually, however, our conversations became a witty (on his part) and very productive exchange. Like a master of reflecting boundaries, he would bounce ideas back to us from an endless variety of angles, especially humorous ones. The amplification of an idea could occur in a time span varying from a coffee conversation to a number of years. Although we would repeatedly go over the same topics, it wasn't exactly like working the beads on a rosary. Every so often an idea would undergo some adjustment or transformation, and something new, perhaps unexpected, would emerge. I don't know whether it was always his way to have short, quick discussions of some central idea, but that is certainly the impression one gets from perusing his comments and problems in *The Scottish Book.* (This famous notebook of problems was jotted down at the Scottish Cafe in Lwow during

the 1930s and first published in this country in 1957. See "Excerpts from *The Scottish Book.*")

Ulam's incredible feel for mathematics was due to a rare combination of intuitions, a common feature of almost all great mathematicians. He had a very good sense of combinatorics and orders of magnitude, which included the ability to make quick, crude, but in-the-ballpark estimates. Those talents, combined with the more ordinary abilities to analyze a problem by means of logic, geometry, or probability theory, already made him very unusual. Besides, he had a good intuition for physical phenomena, which motivated many of his ideas.

Ulam's intuition, as exhibited in numerous problems formulated over a span of more than fifty years, covered an enormous range of subjects. The problems on computing, physical systems, evolution, and biology were stimulated by new developments in those fields. Many others seemed to spring from his head. He usually had some prime examples in mind that motivated his choice of mathematical model or method. In this regard one of his favorite quotes, from Shakespeare's *Henry VIII*, was

> Things done without example
> in their issue
> Are to be feared.

In approaching a complicated problem Stan first searched for simplicity. He had no patience for complicated theories about simple objects, much less complex objects. That philosophical dictum happened to match his personality. He could not hold still for the time it would take to learn, let's say. modern abstract algebraic geometry, nor could he put up with the generalities of category theory. Also, he was familiar with, and early in his career obtained fundamental results in, measure and probability theories. That background led him to approach many problems by placing them in a probabilistic framework. Instead of considering just one possible outcome of a process, one can consider an infinite number of possible outcomes at once by randomizing the process. Then one can apply the powerful tools of probability, such as the laws of large numbers, to determine the likelihood of a given outcome. The famous Monte Carlo method is a perfect example of that approach. In fact, one of the favorite sayings of Erdos and Ulam, both of whom worked in combinatorics (in which the number of outcomes is finite) and probability, was

> The infinite we do right away;
> the finite takes a little longer.

Stan's interest in probability dates back to the early 1930s, when he and Lomnicki proved several theorems concerning its foundations. In particular, they showed how to construct consistent probability measures for systems involving infinite (as opposed to finite) sequences of independent random variables and, more generally, for Markov processes. (In Markov processes probabilities governing the future depend only on the present and are independent of the past.) At about the same time Kolmogorov, independently, proved his consistency theorem, which includes the Ulam and Lomnicki results as well as many more. Those results guarantee the existence of a probability measure on classes of objects generated by various random processes. The objects might be infinite sequences of numbers or more general geometrical or topological objects. such as the homeomorphisms (one-to-one, onto maps) discussed in detail later in this article. Stan's interest in probability continued after World War II, when he and Everett wrote fundamental papers on "multiplicative" processes (better known as branching processes). Those papers were stimulated by the need to calculate neutron multiplication in fission and fusion devices. (David Hawkins. in "The Spirit of Play," discusses some of the earliest work that Stan and he did on branching processes.)

Stan's background in probability made him a leader among the outstanding group of intellects who, during the late 1940s and early 1950s, recognized the potential value of the computer for doing experimental mathematics. They realized that the computer was an ideal tool for analyzing stochastic, or random, processes. While formal theorems gave rules on how to determine a probability measure on a space of objects, the computer opened up the possibility of generating those objects at random. Simply stated constructions that yield complicated objects could be implemented on the computer, and if one was lucky, *demonstrable* guesses could be made about their asymptotic, or long-term, behavior. That was the approach Stan took in studying deterministic as well as random recursions. In addition he invented cellular automata (lattices of cells and rules for evolution at each cell) and used them to simulate growth patterns on the computer.

The experimental approach to mathematics has since become very popular and has tremendously enhanced our vision of complex physical, chemical, and biological systems. Without the fortuitous conjunction of the computer and probability theory, it is very unlikely that we would have reached today's understanding of those nonlinear systems. Such systems present a challenge analogous to that Newton would have faced if the earth were part of a close binary or tertiary star system. (One can speculate whether Newton could have ever unraveled the law of gravitation from the complicated motions of such a system.) At present researchers are trying to formulate limiting laws governing the long-term dynamics of nonlinear systems that are analogous to the major limiting theorems in classical probability theory. The attempt to construct appropriate probability measures for such systems is one of the topics I will discuss in more depth.

Other interests that Ulam maintained throughout his life were logic and set theory. I remember a conference' on large cardinal numbers in New York a few years ago. Stan was the honored participant. More than fifty years earlier he had shown that if a nontrivial probability measure can be defined on all subsets of the real numbers, then the cardinal number, or "size," of the set of all the subsets exceeded the wildest dreams of the time. (See "Learning from Ulam: Measurable Cardinals, Ergodicity, and Biomathematics.") But that large cardinal of his is minuscule compared with the cardinals of today. After listening to some of the conference talks, Stan said that he felt like Woody Allen in *Sleeper* when he woke up after a nap of many years and was confronted with an unbelievably large number on a McDonald's hamburger sign.

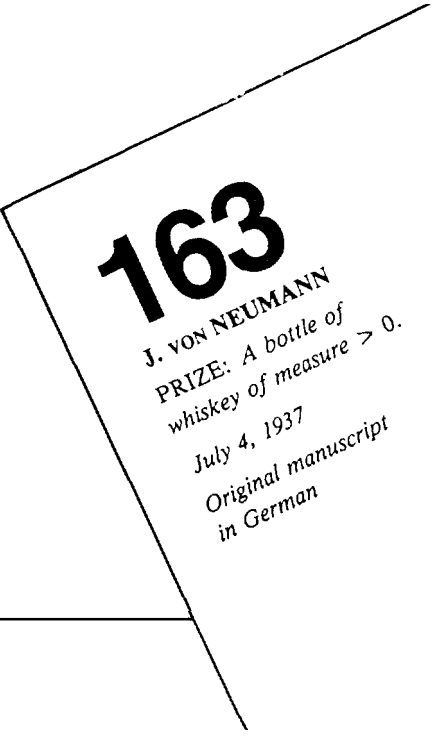
There is a serious aspect to that remark. Stan felt that a split between mathematics and physics had developed during this century. One factor was the trauma that shook the foundations of mathematics when Cantor's set theory was found to lead to paradoxes. That caused mathematics to enter a very introspective phase, which continues to this day. A tremendous effort was devoted to axiomatizing mathematics and raising the level of rigor. Physics, on the other hand, experienced an outward expansion and development. (The situation is somewhat reversed today, as internal issues concerning the foundations of physics receive attention.) As a result, university instruction of mathematicians has become so rigorous and demanding that the mathematical training of scientists has been taken over by other departments. Consequently, instruction in "applied" mathematics, or mathematical methods, is often at a fairly low level of rigor, and, even worse, some of the important mathematical techniques developed during this century have not made their way into the bag of tools of many physical scientists. Stan was very interested in remedying the situation and believed the Center for Nonlinear Studies at Los Alamos could play a significant role.

Stan was associated, either directly or through inspiration, with the three research problems described in Part 111 of this article. Each is an example of how a probabilistic approach and computer simulation can be combined to illuminate features of nonlinear systems. Since some background in modern probability theory is needed to follow the solutions to the problems, Part II provides a tutorial on that subject, which starts with a bit of history and concludes with several profound and useful theorems. Fortunately Mark Kac and Stan Ulam gave a very insightful summary of the development of probability theory in their book *Mathematics and Logic: Retrospect and Prospects.* I have adapted and extended their discussion to meet the needs of this presentation but have retained their broad perspective on the history of mathematics and, in some cases, their actual words.

Excerpts from the SCOTTISH BOOK

These excerpts from ***The Scottish Book: Mathematics from the Scottish Cafe are*** reprinted with permission of Birkhauser Boston. That 1981 edition of problems from "the book" kept at the Scottish Cafe was edited by R. Daniel Mauldin. The two earlier English-1anguage editions of the problems were edited by Stan Ulam, **the first being a 1957 mimeographed version** of Ulam's own translation into English from the languages originally inscribed in "the book."

Problems 18 and 19 are still unsolved, and the work stimulated by Problem 43 has played a major role in understanding the consequences of the axiom of choice.



18

ULAM

Let a steady current flow through a curve in space which is closed and knotted. Does there exist a line of force which is also knotted (knotted = nonequivalent through any homeomorphism of the whole space R_3 with the circumference of a circle)?

153

MAZUR

PRIZE: *A live goose*

November 6, 1936

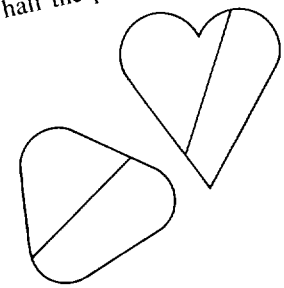
19

ULAM

Is a solid of uniform density which will float in water in every position a sphere?

Commentary. The two-dimensional version of the problem concerns a cylinder of uniform density which floats in every position, having the axis parallel to the water surface, and compatible with Archimedes' law. H. Auerbach ... showed that in the case of density 1/2, the cylinder need not be circular, or even convex, and gave a class of examples. We reproduce his illustration of two of them (Fig. 19.1).

Figure 19.1. Two possible solutions. The line segment rotates within the curve, and in each position cuts off half the area and half the perimeter.



43

MAZUR

PRIZE: One *bottle of wine, S. ULAM*

Definition of a certain game. Given is a set E of real numbers. A game between two players A and B is defined as follows: A selects an arbitrary interval d_1; B then selects an arbitrary segment (interval) d_2 contained in d_1; then A in his turn selects an arbitrary segment d_3 contained in d_2 and so on. A wins if the intersection $d_1, d_2, \ldots . d. \ldots$ contains a point of the set E: otherwise, he loses. If E is a complement of a set of first category, there exists a method through which A can win; if E *is* a set of first category, there exists a method through which B will win.

Problem. It is true that there exists a method of winning for the player A only for those sets E whose complement is, in a certain interval, of first category; similarly, does a method of win exist for B if E is a set of first category?

Addendum. Mazur's conjecture is true.

S. Banach, August 4, 1935

Modifications of Mazur's Game

(1) There is given a set of real numbers E. Players A and B give in turn the digits O or 1. E wins if the number formed by these digits in a given order (in the binary system) belongs to E. For which E does there exist a method of win for player A (player B)'?

U1am

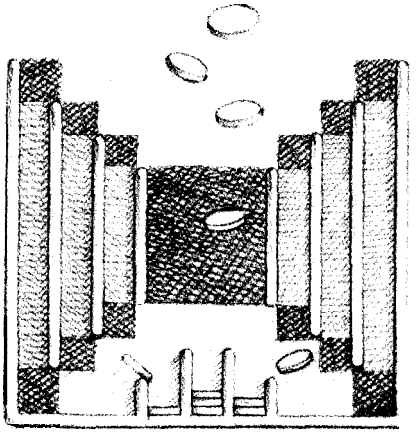
(2) There is given a set of real numbers E. The two players A and B in turn give real numbers which are positive and such that a player always gives a number smaller than the last one given. Player A wins if the sum of the given series of numbers is an element of the set E. The same question as for (l).

Banach

Commentary. The first published paper on genera] finite games with perfect information is Zermelo's .... Here, in Problem 43, we have the first interesting definition of an infinite one. ...

A TUTORIAL *on* PROBABILITY. MEASURE, *and the laws of* LARGE NUMBERS

Part II
PROBABILITY and NONLINEAR SYSTEMS



As mentioned in the introduction, Stan Ulam contributed to the measure-theoretic foundations that allow one to define a probability when the number of possible outcomes is infinite rather than finite. Here I will explain why this extension is so necessary and so powerful and then use it to introduce the laws of large numbers. Those laws are used routinely in probability and its applications (several times, for example, during solution of the problems discussed in Part III). Following the logic of Kac and Ulam I begin at the beginning.*

Early Probability Theory

Probability theory has its logical and historical origins in simple problems of counting. Consider games of chance, such as tossing a coin, rolling a die, or drawing a card from a well-shuffled deck. No specific outcome is predictable with certainty, but all possible outcomes can usually be listed or described. In many instances the number of possible outcomes is finite (though perhaps exceedingly large). Suppose we are interested in some subset of the outcomes (say, drawing an ace from a deck of cards) and wish to assign a number to the likelihood that a given outcome belongs to that subset. Our intuitive notion of probability suggests that that number should equal the ratio of the number of outcomes yielding the event (4, in the case of drawing an ace) to the number of all possible events (52, for a full deck of cards).

This is exactly the notion that Laplace used to formalize the definition of probability in the early nineteenth century. Let A be a subset of the set f) of all possible outcomes, and let $P(A)$ be the probability that a given outcome is in A. For situations such that 0 is a *finite* set and all outcomes in f are *equally probable,* Laplace defined $P(A)$ as the ratio of the number $u(A)$ of elements in A to the total number v(Q) of elements of f; that is,

$$P(A) = \frac{u(A)}{u(f))}.$$

However, the second condition makes the definition circular, for the concept of *probability* then is dependent upon the concept of *equiprobability.* As will be described later, the more modem definition of probability does not have this difficulty.

For now let us illustrate how Laplace's definition reduces the calculation of probabilities to counting. Suppose we toss a fair coin (one for which heads and tails

*The material quoted in this tutorial from *Mathematics and Logic* has been reprinted with permission from Encyclopedia Britannica, Inc.

are equally probable) n times and want to know the probability that we will obtain exactly m heads, where $1 < m < n$. Each outcome of n tosses can be represented as a sequence, of length n, of H's and T's (HTHH . THH, for example), where H stands for heads and T for tails. The set L? of all possible outcomes of n tosses is then the set of all possible sequences of length n containing only H's and T's, The total number of such sequences, v(Q), is 2^{n}. How many of these contain H exactly m times? This is a relatively simple problem in counting. The first H can occur in n positions, the second in n – 1 positions, ... , and the mth in (n – m + 1) positions. So if the H's were an ordered sample $(H_1, H_2, \ldots, H_n)$, the number of sequences with H's would equal $n(n-1)(n-2)\ldots(n-m+1)$. But since all the H's are the same, we have overcounted by a factor of m ! (the number of ways of ordering the H's). So the number of sequences of length n containing m H's is

$$\frac{n(n-1)\ldots(n-m+1)}{m!} = \frac{n!}{m!(n-m)!}.$$

(The number $n!/m!(n-m)!$, often written $\binom{n}{m}$, is the familiar binomial coefficient, that is, the coefficient of $x^m y^{n-m}$ in the expansion of (x + y)"). Since the number of sequences with exactly m H's is $\binom{n}{m}$ and the total number of sequences is 2^n, we have by Laplace's definition that the probability P (m, n) of obtaining m heads in n tosses of a fair coin is

$$P(m,n) = \frac{n!}{m!(n-m)!}\frac{1}{2^n}.$$

Consider now a coin that is "loaded" so that the probability of a head in a single toss is 1/6 (and the probability of a tail in a single toss is 5/6). Suppose again we toss this coin n times and ask for the probability of obtaining exactly m heads. To describe the equiprobable outcomes in this case, one can resort to the artifice of thinking of the coin as a six-faced die with an H on one face and T's on all the others. Using this artifice to do the counting, one finds that the probability of m heads in n tosses of the loaded coin is

$$P(m,n) = \frac{n!}{m!(n-m)!}\left(\frac{1}{6}\right)^m\left(\frac{5}{6}\right)^{n-m}.$$

Suppose further that the coin is loaded to make the probability of H irrational ($\sqrt{2}/2$, for example). In such a case one is forced into considering a many-faced die and passing to an appropriate limit as the number of faces becomes infinitely large. Despite this awkwardness the general result is quite simple: If the probability of a head in one toss is p, $0 \le p \le 1$, and the probability of a tail is $1 - p \equiv q$, then the probability of m heads in n tosses is

$$P(m,n) = \frac{n!}{m!(n-m)!}p^m q^{n-m}.$$

Building on earlier work of de Moivre, Laplace went further to consider what happens as the number of tosses gets larger and larger. Their result, that the number of heads tossed obeys the so-called standard normal distribution of probabilities, was a major triumph of early probability theory. (The standard normal distribution function,

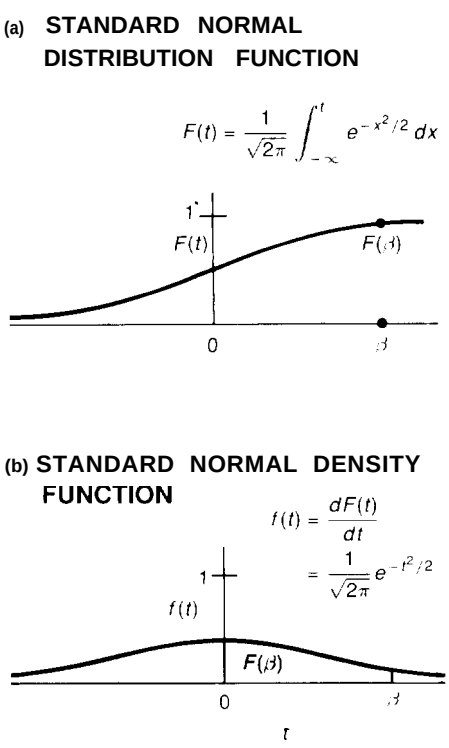


Fig. 1. Almost two centuries ago Laplace showed that the number N_H of heads obtained in a large number n of tosses of a coin (fair or loaded) follows the standard normal distribution of probabilities. More precisely, he showed that the probability of N_H being equal to or less than $np + t\sqrt{np(1-p)}$ (where p is the probability of a head in a single toss and t is some number) can be approximated, for large n, by the standard normal distribution function $F(t)$ shown in (a). The derivative of a distribution function (when it exists) is called a frequency, or density, function. Shown in (b) is the density function $f(t)$ for the standard normal distribution function. Note that the value of the distribution fuction at some particular value of t, say β, is equal to the area under the density function from $-\infty$ to β.

