Ergodic theorem, ergodic theory, and statistical mechanics

Calvin C. Moore

Department of Mathematics, University of California, Berkeley, CA 94720

This perspective highlights the mean ergodic theorem established by John von Neumann and the pointwise ergodic theorem established by George Birkhoff, proofs of which were published nearly simultaneously in PNAS in 1931 and 1932. These theorems were of great significance both in mathematics and in statistical mechanics. In statistical mechanics they provided a key insight into a 60-y-old fundamental problem of the subject—namely, the rationale for the hypothesis that time averages can be set equal to phase averages. The evolution of this problem is traced from the origins of statistical mechanics and Boltzman’s ergodic hypothesis to the Ehrenfests’ quasi-ergodic hypothesis, and then to the ergodic theorems. We discuss communications between von Neumann and Birkhoff in the Fall of 1931 leading up to the publication of these papers and related issues of priority. These ergodic theorems initiated a new field of mathematical-research called ergodic theory that has thrived ever since, and we discuss some of recent developments in ergodic theory that are relevant for statistical mechanics.

George D. Birkhoff (1) and John von Neumann (2) published separate and virtually simultaneous path-breaking papers in which the two authors proved slightly different versions of what came to be known (as a result of these papers) as the ergodic theorem. The techniques that they used were strikingly different, but they arrived at very similar results. The ergodic theorem, when applied say to a mechanical system such as one might meet in statistical mechanics or in celestial mechanics, allows one to conclude remarkable results about the average behavior of the system over long periods of time, provided that the system is metrically transitive (a concept to be defined below). First of all, these two papers provided a key insight into a 60-y-old fundamental problem of statistical mechanics, namely the rationale for the hypothesis that time averages can be set equal to phase averages, but also initiated a new field of mathematical research called ergodic theory, which has thrived for more than 80 y. Subsequent research in ergodic theory since 1932 has further expanded the connection between the ergodic theorem and this core hypothesis of statistical mechanics.

The justification for this hypothesis is a problem that the originators of statistical mechanics, namely the rationale for the hypothesis that time averages can be set equal to phase averages, but also initiated a new field of mathematical success. J. W. Gibbs in his 1902 work (5) argued for his version of the hypothesis based on the fact that using it gives results consistent with experiments. The 1931–1932 ergodic theorem applied to the phase space of a mechanical system that arises in statistical mechanics and to the one-parameter group of homeomorphisms representing the time evolution of the system asserts that for almost all orbits, the time average of an integrable function on phase space is equal to its phase average, provided that the one-parameter group is metrically transitive. Hence, the ergodic theorem transforms the question of equality of time and phase averages into the question of whether the one-parameter group representing the time evolution of the system is metrically transitive.

To be more specific about statistical mechanics systems, consider a typical situation in gas dynamics where one has a macroscopic quantity of a dilute gas enclosed in a finite container. The molecules are in motion, colliding with each other and with the hard walls of the container. The molecules can be assumed for instance to be hard spheres (billiard balls) bouncing off each other or alternately may be assumed to be polyatomic molecules with internal structure and where collisions are governed by short-range repelling potentials. One may also choose to include the effects of external forces, such as gravity on the molecules. We assume that the phase space $M$ consists of a surface of constant energy. This assumption, together with the finite extent of the container, ensures that $M$ is compact and that the invariant measure derived from Liouville’s theorem is finite. The equations of motion, say in Hamiltonian form, can be written in local coordinates as a first-order system of ordinary differential equations

$$\frac{dx_i}{dt} = X_i(x_1, \ldots, x_n).$$

First, the number of variables in the equations is enormous, perhaps on the order of Avogadro’s number, and the equations are quite complex. The system is perfectly deterministic in principle; hence, given the initial positions and momenta of all of the molecules at an initial time, the system evolves...
that a physical measurement takes a short period to perform, but this short time period is a long period for the physical system because, for instance, each molecule will on average experience billions of collisions per second in a typical gas. In fact, this line of reasoning leads to the idea that the result of the measurement is best represented by the limit as $T$ tends to infinity of the time average above. The first problem is, of course, to know that this limit exists, except of course for a negligible set of $x$. Then the assumption of equality of time averages and phase averages would assert that the limit of the time average above is independent of $x$ and equal to the phase average

$$\int_M f(v)\,d\mu(v),$$

for all but a negligible set of $x \in M$ where $\mu$ is the Liouville invariant measure. The significant and useful point is that this phase average can be calculated in many cases, whereas the time average cannot be calculated.

