WHAT IS THE ERGODIC THEOREM?

G. D. BIRKHOFF, Harvard University

The integral of Lebesgue (1901), founded upon Borel measure, has been a dominating weapon in the striking advance of Analysis during the present century. Perhaps the Ergodic Theorem (1931) is destined to hold a central position in this development. Indeed, Wiener and Wintner in a recent article* refer to it as "the only result of real generality established for the solutions of dynamical systems."

To understand the theorem and the nature of its applications it is necessary first of all to say something about (Borel-Lebesgue) measure, *i.e.*, "probability" in the sense sketched by Poincaré in the third volume of his *Méthodes Nouvelles de la Mécanique Céleste*. We restrict ourselves to the case of a line segment of unit length with coördinate x, $0 \leqq x \leqq 1$. Suppose that we have a set of non-overlapping intervals, finite in number and of total length $l < 1$ in this segment. The probability in a certain intuitive sense that a point, *taken at random*, lies in one of these intervals, is l; and the probability that it lies in the complementary set is of course $1-l$.

Now suppose that we are given a point set M containing an infinite number of points, which can be enclosed within an infinite set of non-overlapping intervals of lengths $l_1, l_2, \cdots$ of total length.

$$l_1 + l_2 + l_3 + \cdots = l < 1.$$

Then clearly the probability that a point, taken at random, lies in M, cannot exceed l; and the probability that it lies in the complementary set is at least $1-l$. If now M is of such a nature that it can be enclosed in an infinite set of intervals of total length not exceeding an arbitrarily small quantity ϵ, it is apparent that the probability of a random point falling in M does not exceed ϵ, *i.e.* the probability is 0. Such a set M is said to be of measure 0.

For instance, the set of rational points $x = m/n$ which is everywhere dense on the line segment, is of measure 0. In fact these points may be arranged in order

$$0,\ 1;\ \tfrac{1}{2};\ \tfrac{1}{3},\ \tfrac{2}{3};\ \tfrac{1}{4},\ \tfrac{3}{4};\ \tfrac{1}{5},\ \tfrac{2}{5},\ \tfrac{3}{5},\ \tfrac{4}{5};\ \cdots$$

and the nth one of these points may obviously be enclosed within an interval of length $\epsilon/2^n$. Since we have

$$\frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{8} + \cdots = \epsilon,$$

* On the ergodic dynamics of almost periodic systems, American Journal of Mathematics, vol. 63, 1941. For an introduction to the literature see Eberhard Hopf's "Ergodentheorie," Ergebnisse der Mathematik und ihrer Grenzgebiete, Berlin, Springer, 1937. Our discussion here deals only with the "Ergodic Theorem," and not at all with the "Mean Ergodic Theorem" of von Neumann, which stimulated me to reconsider some old ideas, and so led me to the discovery and proof of the Ergodic Theorem, embodying a strong, precise result which, so far as I know, had never been hoped for.

it is evident that this set of rational points is of measure 0.

More generally, if we have a set M such that it can be enclosed within a set of intervals of length $l_1, l_2, \cdots$ with

$$l_1 + l_2 + \cdots \leqq l + \epsilon$$

while the complementary set $\overline{M}$ can be enclosed similarly within intervals $\bar{l}_1, \bar{l}_2, \cdots$ with

$$\bar{l}_1 + \bar{l}_2 + \cdots \leqq (1 - l) + \epsilon$$

for $\epsilon > 0$ arbitrarily small, then $\overline{M}$ is said to be measureable of measure l; and its complementary set M will then clearly be measurable of measure $1-l$. In this case the probability that a random point falls in M is obviously to be regarded as l.

All ordinary infinite sets specifically defined by analytic methods are found to be measureable in this sense.

The gist of the Ergodic Theorem can now be illustrated by means of our line segment.