BERTRAND'S PARADOX

What is the probability P that a randomly chosen chord of a circle is longer than the side of the equilateral triangle inscribed within the circle?

This question cannot be answered by using Laplace's definition of probability, since the set of all possible chords is infinite, as is the set of desired chords (those longer than the side of the inscribed equilateral triangle). However, the question might be approached in the two ways depicted here and described in the text. Although both approaches seem reasonable, each leads to a different answer!

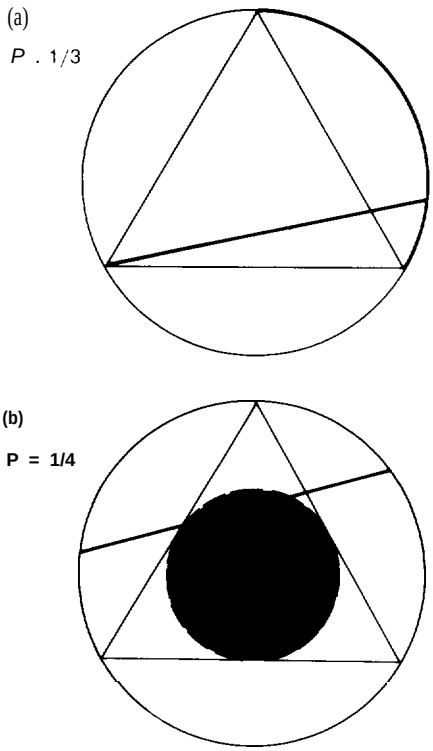


Fig. 2.

call it $F(t)$, is given by

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx;$$

the function $dF/dt = \left(1/\sqrt{2\pi}\right) e^{-t^2/2}$ is called the standard normal density function.)

The de Moivre-Laplace result can be stated as follows. As n gets larger and larger, the probability that N_H, the number of heads tossed, will be less than or equal to $np + t\sqrt{pqn}$ (where t is some number) is approximated better and better by the standard normal distribution function. Symbolically,

$$\lim_{n\to\infty} P(N_H \leq np + t\sqrt{npq}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx.$$

In other words, $P(N_H \leq np + t\sqrt{npq})$ is approximated by the area under the standard normal density function from $-\infty$ to t, as shown in Fig. 1. (In modern terminology N_H is called a random variable; this term and the terms distribution function and density function will be defined in general later.)

The de Moivre-Laplace theorem was originally thought to be just a special property of binomial coefficients. However, many chance phenomena were found empirically to follow the normal distribution function, and it thus assumed an aura of universality, at least in the realm of independent trials and events. The extent to which the normal distribution is universal was determined during the 1920s and 1930s by Lindeberg, Feller, and others after the measure-theoretic foundations of probability had been laid. Today the de Moivre-Laplace theorem (which applies to independent trials, each governed by the same probabilities) and its extension to Poisson schemes (in which each independent trial is governed by different probabilities) are regarded simply as special cases of the very general central limit theorem. Nevertheless they were the seeds from which most of modern probability theory grew.

Bertrand's Paradox

The awkwardness and logical inadequacy of Laplace's definition of probability made mathematicians suspicious of the whole subject. To make matters worse, attempts to extend Laplace's definition to situations in which the number of possible outcomes is infinite resulted in seemingly even greater difficulties. That was dramatized by Bertrand, who considered the problem of finding the probability that a chord of a circle chosen "at random" be longer than the side of an equilateral triangle inscribed in the circle.

If we fix one end of the chord at a vertex of the equilateral triangle (Fig. 2a), we can think of the circumference of the circle as being the set Q of all possible outcomes and the arc between the other two vertices as the set A of "favorable outcomes" (that is, those resulting in chords longer than the side of the triangle). It thus seems proper to take 1/3, the ratio of the length of the arc to the length of the circumference, as the desired probability.

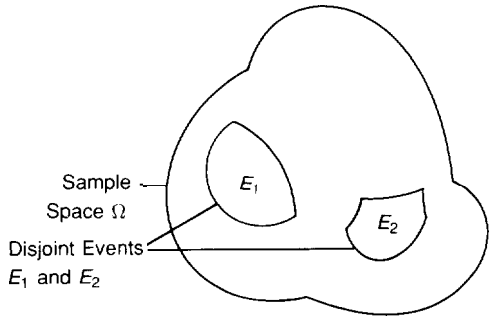
On the other hand we can think of the chord as determined by its midpoint and thus consider the interior of the circle as being the set Q of all possible outcomes. The set A of favorable outcomes is now the shaded circle in Fig. 2b, whose radius is one-half that of the original. It now seems equally proper to take 1/4 for our probability,

the ratio of the area of the smaller circle to that of the original circle.

That two seemingly appropriate ways of solving the problem led to different answers was so striking that the example became known as "Bertrand's paradox." It is not, of course, a logical paradox but simply a warning against uncritical use of the expression "at random." One must specify exactly how something is to be done at random.

Coming as it did on top of other ambiguities and uncertainties, Bertrand's paradox greatly strengthened the negative attitude toward anything having to do with chance and probability. As a result, probability theory all but disappeared as a mathematical discipline until its spectacular successes in physics (in statistical mechanics, for example) revived interest in it early in the twentieth century. In retrospect, the logical difficulties of Laplace's theory proved to be minor, but clarification of the foundations of probability theory had a distinctly beneficial effect on the subject.

AXIOM OF ADDITIVITY



Probability of (E_1 or E_2) =
Probability of E_1 + Probability of E_2

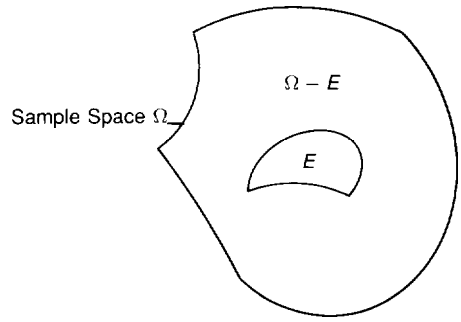
Axioms of Modern Probability Theory

The contemporary approach to probability is quite simple. From the set Ω of all possible outcomes (called the sample space), a collection of subsets (called elementary events) is chosen whose probabilities *are assumed to be given once and for all.* One then tries to calculate the probabilities of more complicated events by the use of two axioms.

Axiom of additivity: If E_1 and E_2 are events, then "E_1 or E_2" is an event. Moreover, if E_1 and E_2 are disjoint events, (that is, the subsets corresponding to E_1 and E_2 have no elements in common), then the probability of the event "E_1 or E_2" is the sum of the probabilities of E_1 and E_2, provided, of course, that E_1 and E_2 can be assigned probabilities. Symbolically,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) \text{ provided } E_1 \cap E_2 = \emptyset.$$

AXIOM OF COMPLEMENTARITY



Probability of (not E) = Probability of ($\Omega - E$) =
1 − Probability of E

Axiom of complementarity: If an event E can be assigned a probability, then the event "not E" also can be assigned a probability. Moreover, since the whole sample space Ω is assigned a probability of 1,

$$P(\text{not } E) = P(\Omega - E) = 1 - P(E).$$

Why these axioms? What is usually required of axioms is that they should codify intuitive assumptions and that they be directly verifiable in a variety of simple situations. The axioms above clearly hold in all situations to which Laplace's definition is unambiguously applicable; they are also in accord with almost every intuition one has about probabilities, except possibly those involved in quantum mechanics (Feynman 1951).

As we will see in the section on measure theory, the axioms of additivity and complementarily have an impressive mathematical content. Nevertheless they are too general and all-embracing to stand alone as a foundation for a theory so rich and fruitful as probability theory. An additional axiom of "countable additivity" is required. That axiom is the basis for the limiting theorems presented below and their application

through approximating forms. Finally, at the heart of the subject is the selection of elementary events and the decision on what probabilities to assign them. Here nonmathematical considerations come into play, and we must rely upon the empirical world to guide us toward promising areas of exploration. These considerations also lead to a central idea in modern probability theory—independence.

The Definition of Independence

Let us return to the experiment of tossing a coin n times. in attempting to construct any realistic and useful theory of coin tossing, we must first consider two entirely different questions: (1) What kind of coin is being tossed? (2) What is the tossing mechanism? The simplest assumptions are that the coin is fair and the tosses are "independent." Since the notion of independence is central to probability theory, we must discuss it in some detail.

Events E and F are independent in the ordinary sense of the word if the occurrence of one has no influence on the occurrence of the other. Technically, the two events (or, for that matter, any finite number of events) are said to be independent if the rule *of multiplication of probabilities is* applicable; that is, if the probability of the joint occurrence of E and F is equal to the product of their individual probabilities,

$$P(E \cap F) = P(E)\, P(F).$$

Kac and Ulam justified this definition of independence as follows:

"In other words, whenever E and F are independent, there should be a *rule* that would make it possible to calculate Prob. $\{E$ and $F\}$ provided only that one knows Prob. $\{E\}$ and Prob. $\{F\}$. Moreover, this rule should be *universal*; it should be applicable to every pair of independent events.

Such a rule takes on the form of a function $f(x, y)$ of two variables x, y, and we can summarize by saying that whenever E and F are independent we have

$$\text{Prob. } \{E \text{ and } F\} = f(\text{Prob. } \{E\}, \text{Prob. } \{F\})$$

Let us now consider the following experiment. Imagine a coin that can be 'loaded' in any way we wish (*i.e.*, we can make the probability p of H any number between 0 and 1) and a four-faced die that can be 'loaded' to suit our purposes also. The faces of the die will be marked 1,2,3,4 and their respective probabilities will be denoted p_1, p_2, p_3, p_4; each p_i is nonnegative and $p_1 + p_2 + p_3 + p_4 = 1$. We must now assume that whatever independence means, it should be possible to toss the coin and the die independently. If this is done and we consider (*e.g.*) the event 'H and (1 or 2)' then on the one hand

$$\text{Prob. } \{\text{H and (1 or 2)}\} = f(p, p_1 + p_2)$$

while on the other hand, since the event 'H and (1 or 2)' is equivalent to the event '(H and 1) or (H and 2),' we also have

$$\text{Prob. } \{\text{H and(1 or 2)}\} = \text{Prob. } \{\text{H and1}\} + \text{Prob. } \{\text{H and 2}\} = f(p, p_1) + f(p, p_2)$$

Note that we have used the axiom of additivity repeatedly. Thus

$$f(p, p_1 + p_2) = f(p, p_1) + f(p, p_2)$$

for all p, p_1, p_2 restricted only by the inequalities

$$0 \le p \le 1, \quad 0 \le p_1, \quad 0 \le p_2, \quad p_1 + p_2 \le 1$$

If one assumes, as seems proper, that f depends continuously on its variables, it follows that $f(x, y) = xy$ and hence the probability of a joint occurrence of independent events should be the *product* of the individual probabilities.

This discussion (which we owe to H. Steinhaus) is an excellent illustration of the kind of informal (one might say 'behind the scenes') argument that precedes a formal definition. The argument is of the sort that says in effect: 'We do not really know what independence is, but whatever it is, if it is to make sense, it must have the following properties ...' Having drawn from these properties appropriate consequences (*e.g.*, that $f(x, y) = xy$ in the above discussion), a mathematician is ready to tighten things logically and to propose a *formal definition*."

Having now defined independence as the applicability of the rule of multiplication of probabilities, let us again derive the probability of obtaining m heads in n tosses of a coin loaded so that p is the probability of a head in a single toss and $q = 1 - p$ is the probability of a tail. If the tosses are assumed to be independent, the probability of obtaining a *specified* sequence of m heads (and $(n - m)$ tails) is $p^m q^{n-m}$ (by the rule of multiplication of probabilities). Since there are $\binom{n}{m}$ such sequences, the probability of the event that exactly m out of n *independent* tosses will be heads is

$$P(n, m) = \binom{n}{m} p^m q^{n-m}.$$

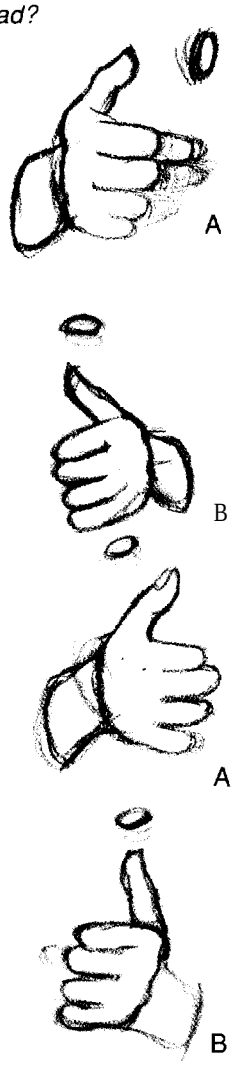
(Here we have applied the axiom of additivity). We have arrived at this formula, first developed almost two centuries ago, by using the modern concept of independence rather than Laplace's concept of equiprobability.

Probability and Measure Theory

As soon as we consider problems involving an infinite (rather than a finite) number of outcomes, we can no longer rely on counting to determine probabilities. We need instead the concept of measure. Indeed, probabilities are measures; that is, they are numerical values assigned to sets in some collection of sets, namely to sets in the sample space of all possible outcomes. The realization, during the early part of this century, that probability theory could be cast in the mold of measure theory made probability theory respectable by supplying a rigorous framework. It also extended the scope of probability theory to new, more complex problems.

Before presenting the general properties of a measure, let us consider two problems involving an infinite number of outcomes. One is the problem that led to Bertrand's paradox, namely, find the probability that a chord of a circle chosen at random is longer than the side of an inscribed equilateral triangle. For that problem the event A, or subset A, of chords that are longer and the sample space Ω of all chords could be depicted geometrically. Thus the relative sizes (measures) of the two sets could be compared even though each was an uncountable set. (The measures of those sets were either lengths or areas.) Another situation in which an infinity of outcomes needs to be considered is the following. Suppose two persons A and B are alternately tossing a coin and that A gets the first toss. What is the probability that A will be the first

What is the probability that A will be the first to toss a head?



to toss a head? This can happen either on the first toss, or on the third (the first two being tails), or on the fifth (the first four being tails), and so on. The event that A will toss the first head is thus decomposed into an infinite number of disjoint events. If the coin is fair and the tosses independent (so that the rule of multiplication applies), then the probabilities of these events are

$$\frac{1}{2}, \frac{1}{2^3}, \frac{1}{2^5}, \cdots,$$

and the probability that A will toss the first head is simply the sum of a geometric series:

$$\frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \cdots = \frac{2}{3}.$$

This result hinges on one very crucial proviso: that we can extend the axiom of additivity to an infinite number of disjoint events. This proviso is the third axiom of modern probability theory.

Axiom of countable additivity: If $E_1, E_2, E_3, \ldots$ is an infinite sequence of disjoint events, then $\bigcup_{i=1}^{\infty} E_i$ is an event and

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

Note that in solving the last problem we not only needed the axiom of countable additivity but also assumed that the probabilities used for finite sequences of trials are well defined on events in the space of infinite sequences of trials. Whether such probabilities could be defined that satisfy the axioms of additivity, complementarity, and countable additivity was one of the central problems of early twentieth-century mathematics. That problem is really the problem of defining a measure because, as we will see below, the axioms of probability are essentially identical with the required properties of a measure.