Another way of phrasing this equality is to use for $f$ the indicator function of a measurable subset $A$ of $M$ (a function that is equal to 1 on $A$ and 0 outside of $A$). Then the time averages above record the fraction of time that the orbit spends in $A$, and the basic hypothesis of statistical mechanics asserts that this is equal for almost all orbits to the Liouville measure of $A$ (assuming that the measure of the total space $M$ has been normalized to 1). However, another way of expressing this equality is to assert that for all but a negligible set of states of the gas, the observed value of a function $f$ will be equal to the average value of $f$ taken over $M$ that is an average value of $f$ taken over an ensemble (to use Gibbs’s language) of all possible states with the same energy. The constant energy surface $M$ with its invariant volume element here is what Gibbs called the microcanonical ensemble.

The ergodic theorems of Birkhoff and von Neumann assert first of all of the existence of the time limit for $T \to \infty$ for any one parameter measure preserving group, and then, assuming that $P_t$ is metrically transitive, they assert the equality

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(P_t(x))\,dt = \frac{1}{M} \int_{M} f(v)\,d\mu(v).$$

The difference between the two theorems is that Birkhoff proved that the convergence of the functions of $x$ on the left side is pointwise almost everywhere (the limit in general can be identified as the conditional expectation of $f$ onto the sigma field of invariant sets, to use the language of probability theory). In the case of metric transitivity, this function is just the constant function equal to the integral of $f$ over $M$. Von Neumann proved that these functions of $x$ converge in mean square [that is in $L^2(\mu)$] to the orthogonal projection of the function $f$ onto the closed subspace of invariant functions, which in the metrically transitive case is one dimensional consisting of the constant functions. Birkhoff assumes the function $f$ is bounded and measurable, whereas von Neumann assumes the more general condition that the function $f$ is square integrable. Although both theorems were originally formulated and proved for measure preserving one parameter groups generated by first-order differential equations on compact manifolds, subsequent work has shown using the same arguments that these results are valid for a much broader class of dynamical systems including one-parameter families of measure preserving transformations of a finite measure space, which may not necessarily be defined by systems of differential equations. Later work also showed that Birkhoff’s theorem holds for an integrable function $f$. Thus, these theorems are theorems about one-parameter groups of automorphisms of measure spaces with no mention of topology. The theorems also clarify what is meant by the informal term in statistical mechanics, negligible set, namely, it is a set of $\mu$ that measures zero. It should be added that the time average does not have to be taken from $-T$ to $T$, but can be taken over any intervals from $T_1$ to $T_2$, as long as the difference $T_2 - T_1$ tends to infinity.

Before moving on to subsequent developments in ergodic theory, it is worth pausing to examine the sequence of events leading to the proofs and publication of the two ergodic theorems: the pointwise ergodic theorem of Birkhoff and the mean ergodic theorem of von Neumann. Much of this was laid out in a subsequent paper of Birkhoff and Koopman (8) in the March 1932 issue of the PNAS. Von Neumann was very much aware of the results of M. H. Stone (9) on spectral theory of one-parameter groups of unitary operators and the results of Koopman (10) that used Stone’s results to analyze one-parameter groups of measure preserving transformations. Koopman had indeed suggested to von Neumann that he might use these results to resolve the problem of equality of time and phase averages, and von Neumann writes that André Weil had made the very same suggestion to him. Von Neumann seized on the notion of metric transitivity, introduced, somewhat ironically, by

Birkhoff and Smith (11) a few years earlier in 1928, and proved his mean ergodic theorem under the hypothesis of metric transitivity. See the article by Mackey (12) for more details.