Suppose that there is given any one-to-one *measure preserving* transformation T of the line segment $0 \leqq x \leqq 1$ into itself; T may have a finite or infinite number of discontinuities. A first simple example is the following: Imagine the line segment $0 \leqq x < 1$ bent into a circle of circumference 1, without any stretching; the first transformation T is merely a rotation of this circle through a certain angle α. A second example is the following: The line segment is divided into the infinite set of intervals,

$$0 \leqq x < \tfrac{1}{2}; \ \tfrac{1}{2} \leqq x < \tfrac{3}{4}; \ \tfrac{3}{4} \leqq x < \tfrac{7}{8}, \cdots$$

and then the second interval is interchanged with the first, the fourth with the third, etc., thus defining the transformation T. In both cases T is evidently of the stated type, and measure is preserved.

The Ergodic Theorem then says: *For any such measure-preserving transformation T, and for each individual point P (except possibly an exceptional set of measure 0), there is a definite probability that its iterates under T, from P on, namely*

$$P, T(P), T^2(P), \cdots \quad \text{and} \quad P, T^{-1}(P), T^{-2}(P), \cdots$$

fall in a given measurable set M.

In other words the proportion of n of these points (beginning with P) which lie in the set M tends toward a definite limit μ_p, as n approaches infinity in either direction.

More generally, a line segment may be replaced by a finite volume M of n-dimensions, $n > 1$, and the points of M may be assigned a variable (integrable) positive weight, $w(P)$. The generalized theorem would then assert that the corresponding weighted means tend toward a limit μ_p. In the simple special case first stated, this weight is 1 for the points of M and 0 for the points not in M.

Or, again, for $n>1$ the discrete transformation T may be replaced by a steady measure-preserving flow T_t in time t, and the analogous theorem holds.

To illustrate this last possibility, suppose that in the square $0 \leqq x < 1$, $0 \leqq y < 1$, the points move with a uniform velocity in a fixed direction, making an angle α with that of the x axis, and leaving the square to return at the homologous point (see the adjoining figure). Evidently such a transformation T_t is area-preserving. Let now M be any selected measurable part of the square, and let P be any point of the square—aside always from a possible exceptional

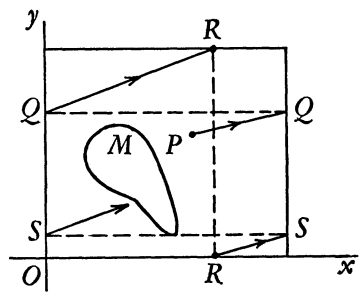

FIG 1.

set of measure 0. On the basis of the same theorem, there is a definite probability in infinite time, $t \geqq 0$ or $t \leqq 0$ that $P_t = T_t(P)$ falls within M, and this probability is the same in both directions. More generally a weight $w(P)$ may be introduced in the case of a "flow" as well as in the discrete case.

In more analytic garb, the theorem states in the two cases respectively that for $n \to \pm\infty, T \to \pm\infty$:

$$\frac{w(P) + w(T(P)) + \cdots w(T^{n-1}(P))}{n} \to \mu_P; \qquad \frac{1}{T}\int_0^T w(P)dP \to \mu_P.$$

The kind of applications to dynamical systems which the Ergodic Theorem affords are exceedingly varied and interesting. Take the simple example of an idealized convex billiard table on which an idealized billiard ball P moves with velocity 1. In the figure let $\phi = \text{arc } OA$, $\phi_1 = \text{arc } OA_1$, $l = AP$, $l^* = AA_1$. We have a transformation $(\theta_1, \phi_2) = T(\theta, \phi)$ defined over a rectangle

$$0 < \theta < \pi; \qquad 0 \leqq \phi \leqq p, \qquad (p = \text{perimeter of table})$$

in the $\theta\phi$-plane, associated with the motion. It is not hard to prove that T is measure-preserving in the sense that the double integral

$$\iint \frac{\sin\theta}{\sin\theta_1}\, d\theta d\phi$$

has the same value when extended over any measurable part of this rectangle

as over its image under T; indeed it would be possible to deform the rectangle so that, over the new region, ordinary areas are preserved.