Measure Theory. The most familiar examples of measures are areas in a plane or volumes in three-dimensional Euclidean space. These measures were first developed by the Greeks and greatly extended by the calculus of Newton and Leibnitz. As mathematics continued to develop, a need arose to assign measures to sets less "tame" than smooth curves, areas, and volumes. Studies of convergence and divergence of Fourier series focused attention on the "sizes" of various sets. For example, given a trigonometric series $\sum a_n \cos nt + b_n \sin nt$, can one assign a measure to the set of *t's* for which the series converges? (Cantor's set theory, which ultimately became the cornerstone of all of modern mathematics, originated in his interest in trigonometric series and their sets of convergence.) For another example, how does one assign a measure to an uncountable set, such as Cantor's middle-third set? (See "Cantor's Middle-Third Set".) Answers to such questions led to the development of measure theory.

The concept of measure can be formulated quite simply, One wants to be able to

assign to a set A a *nonnegative* number $\mu(A)$, which will be called the measure of A, with the following properties.

Property 1: If $A_1, A_2, \ldots$ are disjoint sets that are *measurable*, that is, if each A_i can be assigned a measure $\mu(A_i)$, then their *union* $A_1 \cup A_2 \cup \ldots$ (that is, the set consisting of the elements of $A_1, A_2, \ldots$) is also measurable. Moreover,

$$\mu(A_1 \cup A_2 \cup \ldots) = \mu(A_1) + \mu(A_2) + \cdots .$$

Property 2: If A and B are measurable and A is contained in B ($A \subset B$), then $B - A$ (the set composed of elements that are in B but *not* in A) is also measurable. By property 1 then, $\mu(B - A) = \mu(B) - \mu(A)$.

Two additional properties are assumed for measures on sets in a Euclidean space.

Property 3: A certain set E, the unit set, is assumed to have measure 1: $\mu(E) = 1$.

Property 4: If two measurable sets are congruent (that is, a rigid motion maps one onto the other), their measures are equal.

When dealing with sets of points on a line, in a plane, or in space, one chooses E to be an interval, a square, and a cube, respectively. These choices are dictated by a desire to have the measures assigned to tame sets agree with those assigned to them previously in geometry or calculus.

Can one significantly enlarge the class of sets to which measures can be assigned in accordance with the above properties? The answer is a resounding yes, provided (and it is a crucial proviso) that in property 1 we allow *infinitely many A's.* When we do, the class of measurable sets includes all (well, almost all—perhaps there may be some exceptions . .) the sets considered in both classical and modern mathematics.

Although the concept of countable additivity had been used previously by Poincare, the explicit introduction and development of countably additive measures early in this century by Emile Borel and Henri Lebesgue originated a most vigorous and fruitful line of inquiry in mathematics. The Lebesgue measure is defined on sets that are closed under countably infinite unions, intersections, and complementations. (Such a collection of sets is called a c-r-f ield.) Lebesgue's measure satisfies all four properties listed above. Lebesgue's measure on the real line is equivalent to our ordinary notion of length.

But how general is the Lebesgue measure? Can one assign it to every set on the line? Vitali first showed that even the Lebesgue measure has its limitations, that there are sets on the line for which it cannot be defined. The construction of such nonmeasurable sets involves the use of the celebrated axiom of choice. Given a collection of disjoint sets, one can choose a single element from each and combine the selected elements to form a new set. This innocent-sounding axiom has many consequences that may seem strange or paradoxical. Indeed, in the landmark paper on measurable cardinals mentioned at the beginning of this article, Ulam showed (with the aid of the axiom of choice) that if a nontrivial measure satisfying properties 1 through 3 can be defined on all subsets of the real line, then the cardinality of the real numbers is larger than anyone

CANTOR'S MIDDLE-THIRD SET

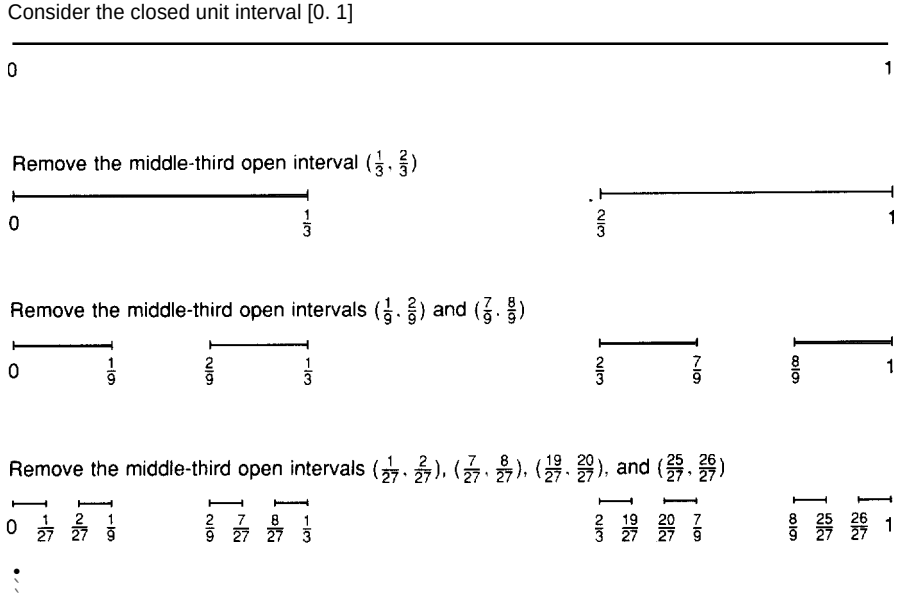
During the last quarter of the nineteenth century, Georg Cantor introduced a series of concepts that now form the cornerstone of all modem mathematics—set theory. Those concepts arose from Cantor's attempt to depict the sets of convergence or divergence of, say, trigonometric series. Many such sets have pathological properties that are illustrated by his famous construction, the "middle-third" set. This set is described by the following recursion. Consider the closed unit interval [0, 1]. First remove the middle-third open interval, obtaining two intervals $[0, 1/3]$ and $[2/3, 1]$. Next remove from each of these intervals its middle-third interval. We now have four closed subintervals each of length 1/9. Continue the process. After n steps we will have 2^n closed subintervals of [0,1] each of length $1/3^n$. From each of these we will remove the middle-third interval of length $1/3^{n+1}$. Continue the process indefinitely. Cantor's middle-third set, K, consists of all numbers in [0,1] that are never removed.

This set possesses a myriad of wonderful properties. For example, K is uncountable and yet has Lebesgue measure zero. To see that K has measure zero, consider the set $\{[0, 1] - \mathrm{K}\}$, which consists of the open intervals that were removed at some stage. At the nth stage 2^{n-1} open intervals of length $1/3^n$ were removed from the remainder. So, by the countable additivity of measure,

$$\mu([0, 1] - \mathrm{K}) = 1/3 + 2/3^2 + \cdots + 2^{n-1}/3^n + \ldots = (1/3)\left(1 + 2/3 + (2/3)^2 + \ldots\right) = 1.$$

Now, from the axiom of complementarity, $\mu(\mathrm{K}) = 0$, which is what we wanted to prove.

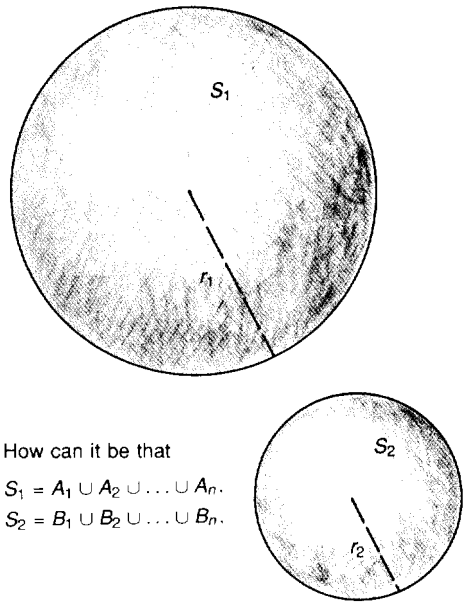
The construction of a nonzero measure on Cantor's middle-third set is discussed in the section of this article entitled Problem 2. Geometry, Invariant Measures, and Dynamical Systems. ■



imagined. (See "Learning from Ulam: Measurable Cardinals, Ergodicity, and Biomathematics.") Another example is the Banach-Tarski paradox.

Banach and Tarski proved that each of two solid spheres S_1 and S_2 of *different* radii can be decomposed into the same finite number of sets, say $S_1 = A_1 \cup A_2 \cup \ldots \cup A_n$ and $S_2 = B_1 \cup B_2 \cup \ldots \cup B_n$, such that all the A_i's and all the B_i's are among themselves pairwise disjoint and yet A_i is congruent to B_i for all i. It is therefore impossible to define meaures for these sets, since their union in one fashion yields a certain sphere and their union in a different fashion yields a sphere of different size! That such a construction is possible rests on the complicated structure, earlier pointed out by Hausdorff, of the group of rigid motions of three-dimensional Euclidean space.

BANACH-TARSKI PARADOX



We close this section on measure theory with a few comments from Kac and Ulam.

"Attempts to generalize the notion of measure were made from necessity. ... For example, one could formulate theorems that were valid for all real numbers *except* for those belonging to a specific set. One wanted to state in a rigorously defined way that the set of these exceptional points is in some sense small or negligible. One could 'neglect' merely countable sets as small in the noncountable continuum of all points but in most cases the exceptional sets turned out to be noncountable, though still of Lebesgue *measure* 0. In the theory of probability one has many statements that are valid 'with probability one' (or 'almost surely'). This simply means that they hold for 'almost all' points of an appropriate set; *i.e.*, for all points except for a set of measure 0. In statistical mechanics one has important theorems that assert properties of dynamic systems that are valid only for *almost all* initial conditions.

One final remark:

The notion or concept of measure is surely close to the most primitive intuition. The axiom of choice, that simply permits one to consider a new set Z obtained by putting together an element from each set of a family of disjoint sets, sounds so obvious as to be nearly trivial. And yet it leads to the Banach-Tarski paradox!

One can see why a critical examination of the logical foundation of set theory was absolutely necessary and why the question of existence of mathematical constructs became a serious problem.

If to exist is to be merely free from contradiction as Poincaré decreed, we have no choice but to learn to live with unpleasant things like nonmeasurable sets or Banach-Tarksi decompositions."

Consistency Theorems for Probability Measures

Now let us return to probability theory and consider the construction of countably additive probability measures. To see that a finitely additive measure cannot always be extended to a countably additive measure, consider the set Ω of integers and take as elementary events the subsets A of Ω such that either the set A is finite or the set $\Omega - A$ is finite. Set

$$\mu(A) = \begin{cases} 0 & \text{if } A \text{ is finite} \\ 1 & \text{if } \Omega - A \text{ is finite.} \end{cases}$$

So, $\mu(\Omega) = 1$ and μ satisfies the axioms of finite additivity and complementarily.

MAPPING ELEMENTARY EVENTS ONTO THE UNIT INTERVAL

Let each elementary event be one of the sets of all infinite binary sequences with the first two digits fixed Then there are four elementary events,

0

$$E_1 = \left\{ \begin{matrix} .000000 \\ \vdots \\ .001111 \end{matrix} \right\} \text{ maps to } [0.\ \tfrac{1}{4})$$

$\frac{1}{4}$

$$E_2 = \left\{ \begin{matrix} .010000 \\ \vdots \\ .011111 \end{matrix} \right\} \text{ maps to } [\tfrac{1}{4},\ \tfrac{1}{2})$$

$\frac{1}{2}$

$$E_3 = \left\{ \begin{matrix} .100000\ldots. \\ \vdots \\ .101111\ldots \end{matrix} \right\} \text{ maps to } [\tfrac{1}{2},\ \tfrac{3}{4})$$

$\frac{3}{4}$

$$E_4 = \left\{ \begin{matrix} .110000\ldots. \\ \vdots \\ .111111\ldots \end{matrix} \right\} \text{ maps to } [\tfrac{3}{4},\ 1]$$

$$\sum_{i=1}^{4} P(E_i) = \text{Length of the unit interval}$$

However, if μ were countably additive, then we would have the contradiction

$$1 = \mu(\Omega) = \mu\left(\bigcup_{n=1}^{\infty}\{n\}\right) = \sum_{n=1}^{\infty}\mu(\{n\}) = 0.$$

Now consider the problem of defining a countably additive probability measure on the sample space Ω of all infinite two-letter sequences (each of which represents the outcome of an infinite number of independent tosses of a fair coin). Take as an elementary event a set E consisting of all sequences whose first m letters are specified (m = 1,2,.. .). Since there are 2"' such elementary events, we use the axiom of finite additivity to assign a probability P of 1 /2'" to each such event. Can this function F', which has been defined on the elementary events, be extended to a countably additive measure defined on the o-field generated by the elementary events? Ulam and Lomnicki proved such an extension exists for any infinite sequence of independent trials. Kolmogorov obtained the ultimate consistency results by giving necessary and sufficient conditions under which an extension can be made from a finitely additive to a countably additive measure, including the case of non-independent trials. These extensions put the famous limiting laws of probability theory, such as the laws of large numbers, on solid ground.

In the case of coin tossing we have chosen our elementary events to be sets of infinite sequences whose first m digits are fixed and have assigned them each a probability of $1/2^m$ in agreement with the finitely additive measure. Now we will show that the measure defined by these choices is equivalent to Lebesgue's measure on the unit interval [0,1] and is therefore a well-defined countably additive measure. First associate the digit 1 with a head and the digit O with a tail and encode each outcome of an infinite number of tosses as an infinite sequence of 1's and 0's (101 10. .., for example), which in turn can be looked upon as the binary representation of a real number t $(0 \leq t \leq 1)$. In this way we establish a correspondence between real numbers in [0,1] and infinite two-letter sequences; the correspondence can be made one-to-one by agreeing once and for all on which of the two infinite binary expansions to take when the choice presents itself. (For instance, we must decide between .01000 . . . and .00111 . . . as the binary representation of 1/4.)

The use of the binary system is dictated not only by considerations of simplicity. As one can easily check, the crucial feature is that each elementary event maps into an interval whose *length is* equal to the corresponding *probability* of the event. In fact, fixing the first m letters of a sequence corresponds to fixing the first m binary digits of a number, and the set of real numbers whose first m binary digits are fixed covers the interval between $\ell/2^m$ and $(\ell+1)/2^m$, where ℓ is $0, 1, 2, \ldots,$ or $2^m - 1$, depending on how the first m digits are fixed. Clearly the length of such an interval, $1/2^m$, is equal to the probability of the corresponding elementary event. Thus the probability measure in the sample space Ω of all infinite two-letter sequences maps into the ordinary Lebesgue measure on the interval [0,1] and is therefore equivalent to it.

The space of all infinite sequences of O's and 1's is *infinite-dimensional* in the sense that it takes infinitely many "coordinates" to describe each "point" of the space. What we did was to construct a certain countably additive measure in the space that was "natural" from the point of view of independent tosses of a fair coin

THE INSTITUTE FOR ADVANCED STUDY
SCHOOL OF MATHEMATICS
FINE HALL
PRINCETON, NEW JERSEY

Oct. 3., 1936.

Dear Ulam,
many thanks for your kind letter – both Mariette and I are delighted to see You soon in Princeton again. Oct. 10. – or any other date thereabouts which suits You – will be very convenient to us – and we expect, that You will stay with us, and stay as long as possible. –

I agree wholeheartedly with Your plans to write an up-to-date presentation of measure-theory. Carathéodorys exposition, which is perhaps the relatively best one existing, is hopelessly obsolete. A thoroughly modern one, as much combinatorial and as little topological as possible making extensive use of finite and infinite direct products, and – above all – interpreting measure much more as probability and much less as volume, would really be a very good thing. At least I often felt how badly such a thing is lacking in the present literature. What would be the style of Your treatise, and its length? I will be very glad if You can let me see any part of your mscr. In the lectures (and 34/35) I gave here on "linear operations" in 1933/34, I tried to deal with measure somewhat in the above spirit, but I was badly handicapped by the fact, that measure was not my primary topic there.

I looking forward, too, with great interest for your mscr. on the general product-operation.

I am expecting to discuss several mathematical questions, when You come here, those You mentioned, and a few others. By then I will have unearthed my two last year's mscr's, too, which You mentioned — we are unpacking now, so the excavations do not proceed very quickly.