According to Birkhoff and Koopman, von Neumann communicated his result personally to both of them on October 22, 1931, and pointed out to them that his result raised the important question of whether a pointwise result might be valid. Birkhoff then went to work, and, by different methods, quickly established his pointwise ergodic theorem. He submitted his paper to PNAS on December 1, 1931, for appearance in the December 1931 issue. One presumes that he sent copies to Koopman and von Neumann, who would have noticed that Birkhoff had not given von Neumann adequate credit and recognition for his result. von Neumann evidently planned to include his ergodic theorem and its proof in a much longer paper he was writing for the Annals of Mathematics, but he then apparently quickly drafted a short paper for PNAS with his proof of the mean ergodic theorem and submitted it to PNAS on December 10, 1931. It appeared in the January 1932 issue. One suspects that these events led Koopman and Birkhoff to write and publish their paper in PNAS 2 months later, which set matters straight and clearly acknowledged von Neumann’s priority. It should also be noted that E. Hopf (13) presented a slightly different proof of the mean ergodic theorem and some improvements on the Birkhoff theorem in a paper, which appeared in the January 1932 issue of PNAS. For whatever reason, the Birkhoff paper and its result has over time become accepted von Neumann adequate credit and recognition for his result.

There are also ergodic theorems for a single measure-preserving map \( P \) and its iterates \( P^n \). They assert the existence of the time limit below and that it converges to the phase average if \( P \) is metrically transitive; that is

\[
\lim_{N \to \infty} \left( N + 1 \right)^{-1} \sum_{n=0}^{N} f(P^n x) = \int f(y) d\mu(y).
\]

The convergence is pointwise for almost all \( x \) for integrable \( f \), and in the mean for \( f \) square summable. One way to conceive of metric transitivity and the ergodic theorem for a single transformation is that for almost all points \( x \), the \( n \) iterates under \( P \) of \( x \) is distributed in some sense evenly throughout the space so that taking the average of a function \( f \) over these points gives a result that is a good approximation to the integral of \( f \) over the space and that the more iterates one includes in the average, the better the approximation. Therefore, it is like a numerical integration scheme.

Finally, we need to define metric transitivity, a concept, as previously noted, that was introduced by Birkhoff and Smith (11). A one-parameter group of measure preserving transformation \( P_t(x) \) (or a single transformation \( P \)) on a measure space \( M \) is metrically transitive provided that any \( \mu \) measurable set invariant under \( P_t \) for all \( t \) (or \( P \)) must have zero measure or its complement must have zero measure. This means that the flow is indecomposable or irreducible in the sense that one cannot decompose it into a union of two disjoint subflows. It also means that there are no measurable functions invariant under the flow (or the transformation \( P \)).

It is heuristically reasonable to argue, owing to the molecular chaos in gas dynamics, that there are no nonconstant continuous invariants or so-called first integrals of the motion. However, more is required for metric transitivity—namely no nonconstant measurable invariants of the motion. In the example from gas dynamics, the total energy is clearly an invariant of the motion, but we have restricted the flow to a surface of constant energy. In addition, total momentum is normally preserved, but although momentum is preserved in collisions between molecules, collisions with the walls do not preserve momentum, so this possible invariant of the motion disappears. Although the term metric transitivity is still in use, current terminology, due to von Neumann, is that any flow or single transformation with this property is simply called ergodic.

It is worth observing that metric transitivity is a necessary and sufficient condition for the validity of the ergodic theorem. To see this, assume the ergodic theorem holds and then apply the statement of the theorem to the indicator function \( f \) of a supposed invariant measurable set \( A \)—that is, \( f \) is equal 1 on \( A \) and 0 on the complement of \( A \); the left side of time averages is always equal to \( f \), but the right side is a constant function. Hence, \( f \) is a constant function, and the alleged invariant set is of measure zero or its complement is of measure zero.

It is interesting to look back at the early history of statistical mechanics to see how the founders of the subject handled the topic of time averages and space averages. Boltzmann (4) coined the terms ergodier or ergodische (which we translate as ergodic) from the Greek εργός (energy) and οδός (path) or energy path. He put forth what he called the ergodic hypothesis, which postulated that the mechanical system, say for gas dynamics, starting from any point, under time evolution \( P_t \), would eventually pass through every state on the energy surface. Maxwell and his followers in England called this concept the continuity of path (3). It is clear that under this assumption, time averages are equal to phase averages, but it is also equally clear to us today that a system could be ergodic in this sense only if phase space were one dimensional. Plancherel (14) and Rosenthal (15) published proofs of this, and earlier, Poincare (16) had expressed doubts about Boltzmann’s ergodic hypothesis. Certainly part of the problem Maxwell and Boltzmann faced was that the mathematics necessary for a proper discussion of the foundations of statistical mechanics, such as the measure theory of Borel and Lebesgue, and elements of modern topology had not been discovered until the first decade of the 20th century and were hence unavailable to them.