Furthermore it is clear that, if we associate with any "state of motion" of

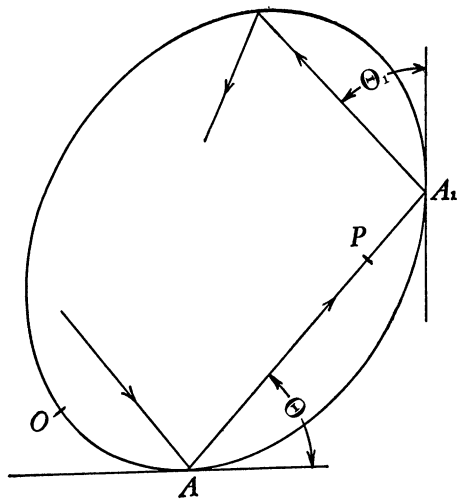

FIG. 2

the billiard ball, as of P, the three coordinates θ, ϕ, l then a steady flow T_t is defined in the corresponding region of three-dimensional $\theta\phi l$-space:

$$0 < \theta < \pi; \qquad 0 \leqq \phi < p, \qquad 0 \leqq l \leqq l^*$$

in which the following volume integral is preserved:

$$\int \left(\int\int \frac{\sin \theta}{\sin \theta_1} \, d\theta d\phi \right) dl.$$

Thus the theorem applies to this flow.

Here are three obvious applications to this simple but typical dyanamical problem:

(1) the average length of n successive chords of the path tends to a definite limit, the same whether the time t increases or decreases;

(2) the average angle θ at n successive collisions tends to a definite limiting value;

(3) the billiard ball tends in the limit to lie in any assigned area of the table a definite proportion of the time.

There is one especially interesting case, which may in fact be the "general case" as far as we know: It may happen that all of the points of our volume behave in essentially the same way in the mean (aside always from the excepted set of measure 0, of course). If they do not so behave, the underlying space can

be subdivided into *invariant* measurable sets; thus for an elliptical table, the motions lying wholly in the ring outside a smaller confocal ellipse form such a closed invariant set; and this is an integrable problem—a limiting case of geodesics on a flattening ellipsoid.

What the Ergodic Theorem means, roughly speaking, is that for a discrete measure-preserving transformation or a measure-preserving flow of a finite volume, probabilities and weighted means tend toward limits when we start from a definite state P (not belonging to a possible exceptional set of measure 0), and, furthermore, the limiting value is the same in both directions.

The Ergodic Theorem applies to manifold deep problems of analysis and of applied mathematics—as well to the solar system as to our simple billiard ball problem! Thus in G. W. Hill's celebrated idealization of the earth-sun-moon problem (the restricted problem of three bodies) we can at once assert (with probability 1) that the moon possesses a true mean angular state of rotation about the earth (measured from the epoch), the same in both directions of the time.

FOCAL CUBICS ASSOCIATED WITH FOUR POINTS IN A PLANE*

M. G. BOYCE, Western Reserve University

1. Introduction. Let A, B, C, D be distinct fixed points and Z a variable point, all in the same plane. The locus of Z such that the directed angles AZB and CZD are equal, or equal when reduced modulo π, will be referred to as the *equal-angle locus* with respect to the four points. Similarly, if the distance ratios AZ/BZ and CZ/DZ are equal, the locus of Z will be called the *equal-ratio locus*. Any ordered set of four points will be said to form an $\mathcal{A}$*-basis* for a given curve if the curve is an equal-angle locus for the four points in that order, and an $\mathcal{R}$*-basis* is defined in the corresponding manner.

Each of the two loci just described has long been known to be a circular cubic which passes through its singular focus. Such a cubic is called a *focal cubic*, since it is the locus of the foci of the conics cut from a cone of second degree by a pencil of planes whose axis is tangent to the cone and perpendicular to a principal section. Quetelet initiated the study of these focal loci in his inaugural dissertation [Gand, 1819] on the case in which the cone is right circular. The focal curve is then identical with the oblique strophoid. Van Rees† soon afterward discussed the general case, Chasles and others followed with additional contributions, and Teixeira‡ in 1908 included a rather comprehensive treatment of focal cubics in his treatise on special curves.

* A preliminary report on this paper was presented to the Ohio Section of the Mathematical Association of America, April 8, 1939.

† K. Van Rees, Memoire sur les focales, Correspondance mathématique et physique de A. Quetelet, vol. 5, Brussels, 1829, pp. 361–378.

‡ F. Gomes Teixeira, Traité des Courbes Speciales Remarquables, vol. 1, Coïmbre, 1908, pp. 45–58.