Expecting to see You soon again, and with the very best greetings from Mariette, too, I am yours as ever

John von Neumann

Dear Ulam,

Many thanks for your kind letter—both Mariette and I are delighted to see you soon in Princeton again. Oct. 10-or any other date thereabouts which suits you—will be very convenient to us—and we expect that you *will* stay with us, and stay as long as possible.—

1 agree wholeheartedly with your plans to write an up-to-date presentation of **measure-theory. Caratheodorys** exposition, which is perhaps the *relatively* best one existing, is hopelessly obsolete. A thoroughly modem one as much combinatorial and as little topological] as possible, making extensive use of finite and **infinite direct products, and—above all—** interpreting measure much more as probability and much less as volume, would really be a very good thing. At least I often felt how badly such a thing is lacking in the present literature. What would be the style of your treatise, and its length? I will be very glad if you can let me see any part of your mscr. In the lectures I gave here on "linear operations" in 1933/34 and 34/35, I tried to deal with measure somewhat in the above spirit, but I was badly handicapped by the fact that measure was not my primary topic there.

I [am] looking forward, too, with great interest for your mscr. on the general product operation.

I am expecting to discuss several mathematical questions, when you come here, those you mentioned, and a few others. By then I will have unearthed my two last year's mscr's, too, which you mentioned—we are unpacking now, so the excavations do not proceed very quickly.

Expecting to see you soon again, and with the very best greetings from Mariette, too, I am yours as ever *John von Neumann*

Note: With the he[p of J. D, Bernstein, Ulam started a book on measure theory while he was at Wisconsin, That collaboration was interrupted by Stan's war years at Los Alamos and was never resumed. The idea of presenting measure theory from the combinatorial probabilistic perspective is now a common practice. A good example is P. Billingsley's Probability and Measure.

This approach immediately suggests extensions to more general infinite-dimensional spaces in which each coordinate, instead of just being O or 1, can be an element of a more general set and need not even be a number. Such extensions, called product measures, were introduced by Lomnicki and Ulam in 1934. (Stan's idea of writing a book on measure theory emphasizing the probabilistic interpretation of measure is the subject of the accompanying letter from von Neumann to Ulam.) Measures for sets of curves have also been developed. The best known and most interesting of these was introduced by Norbert Wiener in the early 1920s and motivated by the theory of Brownian motion. Mathematicians have since found new and unexpected applications of the Wiener measure in seemingly unrelated parts of mathematics. For example, it turns out that the Wiener measure of the set of curves emanating from a point p in space and hitting a three-dimensional region R *is* equal to the electrostatic potential at p generated by a charge distribution that makes the boundary of the "conductor" R an equipotential surface on which the potential is equal to unity. Since the calculation of such a potential can be reduced by classical methods to solving a specific differential equation, we establish in this way a significant link between classical analysis and measure theory.

Random Variables and Distribution Functions

Having introduced the measure-theoretic foundations of probability, we now turn to a convenient formalism for analyzing problems in probability. In many problems the possible outcomes can be described by numerical quantities called random variables. For example, let X be the random variable describing the outcome of a single toss of a fair coin. Thus, set X equal to 1 if the toss yields a head and to 0 if the toss yields a tail. This is an example of an elementary random variable; that is, X is a function with a constant value on some elementary event and another constant value on the complementary event. In general a random variable is a real-valued function defined on the sample space Ω that can be constructed from elementary random variables by forming algebraic combinations and taking limits. For example, N_H, the number of heads obtained in n tosses of a coin, is a random variable defined on the sample space consisting of all sequences of T's and H's of length n; its value is equal to $\sum_{i=1}^{n} X_i$, where $X_i = 1$ if the ith toss is a head and $X_i = 0$ otherwise.

In evaluating the outcomes of a set of measurements subject to random fluctuations, we are often interested in the mean, or expected, value of the random variable being measured. The expected value $E(X)$ (or m) is defined as

$$E(X) \equiv \int_{\Omega} X(\omega)\, dP(\omega),$$

where $X(\omega)$ is the value of X at a point ω in the sample space and $P(\omega)$ is the probability measure defined on the sample space. In the case of a fair coin, $P(X = 1) = 1/2$ and $P(X = 0) = 1/2$, so the expected value of X is a simple sum:

$$E(X) = \sum x_i P_i = (0 \times \frac{1}{2}) + (1 \times \frac{1}{2}) = \frac{1}{2}.$$

The expected value of a random variable X is most easily determined by knowing

its distribution function F. This function, which contains all the information we need to know about a random variable, is defined as follows:

$$F(t) \equiv P(X \leq t),$$

where the set $X \leq t$ is the set of all points ω in Ω such that $X(\omega) \leq t$. The form of this function is particularly convenient. It allows us to rewrite $E(X)$, which is a Lebesgue integral over an abstract space, as a familiar classical integral over the real line:

$$E(X) \equiv \int_{\Omega} X(\omega)\, dP(\omega) = \int_{-\infty}^{\infty} t\, dF(t).$$

Furthermore, if X has a density function $f(t) \equiv dF(t)/dt$, then

$$E(X) = \int_{-\infty}^{\infty} t f(t)\, dt.$$

The expected value is one of the two commonly occurring averages in probability and statistics; the other is the variance of X, denoted by $\sigma^2(X)$ or $\mathrm{var}(X)$. The variance is defined as the expected value of the square of the deviation of X from its mean:

$$\sigma^2(X) = \mathrm{var}(X) \equiv E\left((X - E(X))^2\right) = E(X^2) - \left(E(X)\right)^2.$$

The standard, or root-mean-square, deviation of X is defined as $\sigma(X) = \sqrt{\mathrm{var}(X)}$.

Figures 3 and 4 illustrate two distribution functions, the binomial distribution function for the number of heads obtained in five tosses of a fair coin and a normal distribution function with a positive mean.

The Laws of Large Numbers

A historically important problem in probability theory and statistics asks for estimates on how a random variable deviates from its mean, or expected, value. A simple rough estimate is, of course, its root-mean-square deviation. An estimate of a different nature was obtained by the nineteenth-century mathematician Chebyshev. This estimate, known as Chebyshev's inequality, gives an upper limit on the probability that a random variable Y deviates from its mean $E(Y)$ by an amount equal to or greater than a $(a > 0)$:

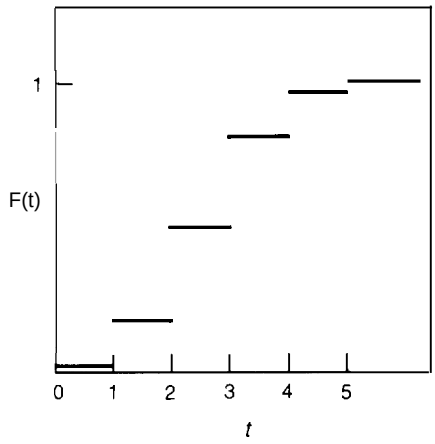
Chebyshev's inequality: $P(|Y - E(Y)| \geq a) \leq \mathrm{var}(Y)/a^2$.

This fundamental inequality will lead us to the famous laws of large numbers, which tell us about average values for infinite sequences of random variables. We begin by returning again to the coin loaded in such a way that p is the probability of a head in a single toss. If this coin is tossed a large number of times n, shouldn't the frequency of heads, N_H/n, be approximately equal to p, at least in some sense?

This question can be answered on several levels. Let X_i be the random variable describing the outcome of the ith toss. Set $X_i = 1$ if the ith toss is a head and $X_i = 0$ if

BINOMIAL DISTRIBUTION FUNCTION

Fig. 3. The distribution function $F(t)$ for the number of heads obtained in n independent tosses of a fair coin is a binomial distribution, so called because the probability of obtaining k heads in n tosses of the coin is given by a formula involving binomial coefficients, namely $\binom{n}{k}\frac{1}{2^n}$. Shown here is the binomial distribution function for the number of heads obtained in five tosses of the coin. The value of $F(t)$ equals the probability that the number of heads is equal to or less than t.

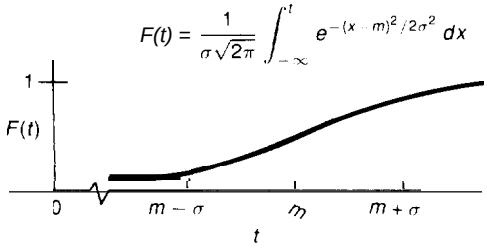


$$F(t) = \sum_{0}^{k} \binom{n}{k} \frac{1}{2^n} \text{ if } k \leq t \leq k+1,$$

$$k = 0, 1, 2, 3, 4$$

$$= 1 \text{ if } t \geq 5$$

(a) **NORMAL DISTRIBUTION FUNCTION**



(b) **NORMAL DENSITY FUNCTION**

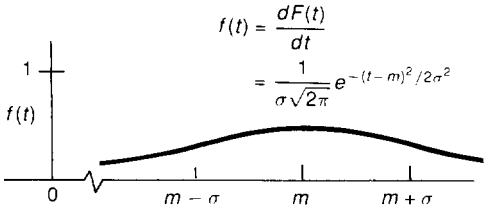


Fig. 4. So many random variables can be described, at least approximately, by the distribution function shown in (a) that it is known as the normal distribution function. Examples of such random variables include the number of heads obtained in very many tosses of a coin and, as a general experimental fact, accidental errors of observation. The value of $F(t)$ equals the probability that the value of the random variable is equal to or less than $(t - m)/\sigma$, where m is the mean, or expected, value of the random variable and σ is its standard deviation. (The mean here is assumed to be positive.) Shown in (b) is the normal density function $f(t) \equiv dF(t)/dt$, which gives the probability that the value of the random variable is $(t - m)/\sigma$.

the ith toss is a tail. Then $N_H = X_1 + \cdots + X_n$. Also, the distribution function for each X_i is the same, namely,

$$F(t) = \begin{cases} 0 & t < 0 \\ 1 - p & 0 \le t < 1 \\ 1 & 1 \le t. \end{cases}$$

(Random variables that have the same distribution function are said to be identically distributed.) Now the expected value of N_H/n is easy to compute:

$$E\left(N_H/n\right) = E\left((X_1 + \ldots + X_n)/n\right) = (1/n)\sum_{i=1}^{n} E(X_i) = (1/n)(np) = p.$$

Thus, on the simplest level our guess is right: The frequency of heads, N_H/n, is approximately equal to p in the sense that the expected value of N_H/n is p. But surely, even in a very long series of tosses, it would be foolish to expect N_H/n to exactly equal p (and N_T/n to exactly equal $1 - p$). What one is looking for is a statement that holds only *in the limit as the number of tosses becomes infinite.*

Bernoulli proved such a theorem: As n gets larger and larger, the probability that N_H/n differs from its expected value p by more than a positive amount ϵ tends to 0:

$$\lim_{n\to\infty} P_n(|N_H/n - p| \ge \epsilon) = 0,$$

where P_n is the probability measure on Ω_n, the space of all sequences of H's and T's of length n. No matter what positive ϵ is chosen, the probability that the difference between the frequency of heads and p, the probability of a head in a single trial, exceeds ϵ can be made arbitrarily small by tossing the coin a sufficiently large number of times.

Let us see how Bernoulli's theorem follows from Chebyshev's inequality. First, notice that $\mathrm{var}(X_i) = p(1 - p)$ for all i. Second, the random variables $X_1, \ldots, X_n$ are independent (the outcome of the ith toss has no influence on the outcome of jth). Now, from the fact that $E(XY) = E(X)E(Y)$ for independent random variables, we get

$$\mathrm{var}(X_1 + \cdot\ \cdot + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_1) = np(1 - p).$$

So, by Chebyshev's inequality

$$P_n(|N_H/n - p| \ge \epsilon) \le np(1 - p)/n^2\epsilon^2 = p(1 - p)/\epsilon^2 n.$$

Thus, for each $\epsilon > 0$

$$\lim_{n\to\infty} P_n(|N_H/n - p| \ge \epsilon) = 0.$$

Notice that the measure-theoretic background of Bernoulli's theorem is trivial (at least as far as coin-tossing is concerned), since the events of interest correspond to finite sets. That is, for each n we need only estimate how many trials of length n there are such that the number of heads differs from np by more than ϵn. Nevertheless, the simple argument just given can be generalized to prove the famous weak law of large numbers.

Weak law of large numbers: Let $X_1, X_2, X_3, \ldots$ be independent, identically distributed random variables such that $\text{var}(X_1) < \infty$. Then for each $\epsilon > 0$

$$\lim_{n\to\infty} P\left(|(X_1 + \ldots + X_n)/n - E(X_1)| \geq \epsilon\right) = 0.$$

In other words, for any positive ϵ the probability that the deviation between the frequency in n trials and the expected value in a single trial exceeds ϵ can be made arbitrarily small by considering a sufficiently large number of trials.

For our coin-tossing example N_H/n approximately equals p in another sense also. Suppose one asks for the probability that the frequency of heads (in the limit as the number of tosses becomes infinite) is actually equal to p. The answer was obtained by Borel in 1909:

$$P\left(\lim_{n\to\infty} N_H/n = p\right) = 1.$$

Notice the complexity of the question. In order to deal with it, the sample space Ω is now by necessity the set of all *infinite* two-letter sequences ω and the subset of interest is the set A of those sequences for which

$$\lim_{n\to\infty} \frac{N_n(\omega)}{n} = p,$$

where $N_n(\omega)$ is the number of H's among the first n letters of the infinite sequence ω. It takes some work just to show that A is an event in the sample space Ω. Unlike the question that led to the weak law of large numbers, this question required the full apparatus of modern probability theory. An extension of Borel's result by Kolmogorov is known as the strong law of large numbers.

Strong law of large numbers: Let $X_1, X_2, X_3, \ldots$ be independent, identically distributed random variables such that $E(|X_1|) < \infty$. Then

$$P\left(\lim_{n\to\infty} (X_1 + \cdots + X_n)/n = E(X_1)\right) = 1.$$

An Application of the Strong Law of Large Numbers. Let us illustrate the power of the strong law of large numbers by using it to answer the following question: What is the probability that, in an infinite sequence of tosses of a fair coin, two heads occur in succession?

We will first answer this question using only the rules governing the probabilities of independent events. In particular, we will use the axioms of countable additivity and complementarity and the rule of multiplication of probabilities. Let A_k be the event that a head occurs on the $(2k-1)$th toss and on the $(2k)$th toss. Each A_k is an elementary event, and $P(A_k) = 1/4$. Now, by the axiom of countable additivity, $\bigcup_{k=1}^{\infty} A_k$ is an event; in particular, it is the event that, for *some* k, heads occur on the $(2k-1)$th and $2k$th tosses. By the axiom of complementarity,

$$P\left(\Omega - \bigcup_{k=1}^{\infty} A_k\right) = P\left(\bigcap_{k=1}^{\infty} (\Omega - A_k)\right) = \lim_{n\to\infty} P\left(\bigcap_{k=1}^{n} (\Omega - A_k)\right).$$

Since the events $A_1, A_2, A_3, \ldots$ are independent, the events $\Omega - A_1, \Omega - A_2, \Omega - A_3, \ldots$ are also independent, and we can apply the rule of multiplication of probabilities:

$$P\left(\Omega - \bigcup_{k=1}^{\infty} A_k\right) = \lim_{n\to\infty} \prod_{k=1}^{n} P(\Omega - A_k) = \lim_{n\to\infty} (3/4)^n = 0.$$

Finally, by the axiom of complementarity, $P\left(\bigcup A_k\right) = 1$; that is, there exists, with probability 1, some k such that the $(2k-1)$th and $(2k)$th tosses are heads.

Now we will answer the same question by using the strong law of large numbers. Let X_i be the random variable such that

$$X_i(\omega) = \begin{cases} 1 & \text{if } \omega \in A_i \\ 0 & \text{if } \omega \notin A_i. \end{cases}$$

Then $X_1, X_2, X_3, \ldots$ is a sequence of independent random variables. Also, they all have the same distribution: $(P(X_i = 1) = 1/4, P(X_i = 0) = 3/4$, and $E(X_i) = 1/4$. Therefore, according to the strong law of large numbers,

$$\lim_{n\to\infty} (X_1 + \ldots + X_n)/n = 1/4 \text{ with probability 1.}$$

This result is stronger than that obtained above. It guarantees, with probability 1, the existence of infinitely many k's such that heads occur on the $(2k-1)$th and $(2k)$th tosses; further, the set of all such k's has an arithmetic density of 1/4.

Borel's theorem marked the beginning of the development of modern probability theory, and Kolmogorov's extension to the strong law of large numbers greatly expanded its applicability. To quote Kac and Ulam:

"Like all great discoveries in mathematics the strong law of large numbers has been greatly generalized and extended; in the process it gave rise to new problems, and it stimulated the search for new methods. It was the first serious venture outside the circle of problems inherited from Laplace, a venture made possible only by developments in measure theory. These in turn were made possible only because of polarization of mathematical thinking along the lines of set theory."

The polarization Kac and Ulam were referring to concerns the great debate at the turn of the century about whether the infinite in mathematics should be based upon Cantor's set theory and its concomitant logical difficulties. The logical problems have been met, and today we use Cantor's theory with ease.