In their influential 1911 article, Ehrenfest and Ehrenfest (17) summarized and discussed problems with the ergodic hypothesis and then proposed instead the quasi-ergodic hypothesis as a replacement. This hypothesis states that some orbit of the flow will pass arbitrarily close to every point of phase space, or in other words this orbit is topologically dense in the phase space. This hypothesis is a far more plausible one than the old ergodic hypothesis, and it does imply that any continuous function invariant under the flow is constant. Some authors [von Plato (18)] have argued that, despite what Boltzmann had written down most of the time in his articles about the ergodic hypothesis, that he probably really meant something like what was later termed the quasi-ergodic hypothesis. However, the quasi-ergodic hypothesis does not imply metric transitivity. For instance, it is not even true that a minimal flow (every orbit is dense) with an invariant measure is metrically transitive [see Furstenberg (19)], for examples. Therefore, although the original ergodic hypothesis was too strong to be plausible, the quasi-ergodic hypothesis was too weak to establish equality of time and phase averages. Further mathematical, progress had to await the concept of metric transitivity and the ergodic theorems of 1931 and 1932. For more details, see the survey article of Mackey (20).

One reaction to the Birkhoff and von Neumann ergodic theorems might be that they do not really solve the problem of equating time average and phase averages but only reduce it to a possibly equally difficult problem of proving metric transitivity. For instance, how can one prove that a one-parameter flow is metrically transitive?
and indeed how do you know metrically transitive systems exist at all. At this point, let us transfer to current terminology and simply call metrically transitive transformations ergodic, as von Neumann suggested.

As to the existence of ergodic transformations, Oxtoby and Ulam (21) showed that on a compact polyhedron \( M \) equipped with a finite Lebesgue–Stieljes measure, the set of all ergodic measure preserving homeomorphisms is a dense \( G \delta \) or a residual set, among all measure preserving homeomorphisms. Hence, not only do ergodic maps exist, but almost all measure preserving homeomorphisms are ergodic in a topological sense. However, there is a different answer for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existence problem. He shows that any one-parameter flow or power for Hamiltonian dynamical systems. Here almost all systems are nonergodic, in the same sense as above (22). The space and the topology are different in these two cases so there is no contradiction.

von Neumann in his Annals of Mathematics paper (23) provides an intriguing and powerful answer to the existen...
take on only a finite number of values so
the flow on a phase space of fixed energy
cannot be ergodic. However, if one has
a billiard table of more complex geometry,
the situation becomes more interesting. For
instance, if the billiard table is a polygon,
then Kerckhoff et al. (30) show that for
topologically almost all polygons, the bil-
liard flow on a phase space of constant en-
ergy is ergodic. In particular, one has to
stay away from rational polygons where
all of the angles are rational multiples of
π. The authors point out an interesting cor-
ollary of this result, which is that a mechan-
ical system of two particles of masses \( m_1 \)
and \( m_2 \) moving along a finite track without
friction, bouncing off each other elastically
and bouncing off the fixed end points is
ergodic on a phase space of constant energy
for topologically almost all values of the
ratio \( m_1/m_2 \). This is sort of like a one-
dimensional gas.

Finally, Simanyi (31) has established that a
system consisting of two hard spheres
contained in a cube of any dimension at
least two bouncing off each other and off
the hard walls is ergodic on any surface
of constant energy. This appears to be a rig-
orous ergodicity result in a situation that
comes closest to an actual gas. However, this
whole array of theorems, of which we have
mentioned only a few, suggest that the hy-
pothesis of ergodicity (or metric transitivity)
for a physical system like that of gas dynamics
mentioned at the outset of this essay is very
plausible.