The Monte Carlo Method. One of Stan Ulam's great ideas, which was first developed and implemented by von Neumann and Metropolis, was the famous Monte Carlo method. It can be illustrated with Chebyshev's inequality. Suppose that we need to quickly get a rough estimate of $\int_1^\infty (\sin x)/x^3\, dx$. Setting $t = 1/x$, the problem then is to estimate $\int_0^1 t \sin(1/t)\, dt$. Let $y_1, \ldots, y_n$ be independent random variables each uniformly distributed on [0,1]. That is, for all i, $P(a < y_i < b) = b - a$, where (a, b) is a subinterval of [0,1]. Now set $f(t) = t \sin(1/t)$ and for each i let $X_i = f(y_i)$. Then

$X_1, \ldots, X_n$ is a sequence of independent identically distributed random variables. Also,

$$|E(X_i)| = \left| \int_0^1 t \sin(1/t)\, dt \right| < \int_0^1 |t \sin(1/t)|\, dt < 1,$$

and

$$\mathrm{var}(X_1) = \int_0^1 t^2 \sin^2(1/t)\, dt - \left(E(X_i)\right)^2 < \int_0^1 t^2 \sin^2(1/t)\, dt < \int_0^1 t^2 dt = 1/3$$

By Chebyshev's inequality we have

$$P\left(\left| (1/n) \sum_{i=1}^{n} X_i - \int_0^1 t \sin(1/t)\, dt \right| \geq a \right) \leq \mathrm{var}\left((1/n) \sum_{i=1}^{n} X_i \right) / a^2 \leq \mathrm{var}(X_1)/na^2.$$

Thus if n is large, $(1/n)\sum_{i=1}^{n} X_i$ is, with high probability, a good estimate of the value of the integral. For example, if $a = 0.005$ and $n = 134,000$, then

$$P\left(\left| (1/n) \sum_{i=1}^{n} X_i - \int_0^1 t \sin(1/t)\, dt \right| \geq 0.005 \right) < 0.1.$$

In other words, if we chose 134,000 numbers $y_1, \ldots, y_{134,000}$ independently and at random from [0,1], then we are 90 percent certain that $(1/134,000)\sum_{i=1}^{134,000} y_i \sin(1/y_i)$ differs from the integral by no more than 0.005. So, if we can statistically sample the unit interval with numbers $y_1, \ldots, y_n$, then

$$(1/n) \sum_{i=1}^{n} y_i \sin(1/y_i) \approx \int_1^{\infty} \frac{\sin t}{t^3}\, dt.$$

(The reader may well wonder why such a large number of sample points is required to be only 90 percent certain of the value of the integral to within only two decimal places. The answer lies in the use of Chebyshev's inequality. By using instead the stronger central limit theorem, which will be introduced below, many fewer sample points are needed to yield a similar estimate.)

The Monte Carlo method is a wonderful idea and, of course, tailor-made for computers. Although it might be regarded simply as an aspect of the more ancient statistical sampling technique, it had many exciting new aspects. Three of these are (1) a scope of application that includes large-scale processes, such as neutron chain reactions; (2) the capability of being completely implemented on a digital computer; and (3) the idea of generating random numbers and random variables. How do we mechanically produce numbers $y_1, \ldots, y_n$ in [0,1] such that the y_i's are independent and identically distributed? The answer is we don't. Instead, so-called pseudo-random numbers are generated. Many fascinating problems surfaced with the advent of Monte Carlo. Dealing with them is one of the major accomplishments of the group of intellects gathered at Los Alamos in the forties and the fifties. (See "The Beginning of the Monte Carlo Method.")

Central Limit Theorem

We close this tutorial by returning to the de Moivre-Laplace theorem and interpreting it in the modern context. Let X_i be a random variable describing the outcome of the ith toss of a coin; set $X_i = 1$ if the ith toss is a head and $X_i = 0$ otherwise. Let S_n be the number of heads obtained in n tosses; that is $S_n = X_1 + \cdots + X_n$. Then the de Moivre-Laplace theorem can be stated as follows:

$$\lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx.$$

Now $np = nE(X_1) = E(S_n)$ and $\sqrt{np(1-p)} = \sqrt{n}\sigma(X_1) = \sigma(S_n)$. So if we "renormalize" S_n by setting $Y_n = \big(S_n - E(S_n)\big)/\sigma(S_n)$, each Y_n has a mean of 0 and a standard deviation of 1. Then the equation above tells us that the distribution function of Y_n tends to the standard normal distribution. The central limit theorem is a generalization of this result to any sequence of identically distributed random variables. We state the central limit theorem formally.

Central limit theorem: Let $X_1, X_2, X_3, \cdots$ be a sequence of independent, identically distributed random variables with $E(X_1) = m$ and $\text{var}(X_i) = \sigma^2 < \infty$. Set $S_n = X_1 + \cdots + X_n$. Then

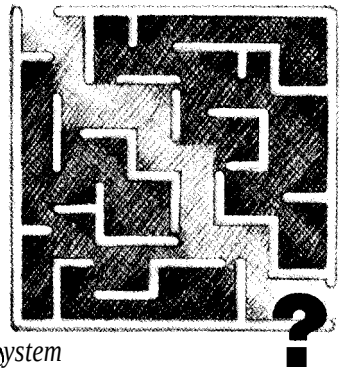
$$\lim_{n\to\infty} P\left(\frac{S_n - nm}{\sqrt{n}\sigma} \leq t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx.$$

Thus the normal distribution is the universal behavior in the domain of independent trials under renormalization. Its appearance in so many areas of science has led to many debates as to whether it is a "law of nature" or a mathematical theorem.

Thanks to the developments in modern probability theory, we begin our investigations with many powerful tools at our disposal. Those tools were forged during a period of tremendous upheavals and turmoil, a time when very careful analysis carried the day. At the heart of that analysis lay the concept of countable additivity. Stan Ulam played a seminal role in developing these tools and presenting them to us.

PROBABILISTIC APPROACHES to NONLINEAR PROBLEMS

Part 111
PROBABILITY and NONLINEAR SYSTEMS



Problem 1. *Energy Redistribution: An Exact Solution to a Nonlinear, Many-Particle System*

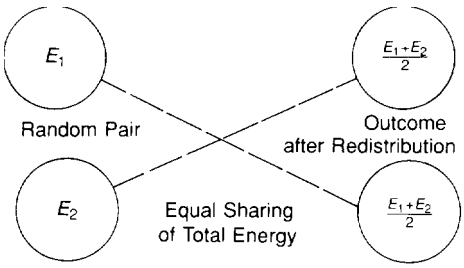
Ulam's talent for seeing new approaches to familiar problems is evident in one he posed concerning the distribution of energy in physical systems. Will the energy distribution of an isolated system of N interacting particles always evolve to some limiting energy distribution? And, if so, what is the distribution? (Note that this question differs from the one asked in statistical mechanics. There one assumes that at equilibrium the system will have the most probable distribution. One then derives that the most probable distribution is the Boltzmann distribution, the density of which is $e^{-E/kT}$.)

Obviously, following the evolution of a system of N interacting particles in space and time is a very complex task. It was Stan's idea to simplify the situation by neglecting the spatial setting and redistributing the energy in an abstract random manner. What insights can one gain from such a simplification? One can hope for new perspectives on the original problem as well as on the standard results of statistical mechanics. Also, even if the simplification is unrealistic, one can hope to develop some techniques of analysis that can be applied to more realistic models. In this case David Blackwell and I were able to give an exact analysis of an abstract, highly nonlinear system by using a combination of the machinery of probability theory and higher order recursions (Blackwell and Mauldin 1985). We hope that the technique will be useful in other contexts.

Let us state the problem more clearly and define what we mean by redistributing energy in an "abstract random manner." Assume we have a vast number of indistinguishable particles with some initial distribution of energy, and that the average energy per particle is normalized to unity. Further, let us assume the particles interact only in pairs as follows: At each step in the evolution of the system, pair all the particles at random and let the total energy of each pair be redistributed between the members of the pair according to some fixed "law of redistribution" that is independent of the pairs. Iterate this procedure. Does the system have a limiting energy distribution and, if so, how does it depend on the redistribution law?

The Simplest Redistribution Law. To begin we will consider the simplest redistribution law: each particle in a random pair gets one-half the total energy of the pair, If the number of particles in the system is finite, it is intuitively clear that under iteration the total energy of the system will tend to become evenly distributed—all the particles

SIMPLEST LAW FOR ENERGY REDISTRIBUTION



LIMITING ENERGY DISTRIBUTION

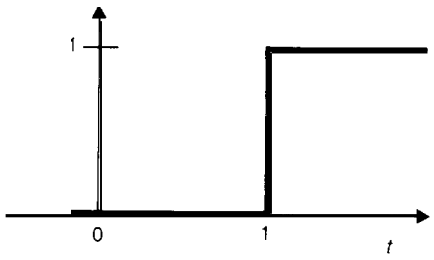


Fig. 5. Consider a system of *N* particles with some arbitrary initial distribution of energy. Assume that the initial mean energy is 1 and that the particles interact in pairs. Assume further that the total energy of an interacting pair is redistributed so that each member of the pair acquires one-half the total energy of the pair. Then with probability 1 the system reaches a limiting energy distribution described by a step function with a step height of 1 at $t = 1$. That is, the probability that the energy per particle is less than t equals 0 for $t < 1$ and equals 1 (the initial mean energy) for $t \geq 1$.

will tend to have the same energy. So, a system with only finitely many particles has a limiting distribution of energy, namely, a step function with a jump of size 1 at $t = 1$, and moreover, no matter what the initial distribution of energy is, the system tends to this distribution under iteration.

Even for a system with a continuum of particles, our observations for the finite case still hold. In order to see this, we formalize the problem in terms of probability theory.

Let X be a random variable corresponding to the initial energy of the particles. Thus, the distribution function F_1 associated with X is the initial distribution of energy: $F_1(t) = P(X < t)$ is the proportion of particles with energy less than t. Our arguments and analysis will be based only on the knowledge of the energy distribution function and how it is transformed under iteration by the redistribution law. In terms of distribution functions, our normalization condition, that the average energy per particle is unity, means that the expected value of X, $\int_0^\infty t\, dF_1(t)$, equals 1.

We seek a random variable $T(X)$ corresponding to the energy per particle after applying the redistribution law once. To say that the indistinguishable particles are paired *at random* in the redistribution process means that, given one particle in the pair, we know nothing about the energy of the second except that its distribution function should be the initial distribution function F_1. In other words, we can describe the energy of the randomly paired particles by two *independent* random variables X_1 and X_2, each having the same distribution as X. Thus the simplest redistribution law, according to which paired particles share the total energy of the pair equally, can be expressed in terms of $T(X)$, X_1, and X_2 as

$$T(X) = \frac{X_1 + X_2}{2}.$$

The new distribution of energy, call it F_2, that describes the random variable $T(X)$ will be a convolution of the distributions of $X_1/2$ and $X_2/2$. Since X_1 and X_2 both have the distribution F_1, the distribution F_2 of $T(X)$ is given by

$$F_2(t) = P\left(\frac{X_1 + X_2}{2} \leq t\right) = P(X_1 + X_2 \leq 2t) = \int_{-\infty}^{t} F_1(2t - x)\, dF_1(x).$$

To carry out the second iteration, we repeat the process. The energy $T^2(X) = T\big(T(X)\big)$ will have the same distribution as $(Y_1+Y_2)/2$, where Y_1 and Y_2 are *independent* and each is distributed as $T(X)$. In other words, if we let X_1, X_2, X_3, and X_4 be independent and distributed as X_1, then Y_1 is distributed as $(X_1 + X_2)/2$, and Y_2 is distributed as $(X_3 + X_4)/2$. The energy is distributed as $T^2(X) = (X_1 + X_2 + X_3 + X_4)/4$.

After n iterations the energy per particle will have the same distribution as $T^n(X) = (X_1 + \cdots + X_{2^n})/2^n$, where the X_i's are independent and distributed as X. This expression for $T^n(X)$ is exactly the expression that appears in the strong law of large numbers (see page 71). Therefore the strong law tells us that the limiting energy of each particle ω as $n \to \infty$ is

$$\lim_{t\to\infty} T^n\big(X(\omega)\big) = \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_{2^n}(\omega)}{2^n} = E(X_1) = 1, \text{ almost surely,}$$

where $E(X_i)$ is the expected value of the initial distribution Thus, after n iterations of

this random process, the energies of almost all particles converge to unity. In terms of distribution functions, we say that in the space of all "potential" actualizations of this iterative random process, almost surely, or with probability 1, the limiting distribution of energy will be a step function with a jump of size 1 at $t = 1$ (Fig. 5).

Notice that for this simplest redistribution law (1) the redistribution operator T is a simple linear operator and (2) even so, the strong law of large numbers is needed to determine the limiting behavior.

More Complicated Redistribution Laws. Stan proposed more interesting laws of redistribution. The redistribution operator T for each of these laws is nonlinear, and different techniques are needed to analyze the system. For example, after pairing the particles, choose a number α between 0 and 1 at random. Then instead of giving each particle one-half the total energy of the pair, let us give one particle α times the total energy of the pair and give the other particle $(1 - \alpha)$ times the total energy. The energy $T(X)$ will then have the same distribution as $U(X_1 + X_2)$, where U is uniformly distributed on [0,1] (that is, all values between 0 and 1 are equally probable) and U, X_1, and X_2 are independent. What happens to this system under iteration is a much more complicated matter. For one thing, unlike the redistribution operator in the simplest case, the operator T is now highly *nonlinear* and the law of large numbers is not available as a tool. A new approach is required. To get an idea of what to expect, Stan first used the computer as an experimental tool. From these studies he correctly guessed the limiting behavior (Ulam 1980): no matter what the initial distribution of energy is, we have convergence to the *exponential* distribution (Fig. 6).

Let me indicate how Blackwell and I proved this conjecture. We used a classical *method of moments* together with an analysis of a *quadratic recursion*. For now let us assume that a stable limiting distribution exists and let X have this distribution. Then $T(X) = U(X_1 + X_2)$ has the same distribution. So, calculating m_n, the nth moment of X (that is, the expected value of X^n), we have

$$m_n = E(X^n) = E\left(T(X)^n\right) = E\left((U(X_1 + X_2))^n\right) = E\left(U^n(X_1 + X_2)^n\right).$$

By independence and the binomial theorem

$$m_n = E(U^n)E\left((X_1 + X_2)^n\right) = \frac{1}{n+1}E\left(\sum_{p=0}^{n}\binom{n}{p}X_1^p X_2^{n-p}\right) = \frac{1}{n+1}\sum_{p=0}^{n}\binom{n}{p}E(X_1^p X_2^{n-p}).$$

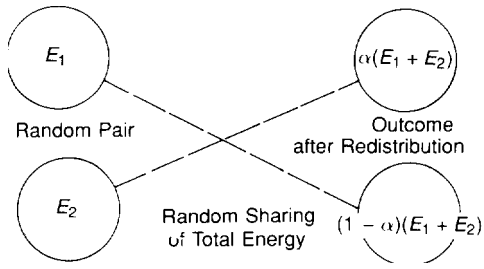
Since X_1 and X_2 are independent, the expected value of each product is equal to the product of the expected values, $E(X_1^p X_2^{n-p}) = E(X_1^p)E(X_2^{n-p})$. Substituting this into the equation above and using the definition of moments, we have

$$m_n = \frac{1}{n+1}\sum_{p=0}^{n}\binom{n}{p}m_p m_{n-p} = \frac{2}{n+1}m_0 m_n + \sum_{p=1}^{n-1}\binom{n}{p}m_p m_{n-p}.$$

Using the fact that $m_0 = 1$, we solve for m_n:

$$m_n = \frac{1}{n-1}\sum_{p=1}^{n-1}\binom{n}{p}m_p m_{n-p}.$$

RANDOM LAW FOR ENERGY REDISTRIBUTION



LIMITING ENERGY DISTRIBUTION

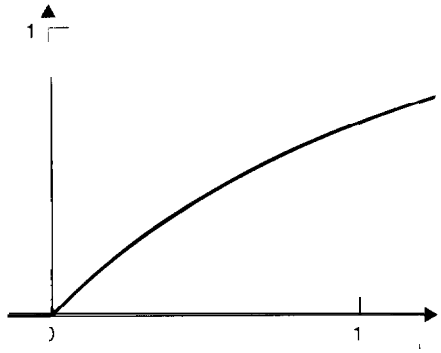


Fig. 6. Consider a system identical to the one described in Fig. 5 except that the total energy of an interacting pair is redistributed randomly between the members of the pair. In particular, assume that one particle receives a randomly chosen fraction α of the total energy and the other particle receives the remainder. The system still reaches a limiting energy distribution, one equal to 0 for $t < 0$ and equal to $1 - e^{-t}$ for $t \geq 0$.

This is a quadratic recursion formula. Substituting the initial condition $m_1 = 1$, we find that $m_2 = 2$ and $m_3 = 6$. An induction argument shows that $m_n = n!$ for all n. But $n!$ is the nth moment of the exponential distribution! Of course, our assumption is that a stable distribution and all its moments exist. It takes some work to prove that this assumption is indeed true and that no matter what initial distribution one starts with, the distribution of the iterates converges to the exponential.

It should not be too surprising that our result agrees in its general form with the Boltzmann distribution of statistical mechanics. After all, both are derived from similar assumptions. The Boltzmann distribution is derived from the assumptions that (1) energy and the number of particles are conserved, (2) all energy states are equally probable, and (3) the distribution of energy is the most probable distribution. In our problem we also assumed conservation of energy and number of particles. Moreover, taking U in our redistribution law to be the uniform distribution makes all energy states equally probable. The difference is that the iteration process selects the most probable distribution with no a priori assumption that the most probable distribution will be reached.

We can go further and replace U by any random variable with a symmetric distribution on [0,1]. The symmetric condition insures that the particles are indistinguishable. We call the distribution of U the redistribution law. Again, one obtains a quadratic recursion formula. Blackwell and I analyzed this formula and showed that for every such U the system tends toward a stable limiting distribution. In other words, there is an attractive fixed point in the space of all distributions. Moreover, there is a one-to-one correspondence between the stable limiting distribution and the redistribution law that yields it.

Momentum Redistribution. There is a corresponding momentum problem. Assume we have a vast number of indistinguishable particles (all of unit mass) with some initial distribution of momentum. Let us assume that the particles interact in pairs as follows. At each step in the evolution of the system, pair all the particles at random and let the total momentum of each pair be redistributed between the members of the pair according to some law of redistribution that is independent of the pairs. Of course, we wish to conserve energy and momentum. These conservation laws place severe constraints on the possibilities. If $\mathbf{v}_1$ and $\mathbf{v}_2$ are the initial velocity vectors of two particles in a pair and $\mathbf{v}_1'$ and $\mathbf{v}_2'$ are the velocity vectors after collision, then by momentum conservation

$$\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}_1' + \mathbf{v}_2'$$

and by energy conservation

$$\|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2 = \|\mathbf{v}_1'\|^2 + \|\mathbf{v}_2'\|^2.$$

Consider this process in the center-of-mass frame of reference. Let λ_i be the fraction of the total kinetic energy that particle i has after collision and let $\mathbf{u}_i$ be the unit vector in the direction of the velocity of particle i. Then

$$\lambda_1 + \lambda_2 = 1$$

and

$$\sqrt{\lambda_1}\mathbf{v}_1 + \sqrt{\lambda_2}\mathbf{v}_2 = 0.$$

From these equations it follows that $\lambda_1 = \lambda_2 = 1/2$ and $\mathbf{v}_2 = -\mathbf{v}_1$. What this means is that all we can do is choose in the center-of-mass frame a new direction vector for one of the two colliding particles. Everything else is then determined. The other particle goes in the opposite direction, and the total kinetic energy in the center-of-mass frame is divided evenly between the two particles. Thus, the only element of randomness is in how the new direction vector is chosen. If all directions are assumed to be equiprobable, then it can be shown that no matter what the initial distribution of velocity is, the system tends under iteration to a limiting distribution that is the standard normal distribution in three-dimensional Euclidean space $\Re^3$. We have thus rederived the Maxwell-Boltzmann distribution of velocities. Here again we can go further and consider more complicated redistribution laws.

Suppose one allows ternary collisions instead of binary collisions. Then there are more degrees of freedom, and the problem again becomes interesting mathematically. The results of our analysis show that the situation is much like the redistribution of energy in that the limiting distribution of velocity depends on the law of redistribution of velocity.

Problem 2. *Geometry, Invariant Measures, and Dynamical Systems*

The intimate relationship among geometry, measures, and dynamical systems that was elucidated in the last century continues to deepen and hold our attention today. Poincaré made several monumental contributions to this development in his treatise *Les Méthodes Nouvelles de la Mécanique Céleste*. One major issue he considered concerned the stability of motion in a gravitational field such as that of our solar system. Would small perturbations from any given set of initial orbits lead to a collision of the planets? A tremendous amount of work had been done on this dynamical system, but the governing system of differential equations remained unsolved. Faced with this situation, Poincaré made a wonderful flanking maneuver by introducing "qualitative" methods that involved measures.

For the setting consider the motion of N bodies and the corresponding phase space S, whose $6N$ coordinates code the position and momentum of each of the N bodies. The phase space is a subset of Euclidean $6N$-space and each point of S corresponds to a state of the system. Consider T, the time-one map of S. That is, if s is the initial state of the system, then $T(s)$ is the state of the system one time unit later. Now, various notions of stability can be given in terms of the properties of T. One of these is recurrence, or, as Poincaré said, "stabilité à la Poisson." A state s is said to be recurrent provided that if the system is ever in s, then it will return arbitrarily close to s infinitely often. Formally, s is recurrent provided that for every open region U about s there are infinitely many positive integers n such that $T^n(s)$ is in U. Poisson had earlier attempted to show this kind of stability for the restricted three-body problem. Poincaré used the fundamental tenet of measure theory, countable additivity, to prove that the set of all points s in the phase space for which recurrence *does not* occur is of measure zero.

Recurrence Theorem: Let $B = (s \in S \,|\, s$ is not recurrent). Then B has measure zero.

Poincare's proof of this theorem (see "The Essence of Poincare's Proof of the Re-

POINCARÉ'S PROOF *of the* RECURRENCE THEOREM

Let us indicate the essential ingredients of the argument that B, the set of points in phase space that are not recurrent, has measure zero. Assume S is a surface of constant energy and the volume (measure) of S is finite, $v(S) < \infty$. Let $U_1, U_2, U_3, \ldots$ be an infinite sequence of open balls in S such that each point of S lies in one of the balls (no matter how small the radii of the balls may be). For each n let B_n be the set of points in U_n that are *not* recurrent; that is, B_n consists of all points $s \in U_n$ such that $T^p(s) \in U_n$ for only finitely many positive integers p. Now consider the set $B = \bigcup_{n=1}^{\infty} B_n$, that is, the set of *all* points that are not recurrent. Since the measure v is *assumed to be countably additive,* we have $v(B) \leq \sum_{n=1}^{\infty} v(B_n)$. Poincaré also assumed that the notion of volume could be extended to sets B_n that are more complicated than open regions.

Given these assumptions we can prove that B has measure zero if we show that $v(B_n) = 0$ for each n.

The argument goes like this. Fix n and let $U = U_n$. Let

$$C = U - \bigcup_{p=1}^{\infty} T^{-p}(U).$$

It is easy to show that, for each $k, s \in T^{-k}(C)$ if and only if $T^k(s) \in U$ and $T^i(s) \notin U$ for all $i > k$. Consequently,

$$T^{-i}(C) \cap T^{-j}(C) = \emptyset \text{ if } 0 \leq i < j,$$

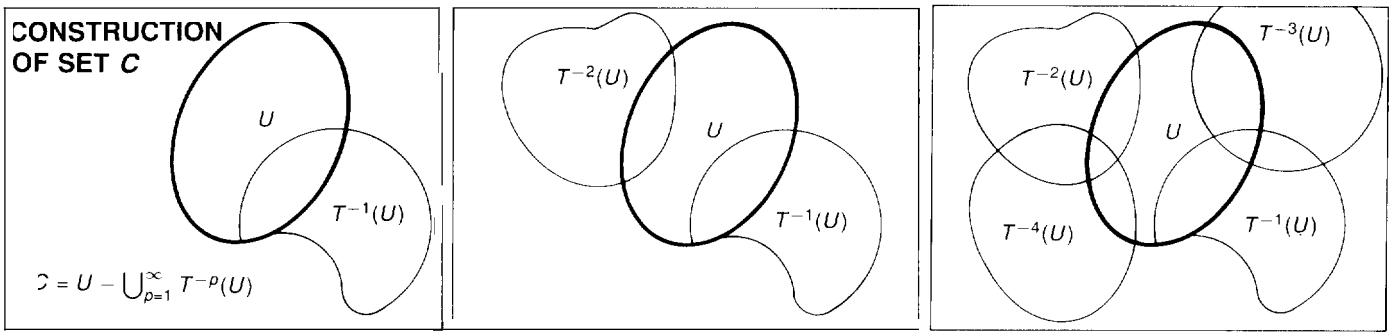
and

$$B_n = \bigcup_{k=0}^{\infty} T^{-k}(C).$$

Therefore, since the sets $T^{-k}(C)$ are pairwise disjoint and the measure is countably additive,

$$v(B_n) = \sum_{k=0}^{\infty} v\left(T^{-k}(C)\right).$$

Since T is volume-preserving, the sets $T^{-k}(C)$ all have the same measure, a. If $a > 0$, then $v(S)$ would be infinite, which is a contradiction. Thus each B_n has measure zero, and therefore B also has measure zero. ■



currence Theorem") is a shining jewel that made clear to the mathematical world the importance of countable additivity in the development of measure.

But what measure did Poincare have in mind here? Af'ter all, there is an entire grab bag of measures on the subsets of S. In the case of the N -body problem, since the system is a Hamiltonian system, the geometry of the phase space clearly indicates the correct measure. Let us see why. Liouville had proved the seminal result that if the map T that describes the time evolution of the system is a Hamiltonian, then T is volume-preserving in the phase space. That is, if U is an open set or region, then $v(U) = v\left(T(U)\right)$, where $v(E)$ is the volume of E. Poincaré carried out his analysis on a "surface of constant energy." Since the N-body problem is a conservative system, the function T leaves the total energy invariant and therefore maps each such surface nto itself. Moreover, since T is a Hamiltonian, it is volume-preserving on this surface. Consequently, the geometric structure of the surface determines the appropriate measure o use. Since the surface is a manifold, by definition there is a positive integer m such hat each point of S lies in a region that is geometrically the same as a piece of Euclidean n-dimensional space. So, the measure to use on the manifold S is the one we naturally ssociate with Euclidean m-dimensional space, namely, m-dimensional volume.

Geometry and Dynamical Systems

To summarize, the N-body problem is a classical dynamical system in which the time-one map T is a continuous one-to-one map of the phase space X onto itself. The inverse map, T^{-1}, is also continuous. Thus, T is a *homeomorphism*. There is a natural measure on the phase space X that is invariant under T. From one point of view, this measure is the volume element corresponding to the dimension of the phase space. From another viewpoint the natural invariant measure expresses the fact that the system is a Hamiltonian system. In the phase space X a surface S of constant energy forms an invariant set, and again there is an invariant measure on S corresponding to our ordinary notion of volume. The set B of all points that are not recurrent is also an invariant set with respect to T. However, it is not at all clear that we can define some natural invariant measure on B that is both nonzero and invariant under T. Many dynamical systems being studied today "live" on invariant sets that, like B, are not manifolds. Instead they are "pathological" sets, sets that at one time were thought to be the private domain of the purest and most abstract mathematicians. The examples range from Cantor sets to nowhere-differentiable curves to indecomposable continua. Many of these pathological invariant sets are "strange attractors" of dynamical systems; the system is "attracted" in the sense that it will eventually end up on the set from any starting point. (The discovery of one of the first strange attractors is described in the section Cubic Maps and Chaos of the article "Iteration of Maps, Strange Attractors, and Number Theory—An Ulamian Potpourri.")

Properties of Invariant Sets. Let us now indicate some of the problems and techniques used in studying such sets in the context of *dynamical systems*. We will consider discrete dynamical systems, that is, systems in which the time evolution is described by discrete steps. We consider a function T that maps a space X into itself and the iterates of T, that is, $T^1, T^2, T^3, \ldots$, where $T^{n+1}(x) = T\left(T^n(x)\right)$. We are interested in an *invariant* set—a subset M of X such that $T(M) \subset M$. The simplest invariant set consists of a fixed point x such that $T(x) = x$; a more complicated invariant set is a periodic orbit, a set consisting of the points $x, T(x), \ldots, T^{n-1}(x)$, and $T^n(x) = x$. Invariant sets are further classified according to how points near the invariant set behave under T. An invariant set M is called an *attractor* if there is a region U surrounding M such that if $x \in U$, then $T^n(x)$ gets closer and closer to M as n increases. On the other hand, M is called a *repeller* if there is a region U surrounding M such that if $x \in (U - M)$, then $T^n(x)$ is not in M for n sufficiently large. For example, if X is the real number line, then 0 is an attracting fixed point for $T(x) = x/2$ and a repelling fixed point for $T^{-1}(x) = 2x$. The intrinsic properties of an invariant set are also of interest. For example, one might want to know whether there is a point x of M such that the *orbit* of x, that is, $x, T(x), T^2(x), \ldots$, is dense in M. If T is an irrational rotation of the plane, then the unit circle is invariant and the orbit of every point on the circle is dense in the circle. Another possibility is that T is *topologically mixing* on M; that is, for every region U of M there is some n such that $M \subset T^n(U)$.

One central problem we will look at in some depth is the construction of "natural" or useful invariant measures for the sets M. In particular we want a measure μ such that $\mu(X - M) = 0$ and $\mu\left(T^{-1}(B)\right) = \mu(B)$ for each measurable subset B of M. That is, the measure is zero for points outside the invariant set M and is invariant with respect to the inverse of T.

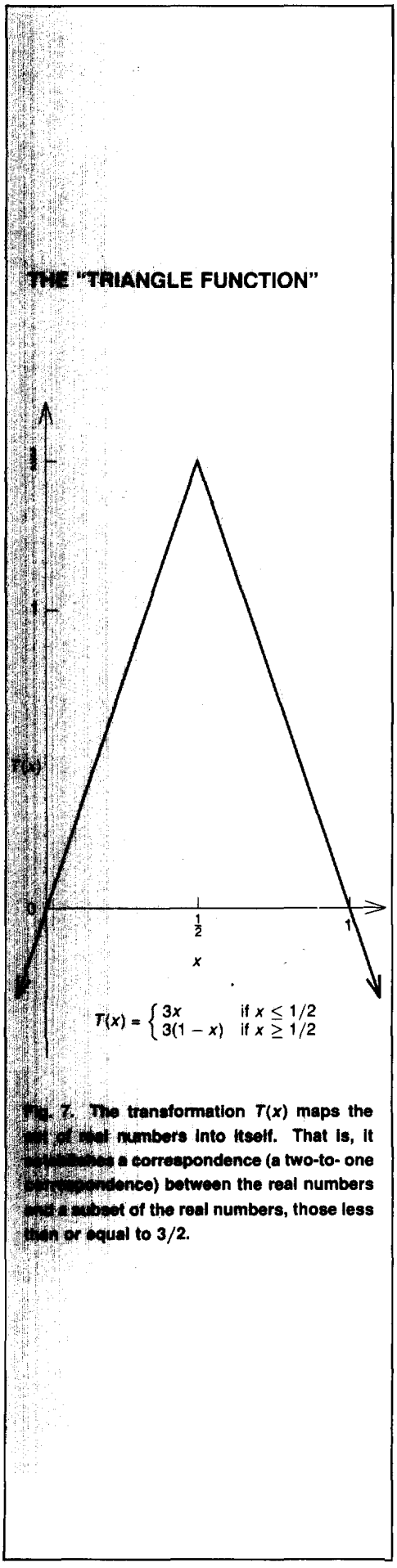


Fig. 7. The transformation $T(x)$ maps the set of real numbers into itself. That is, it establishes a correspondence (a two-to- one correspondence) between the real numbers and a subset of the real numbers, those less than or equal to 3/2.

Cantor's Set as an Invariant Set. Let us consider a simple example of a map whose invariant set is Cantor's middle-third set. Let X be the real number line and let $T(x) = (3/2)(1 - |2x - 1|)$. Then T is a two-to-one map of X into itself, the "triangle function" whose graph is shown in Fig. 7. This transformation can also be written in the following form:

$$T(x) = \begin{cases} 3x & \text{if } x \leq 1/2 \\ 3(1-x) & \text{if } x \geq 1/2. \end{cases}$$

Now consider what happens to x under the iterates of T. If $x < 0$, then $T^n(x) = 3^n x$ and $T^n(x) \to -\infty$. If $1 < x$, then $T(x) < 0$ and higher iterates are given by $3^n(1-x)$. Again, $T^n(x) \to -\infty$. Thus, the iterates of all points outside the interval [0,1] are repelled. On the other hand, $x = 0$ is a fixed point, and, since $T(1/4) = 3/4$ and $T(3/4) = 1/4$, the set $\{1/4, 3/4\}$ forms a periodic orbit of order 2. It turns out that there is a natural invariant set under the iterates of T that lies in the interval [0,1]. To find it we consider successive iterations of T and keep track of the parts of the interval [0,1] that are mapped outside the interval by each interation. The first few iterations of T are illustrated in Fig. 8 and are described below. If x is in the open interval $(1/3, 2/3)$, $T(x) > 1$, and thus T maps this open interval out of the interval [0,1]. The two intervals $J_1 = [0, 1/3]$ and $J_2 = [2/3, 1]$ are each mapped onto [0,1]. Thus, $J_1 \cup J_2$ consists of all points remaining in the interval [0,1] after one iteration. What points of J_1 remain in [0,1] after the second iteration? The middle third of J_1, namely $(1/9, 2/9)$ is mapped out of the interval [0,1] by the second iteration of T, and the two subintervals $J_{11} = [0, 1/9]$ and $J_{12} = [2/9, 1/3]$ make up the points of J_1 that remain in $[0, 1]$ after two iterations of T. Similarly, the middle third of $J_2, (7/9, 8/9)$, is mapped out of [0,1] by T^2, and the two subintervals of J_2, $J_{21} = [2/3, 7/9]$ and $J_{22} = [8/9, 1]$, make up the points of J_2 that remain in [0,1] under T^2. Continuing this analysis, we find that the points of [0,1] that remain in [0,1] after n iterations of T consist of 2^n intervals. Moreover, they are precisely the same 2^n intervals that appear in the construction of Cantor's famous middle-third set. Thus, Cantor's middle-third set, call it M, is invariant under T, and if $x \notin M$, then for some k, $T^k(x)$ is not in [0,1]. Thus, if $x \notin M, T^n(x) \to -\infty$. The Cantor set is a repellent invariant set of T, and this map is also topologically mixing on M.

Hausdorff Measure and Dimension. If we think of T as an analog of a dynamical system whose motion in phase space is restricted to a Cantor set we might like to find a natural measure on this set. Our problem is: Which one of the many possible invariant measures is useful? One clue for determining the appropriate measure for the N-body problem was the fact that the phase space is a manifold and we therefore know the *dimension* of the space. We could then use the corresponding volume in the Euclidean space of that dimension to guide us to the correct measure. But what do we do with the Cantor set of our example? What is its dimension? In the early part of this century Felix Hausdorff developed an approach for determining the dimension of a general metric space (a space with a notion of a metric, or distance, between points) in terms of measures associated with the metric. It is perhaps surprising at first that the dimension of a space may not be an integer. Such spaces have been christened fractals by Mandelbrot, and he has provided many examples of their occurrence in physical phenomena. The idea behind Hausdorff's generalization of dimension is very simple

CONSTRUCTION OF INVARIANT SET FOR $T(x)$

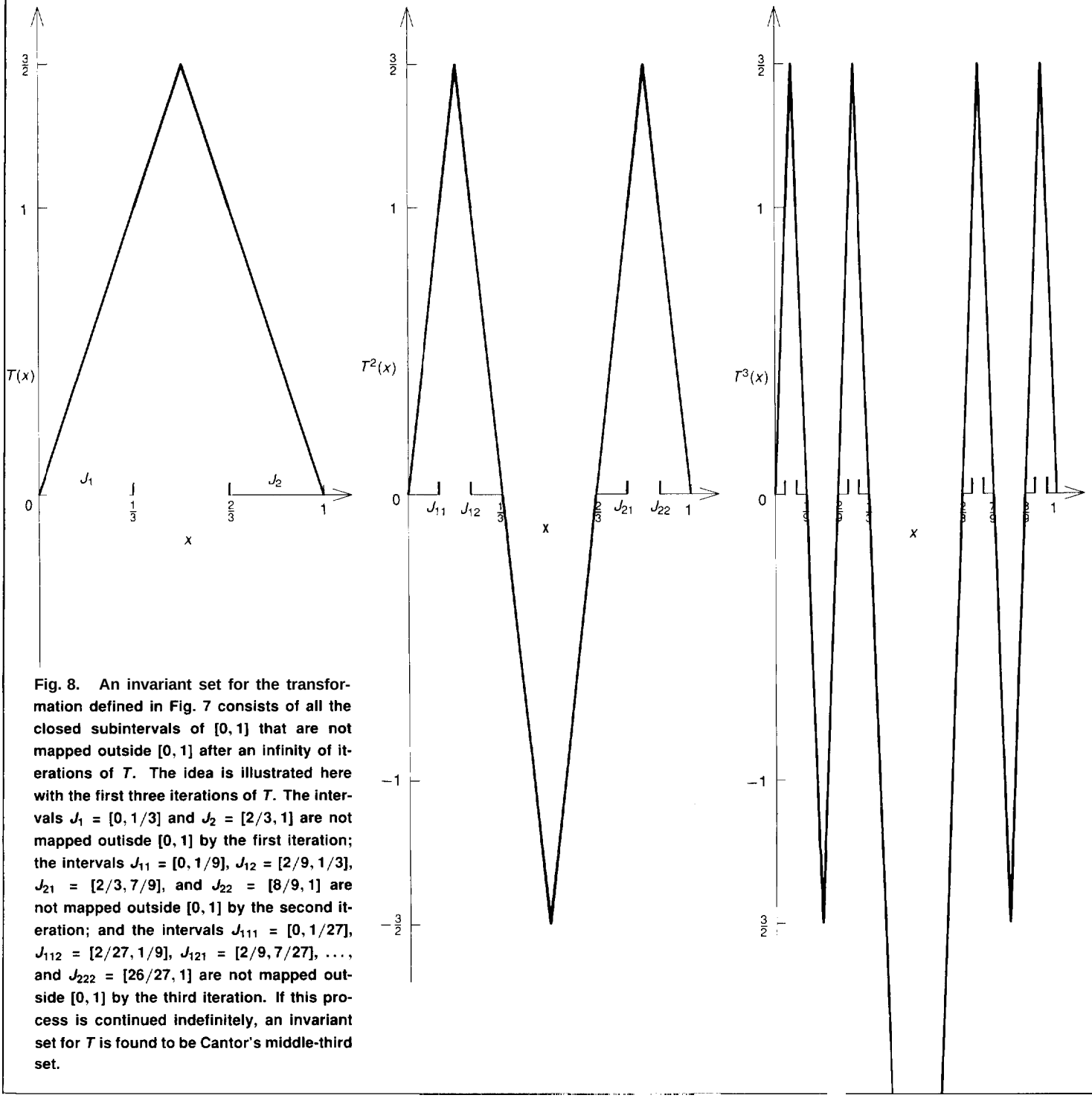


Fig. 8. An invariant set for the transformation defined in Fig. 7 consists of all the closed subintervals of [0, 1] that are not mapped outside [0, 1] after an infinity of iterations of T. The idea is illustrated here with the first three iterations of T. The intervals J_1 = [0, 1/3] and J_2 = [2/3, 1] are not mapped outisde [0, 1] by the first iteration; the intervals J_{11} = [0, 1/9], J_{12} = [2/9, 1/3], J_{21} = [2/3, 7/9], and J_{22} = [8/9, 1] are not mapped outside [0, 1] by the second iteration; and the intervals J_{111} = [0, 1/27], J_{112} = [2/27, 1/9], J_{121} = [2/9, 7/27], ..., and J_{222} = [26/27, 1] are not mapped outside [0, 1] by the third iteration. If this process is continued indefinitely, an invariant set for T is found to be Cantor's middle-third set.

and is based on the idea of *self-similarity* or *scaling*.

Let's take the simplest example, the unit square. We could say that the dimension of the unit square is 2 for the following reason. Consider any scaling transformation $f(x) = \lambda x$, where x is a point in the plane. The transformation f is called a similarity map of the plane and the image of the unit square under f will be a square whose area is λ^2. The *power* to which we raise the scaling exponent to obtain the measure of the image set is the dimension of the original set. Exactly the same reasoning shows that the unit cube in Euclidean n space has dimension n.

The generalization to more complicated metric spaces is straightforward. Consider a general metric space X. A map f is a *similarity* map of a subset E of X if the distance between points in E scale by a factor r under the action of the map. In other words there is a number r such that for all x and y in E, $\text{dist}\big(f(x),f(y)\big) = r\,\text{dist}(x,y)$. Hausdorff defined for each number $\beta \geq 0$ a measure H^β on X that obeys the scaling law of Hausdorff measures.

Scaling law of Hausdorff measures: If $E \subset X$ and f is a similarity map of E onto $f(E)$ with similarity ratio r, then $H^\beta\big(f(E)\big) = r^\beta H^\beta(E)$.

While the measures H^β are defined on the metric space for all values of $\beta > 0$, Hausdorff showed that there is one and only one measure H^α for which a "jump" occurs. He called α the dimension of the metric space.

Hausdorff dimension theorem: For each metric space X, there is a number α such that if $\beta < \alpha$, then $H^\beta(X) = \infty$ and if $\alpha < \beta$, then $H^\beta(X) = 0$. The number α is called the Hausdorff dimension of X.

How do Hausdorff's definitions of measure and dimension compare with our ordinary notions in Euclidean space? It turns out that the Hausdorff dimension of n-dimensional Euclidean space is n (which it should be, of course) and the associated Hausdorff measure H^n is the same as our usual definition of volume element. Thus, H^α is a natural generalization to a space of dimension α of our ordinary notions of measure, or volume element, in Euclidean space. Once the Hausdorff dimension α of a space is known, we have a natural measure on the space, namely H^α. So the first problem is to determine the dimension of the space under consideration.

Hausdorff Dimension of Cantor's Middle-Third Set. As an example, we will show that the self-similarity properties of the middle-third Cantor set C define its Hausdorff dimension as $\log 2/\log 3$. (In fact, Hausdorff proved this in his original paper.)

Consider the two similarity maps $f_1(x) = x/3$ and $f_2(x) = x/3 + 2/3$. Then $f_1(C) = C \cap [0, 1/3]$ and $f_2(C) = C \cap [2/3, 1]$. So $C = f_1(C) \cup f_2(C)$. Since $f_1(C)$ and $f_2(C)$ are disjoint and H^α is a measure,

$$H^\alpha(C) = H^\alpha\big(f_1(C)\big) + H^\alpha\big(f_2(C)\big).$$

By the scaling law, $H^\alpha\big(f_1(C)\big) = (1/3)^\alpha H^\alpha(C)$ and $H^\alpha\big(f_2(C)\big) = (1/3)^\alpha H^\alpha(C)$. Therefore

$$H^\alpha(C) = (1/3)^\alpha H^\alpha(C) + (1/3)^\alpha H^\alpha(C) = (2/3^\alpha) H^\alpha(C).$$

Cancelling $H^\alpha(C)$, we have

$$1 = 2/3^{\alpha}, \text{ or } \alpha = \log 2/\log 3.$$

We conclude that the Hausdorff dimension of C is $\log 2/\log 3$. Of course, this is only a heuristic argument (because we cannot cancel $H^{\alpha}(C)$ unless $H^{\alpha}(C)$ is positive and finite), but it can be justified.

Returning to our example $T(x) = (3/2)(1 - |2x - 1|)$, we have shown that the invariant set M is Cantor's middle-third set and that the Hausdorff dimension of M is $\alpha = \log 2/\log 3$. In fact $\mu = H^{\alpha}$, Hausdorff's volume element in dimension α, is an invariant measure on M.

Our analysis of this example is typical of the analyses of many discrete dynamical systems. We found an invariant set M that is constructed by an algorithm that analyzes the behavior of points near M. The first application of the algorithm yields nonoverlapping closed regions $J_1, \ldots, J_n$ the second yields nonoverlapping subregions $J_{i1}, \ldots, J_{in}$ in each J_i, and so forth. Finally, the invariant set M is realized as

$$M = \bigcap_{k=1}^{\infty} \left(\bigcup_{i_j \leq n} J_{i_1 \ldots i_k} \right).$$

In this example the construction is *self-similiar*; that is, there are scaling ratios $t_1, \ldots, t_n$ such that a region at iteration k, $J_{i_1 \ldots i_k}$ and a subregion at level $k+1$, $J_{i_1 \ldots i_{k+1}}$, are geometrically similar with ratio $t_{i_{k+1}}$. (In our example $t_1 = t_2 = 1/3$.) When such similarity ratios exist, one can use a fundamental formula due to P. A. P. Moran for calculating the Hausdorff dimension of the invariant set.

Theorem: If $M = \bigcap_{k=1}^{\infty} \left(\bigcup_{i_j \leq n} J_{i_1 \ldots i_k} \right)$, then $\dim(M) = \alpha$, where α is the solution of $t_1^{\alpha} + \cdots + t_n^{\alpha} = 1$. Moreover, $0 < H^{\alpha}(M) < +\infty$.

That is, α is the Hausdorff dimension of M, and H^{α} is a well-defined finite measure on M.

Random Cantor Sets. One of my current interests centers on analyzing the invariant sets obtained when the dynamical system experiences some sort of random perturbation. The perturbation introduces a perturbation in the algorithm used to construct the invariant set. Thus we randomize the algorithm, and the scaling ratios $t_1, t_2, \ldots, t_n$, instead of having fixed or deterministic values as before, are now random variables that have a certain probability distribution. One theorem of Williams and mine (Mauldin and Williams 1986) is that the Hausdorff dimension of the final "perturbed" set M is, with probability 1, the solution of

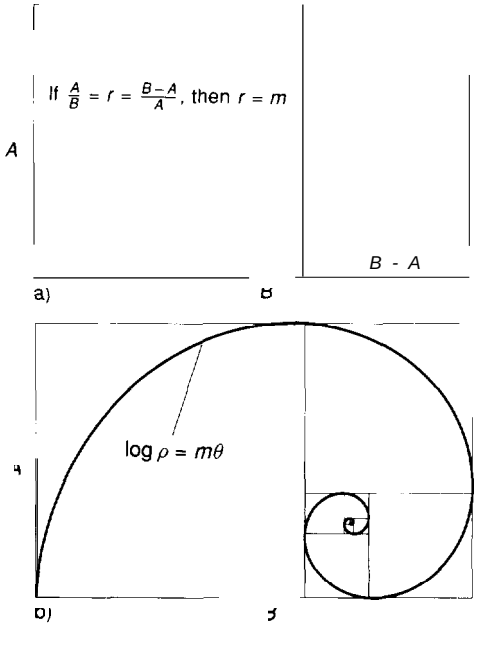
$$E\left(t_1^{\alpha} + \cdots + t_n^{\alpha}\right) = 1,$$

where $E(t_1^{\alpha} + t_2^{\alpha} + \cdots)$ is the expected value of the sum of the αth powers of the scaling ratios. Note that this formula reduces to Moran's formula in the deterministic case.

As an example suppose our randomly perturbed system produces Cantor subsets of [0,1] as follows. First, choose x at random according to the uniform distribution on [0,1]. Then between x and 1 choose y at random according to the uniform distribution on $[x, 1]$. We obtain two intervals $J_1 = [0, x]$ and $J_2 = [y, 1]$. Now in each of these intervals repeat the same procedure (independently in each interval). We obtain two

THE GOLDEN MEAN

Fig. 9. (a) Consider a rectangle with sides of length *A* and *B*, *A* < *B*. Let *r* denote the ratio of *A* to *B*. Divide this rectangle into a square of side *A* and a new rectangle. If the ratio of the lengths of the sides of the new rectangle, (*B* − *A*)/*A*, also equals *r*, then both the original rectangle and the new rectangle are golden rectangles and *r* is equal to the golden mean *m*. (The numerical value of *m*, $(\sqrt{5}-1)/2$, is obtained by solving the two simultaneous equations $r = A/B$ and $r = (B-A)/A$.) (b) The process of dividing a golden rectangle into a square and a new golden rectangle can, of course, be continued indefinitely. It can be shown that the logarithmic spiral given in polar coordinates by $\log\rho = m\theta$ passes through two opposite vertices of each successively smaller square. This fact may help explain why the Hausdorff dimension of the random Cantor sets described in the text is equal to the golden mean.



subintervals of J_1, J_{11} and J_{12}, and two subintervals of J_2, J_{21} and J_{22}. Continue this process. We will obtain a random Cantor set, and its Hausdorff dimension α is, with probability 1, the solution of $E\left(t_1^\alpha + t_2^\alpha\right) = 1$, or

$$\int_0^1 \left(x^\alpha + \frac{1}{1-x}\int_x^1 (1-y)^\alpha dy \right) dx = 1.$$

A little calculus shows that

$$\alpha = \frac{\sqrt{5}-1}{2}, \text{ the golden mean!}$$

A problem left for the reader: Why should the golden mean (Fig. 9) arise as the dimension of these randomly constructed Cantor sets?

Problem 3. *Computer Experiments and Random Homomorphisms*

One topic Stan and I discussed several times was whether one could "randomize" dynamical systems in some way. Is it possible to define a probability measure on a wide class of dynamical systems such that meaningful statements could be made, for instance, about the probability that a system would become turbulent or about the expected time to the "onset of chaos"? To get started on this very ambitious problem, we discussed how we would go about generating homeomorphisms at random. For simplicity, let us generate homeomorphisms of the unit interval [0,1] onto itself. Thus, we wish to build continuous, strictly increasing maps h with $h(0) = 0$ and $h(1) = 1$. One algorithm for doing this randomly follows.

Set $h(0) = 0$ and $h(1) = 1$. Choose $h(1/2)$ according to the uniform distribution on [0,1]. Continue by choosing $h(1/4)$ and $h(3/4)$ according to the uniform distribution on $[0, 1/2]$ and $[1/2, 1]$, respectively. In general, once the values of $h(i/2^n)$ have been determined for $i = 0, 1, \ldots, 2^n$, choose $h\left((2i+1)/2^{n+1}\right)$ according to the uniform distribution on $\left[h(i/2^n), h(i+1)/2^n\right]$. This simple algorithm is easily implemented on a computer. (It needs no more than fifty lines of FORTRAN.) If the computer's random-number generator is fairly good, general properties of these functions can be guessed. However, to show that this algorithm defines an associated probability measure P on Ω, the set of all homeomorphisms of [0,1] onto [0,1], is no small task. First we need to define a class of elementary events and the probabilities associated with them. An elementary event in the sample space Ω comes naturally from the random algorithm. For a positive integer n, consider the dyadic grid on [0,1] given by the points $1/2^n$, $2/2^n, \ldots, (2^n-1)/2^n$. Over each grid point $i/2^n$ construct a "gate", an interval (a_i, b_i) such that $a_i < b_i \leq a_{i+1}$. An elementary event consists of all elements h of Ω that pass through all the gates: $a_i < h(i/2^n) < b_i$, for $i = 1, 2, \ldots, 2^n - 1$ (Fig. 10).

The probability assigned to an elementary event is defined by induction on n. For example, if $n = 1$, an elementary event consists of all h that pass through a single gate: $a < h(1/2) < b$. Since the random algorithm chooses $h(1/2)$ uniformly, the probability assigned to this event is the length of the interval, $b - a$. If $n > 1$, the probability of an elementary event is determined from the conditional probabilities given by the algorithm. For example, the distribution function of the random variable $h(3/4)$ is $P\left(h(3/4) \leq t\right)$. To calculate this distribution function, we first find the

conditional probability that $h(3/4) \leq t$, given that $h(1/2) = s$. It follows directly from the construction algorithm that

$$P\left(h(3/4) \leq t | h(1/2) = s\right) = \begin{cases} 1 & \text{if } 1 \leq t \\ (t-s)/(1-s) & \text{if } s < t < 1 \\ 0 & \text{if } t \leq s. \end{cases}$$

So,

$$\begin{aligned} P\left(h(3/4) \leq t\right) &= \int_0^1 P\left(h(3/4) \leq t | h(1/2) = s\right) ds \\ &= \int_0^t P\left(h(3/4) \leq t | h(1/2) = s\right) ds \\ &= \int_0^t \left((t-s)/(1-s)\right) ds \\ &= t + (1-t)\ln(1-t). \end{aligned}$$

The distribution of $h(3/4)$ is shown in Fig. 11.

The exact formulas for the probabilities assigned to various elementary events are quite complicated. What is required is to determine that probabilities of the form

$$P\left(h(1/2^n) \leq t_1, h(2/2^n) \leq t_2, \ldots, h\left((2^n-1)/2^n\right) \leq t_{2^n-1}\right)$$

satisfy Kolmogorov's consistency theorem. We have shown that these conditions are indeed satisfied and therefore a probability measure P is defined on the homeomorphisms of [0,1]. To see what these homeomorphisms look like, we used the computer. Figure 12 shows a few samples from our computer studies in which the values of $h(i/2^n)$ are computed for $n = 10$.

S. Graf, S. C. Williams, and I studied this method in detail (Graf, Mauldin, and Williams 1986). For example, we examined a large number of the computer studies and guessed that with probability 1 the derivative of a random homeomorphism at the origin is 0. This conjecture turned out to be correct. The argument is essentially the following. First, since h is increasing and $h(0) = 0$, it is enough to show that

$$\lim_{n\to\infty} \frac{h(1/2^n) - h(0)}{1/2^n} = \lim_{n\to\infty} 2^n h(1/2^n) = 0.$$

Second, set

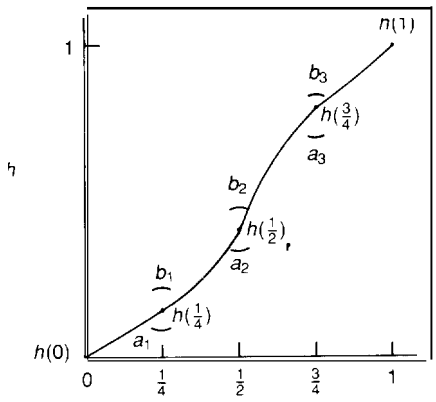
$$\Psi_n(h) = \frac{h(1/2^n)}{h(1/2^{n+1})},$$

where $n = 1, 2, 3, \ldots$. It is intuitively clear and can be proved that $\Psi_1, \Psi_2, \Psi_3, \ldots$ are independent random variables, all uniformly distributed on [0,1]. Set $X_n = \ln \Psi_n$. The X_n's are independent and identically distributed, and $E(X_n) = \int_0^1 \ln t\, dt = -1$. Therefore, by the strong law of large numbers,

$$\lim_{n\to\infty} (1/n) \sum_{p=1}^{n} X_p = -1.$$

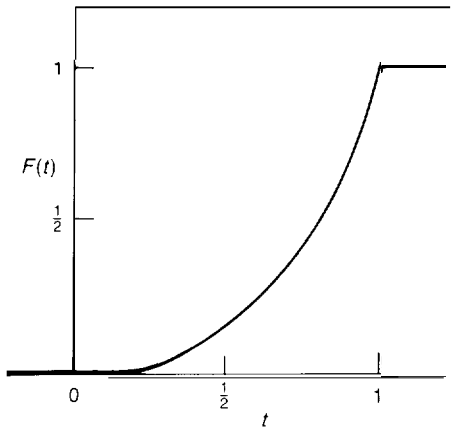
CONSTRUCTION OF ELEMENTARY EVENTS

Fig. 10. In the study of random homeomorphisms described in the text, an elementary event is defined as the set of all homeomorphisms h that pass through $2^n - 1$ "gates" consisting of open intervals (a_i, b_i) over the grid points $i/2^n$ ($i = 1, 2, \ldots, 2^n - 1$). The a_i's and b_i's are restricted by the conditions $a_i < b_i < a_{i+1}$. Shown here is one possible set of gates for $n = 2$ and a member of the corresponding elementary event.



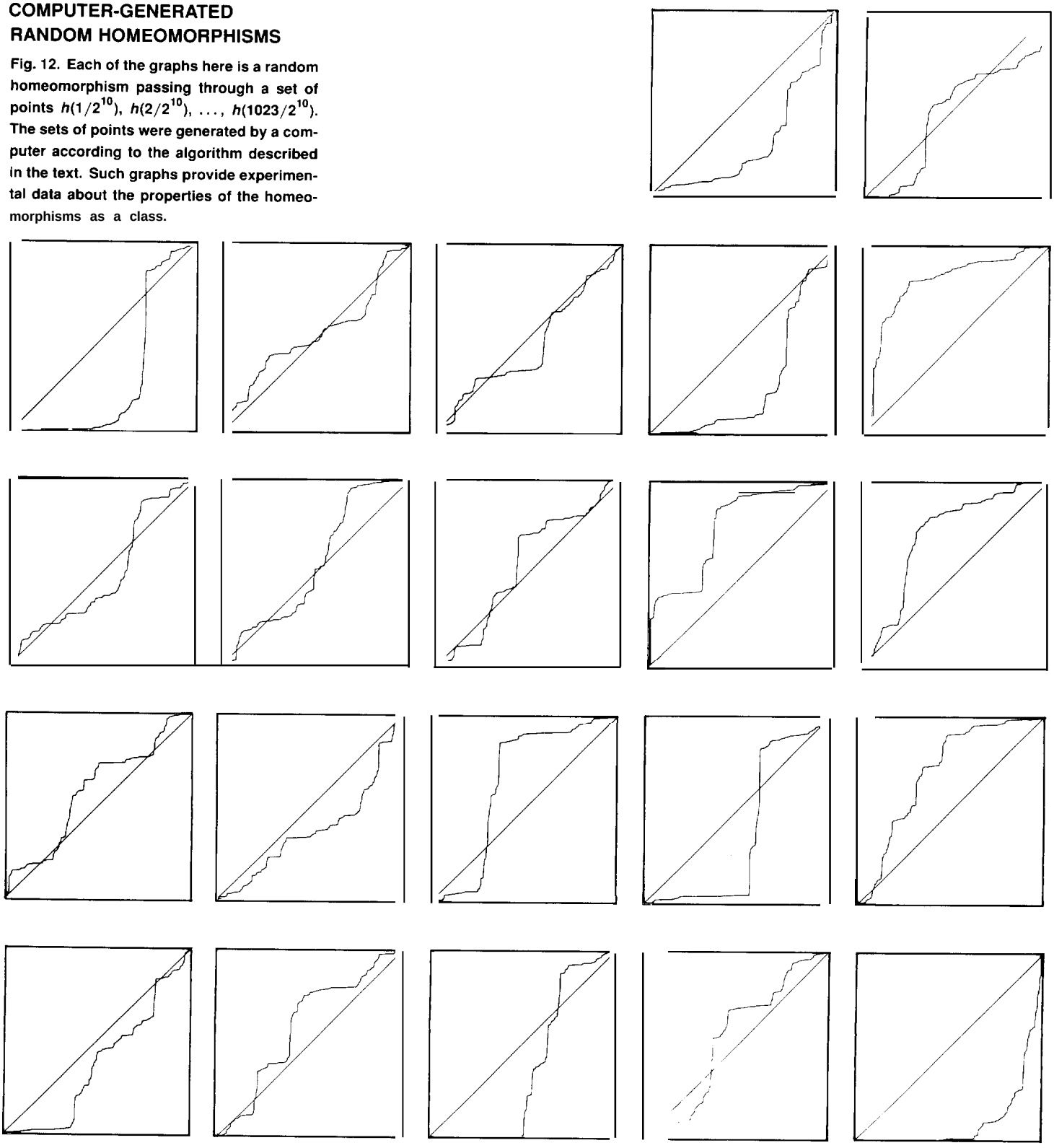
DISTRIBUTION FUNCTION FOR $h(3/4)$

Fig. 11. As demonstrated in the text, $F(t) \equiv P\left(h(3/4) \leq t\right)$ equals 0 for $t < 0$ and equals $t + (1-t)\ln(1-t)$ for $t \geq 0$. Shown here is the graph of that distribution function.



COMPUTER-GENERATED RANDOM HOMEOMORPHISMS

Fig. 12. Each of the graphs here is a random homeomorphism passing through a set of points $h(1/2^{10})$, $h(2/2^{10})$, ..., $h(1023/2^{10})$. The sets of points were generated by a computer according to the algorithm described in the text. Such graphs provide experimental data about the properties of the homeomorphisms as a class.



Multiplying both sides by n we have, with probability 1.

$$-\infty = \lim_{n\to\infty} \sum_{p=1}^{n} X_p = \lim_{n\to\infty} \sum_{p=1}^{n} \ln \Psi_p = \lim_{n\to\infty} \ln \prod_{p=1}^{n} \Psi_p .$$

Exponentiating we get

$$0 = \lim_{n\to\infty} \prod_{p=1}^{n} \Psi_p = \lim_{n\to\infty} 2^n h(1/2^n),$$

which is what we wanted to show.

We have also shown that, with probability 1, a random homeomorphism has a derivative of O almost everywhere, that is, everywhere except for a subset of [0,1] with Lebesgue measure O. Consequently, with probability 1, a random homomorphism is not smooth. Therefore this approach will not yield answers to questions concerning the transition from smooth to turbulent, or chaotic, behavior. As often happened with Stan's problems, the original question, which was motivated by physics, would eventually become a purely mathematical problem.

By the way, our original studies on an Apple computer illustrate the pitfalls of working with numerical results. From looking at the graphs we guessed that the set of fixed points for these homeomorphisms is a Cantor set. When we were unable to prove this conjecture, Tony Warnock conducted more highly resolved computer studies on a Cray, The results suggested not that the fixed points are a Cantor set but rather that a high proportion of the random homomorphisms have an odd number of fixed points (see the accompanying table). This time we guessed that, with probability 1, a random homeomorphism has a finite odd number of fixed points. Indeed we were able to prove this; however, the proof is too complicated to outline here.

A few closing comments on this problem. First, the procedure for generating a random homeomorphism can also be viewed as a procedure for generating a distribution function at random. Thus, we have a probability measure on the space of probability measures! This viewpoint was thought of and developed earlier by Dubins and Freedman. Second, Stan and I did consider the generation of random homeomorphisms on other spaces. For example, the algorithm for generating homeomorphisms of the circle reads almost exactly like that for generating homomorphisms of the interval. (However, in that case we don't know whether there is a positive probability of generating homomorphisms with no periodic points. This is an interesting possibility.) Third, it is possible to bootstrap oneself up from generating homomorphisms of the interval to generating homeomorphisms of the square, the cube, and so on. These possibilities are described in Graf, Williams, and Mauldin 1986. Finally Stan had some wild ideas about "crossing" random homomorphisms with something like Brownian motion to produce flows at random.

That wildness was the joy of being with Stan Ulam. His boundless imagination opened up one's mind to the endless possibilities of creating. It was my good fortune to have known Stan for some ten years as a deep personal friend, a most stimulating collaborator, and an endless source of inspiration. ■

FIXED POINTS OF RANDOM HOMEOMORPHISMS

Listed **here are computer-generated sets of data on the number of fixed points possessed by each of (a) 5000 and (b) 10,000 of the random homeomorphisms (h's) defined in the text. Note the predominance of homeomorphisms with odd numbers of fixed points. That observation led us to conjecture, and to prove, that, with probability 1, any such random homeomorphism has a finite odd number of fixed points.**

Number k of Fixed Points	Number of h's with k Fixed Points (a)	(b)
0	185	510
1	1332	
2	196	**544**
3	876	**1835**
4	179	**418**
5	605	**1138**
6	143	283
7	410	751
8	114	174
9	259	464
10	75	136
11	187	276
12	52	95
13	114	190
14	32	50
15	61	80
16	20	25
17	48	59
18	23	13
19	38	25
20	9	9
21	6	21
22	7	5
23	19	12
24	3	3
25	1	2
26	2	1
27	2	1
28	0	1
29	0	2

Further Reading

The first five works are general; the remainder are those cited in reference to specific topics.

A. I. Khinchin. 1949. *Mathematical Foundations of Statistical Mechanic ,s. New* York: Dover Publications, Inc.

Mark Kac. 1959. *Probability and Related Topics in Physical Sciences. New* York: Interscience Publishers, Inc.

Mark Kac and Stanislaw M. Ulam. 1968. *Mathematics and Logic: Retrospect and Prospects. New* York: Frederick A. Praeger, Inc. Also in Volume 1 of *Britannica Perspectives.* Chicago: Encyclopedia Britannica, Inc.

R. Daniel Mauldin, editor. 1981. *The Scottish Book: Mathematics from the Scottish Café.* Boston: Birkhäuser Boston.

R. D. Mauldin and S. M. Ulam. 1987. Problems and games in mathematics. *Advances in Applied Mathematics 8: 281–344,*

Richard P. Feynman. 1951. The concept of probability in quantum mechanics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,* edited by Jerzy Neyman. Berkeley and Los Angeles: University of California Press,

David Blackwell and R. Daniel Mauldin. 1985. Ulam's redistribution of energy problem. *Letters in Mathematical Physics 60: 149.* (This entire issue is devoted to Stan L-lam.)

S. Ulam. 1980. On the operations of pair production, transmutations, and generalized random walk. *Advances in Applied Mathematics I: 7–21,*

R. D. Mauldin and S, C. Williams. 19X6. Random recursive constructions. *Transactions of the American Mathematical Society 295: 325–346.*

S. Graf, R. Daniel Mauldin, and S. C. Williams. 1986. Random homeomorphisms. *Advances in Mathematics* 60: 239

R. Daniel **Mauldin** received his Ph.D. in mathematics from the University of Texas in 1969, In 1977, after eight years at the University of Florida, he joined the faculty at North Texas State University, where he is currently the Decker Science Fellow. He is a frequent visitor to the Laboratory, Some of his current research interests involve deterministic and random recursions and the asymptotic geometrical and measure-theoretic properties of objects defined by these processes. He is a member of the American Mathematical Society and an editor of its Proceedings,