1

Introduction and preliminaries

Without going into the details (to which the rest of the book is devoted), we mention some of the basic questions, examples, and constructions of ergodic theory, in order to provide an indication of the content and flavor of the subject as well as to establish reference points for terminology and notation. The final section presents a few facts from measure theory and functional analysis that will be used repeatedly.

1.1. The basic questions of ergodic theory

Ergodic theory is the mathematical study of the long-term average behavior of systems. The collection of all states of a system forms a space X. The evolution of the system is represented by a transformation $T: X \to X$, where Tx is taken as the state at time 1 of a system which at time 0 is in state x. If one prefers a continuous variable for the time, he can consider a one-parameter family $\{T_t : t \in \mathbb{R}\}$ of maps of X into itself. When the laws governing the behavior of the system do not change with time, it is natural to suppose that $T_{s+t} = T_s T_t$, so that $\{T_t : t \in \mathbb{R}\}$ is a *flow*, or group action of $\mathbb{R}$ on X. A single (invertible) transformation $T: X \to X$ also determines the action of a group, namely the integers $\mathbb{Z}$, on X. The actions of arbitrary groups, and even of semigroups in case the transformations may not all be invertible, are worthy of study, but we will be interested mainly in the action of the powers of a single transformation and, occasionally, of a flow.

In order to analyze a system mathematically, one needs to have some structure on X and restrictions on T. There are three major cases:

(1) X is a differentiable manifold and T is a diffeomorphism, the case of *differentiable dynamics*;
(2) X is a topological space and T is a homeomorphism, the case of *topological dynamics*;

(3) X is a measure space and T is a measure-preserving transformation, the case of *ergodic theory*.

Of course the three cases overlap extensively, and a single example can be viewed in different lights; in fact, some of the most interesting problems concern the relationships among the three areas.

Let us define more carefully the case (3) of most interest for us. Let $(X, \mathcal{B}, \mu)$ be a complete probability space, that is, a set X together with a σ-algebra $\mathcal{B}$ of measurable subsets of X and a countably additive nonnegative set function μ on $\mathcal{B}$ such that $\mu(X) = 1$ and such that $\mathcal{B}$ contains all subsets of sets of measure 0. Let $T: X \to X$ be a one-to-one onto map such that T and T^{-1} are both measurable: $T^{-1}\mathcal{B} = T\mathcal{B} = \mathcal{B}$. Since sets of measure 0 don't matter, we don't care if T only becomes well-defined and one-to-one onto after a set of measure 0 is discarded from X. Assume further that $\mu(T^{-1}E) = \mu(E)$ for all $E \in \mathcal{B}$. A map T satisfying these conditions is called a *measure-preserving transformation* (abbreviated m.p.t.), and the systems $(X, \mathcal{B}, \mu, T)$ will be our fundamental objects of study.

(Sometimes one wishes to consider possibly noninvertible maps $T: X \to X$ such that $T^{-1}\mathcal{B} \subset \mathcal{B}$ and $\mu T^{-1} = \mu$, and many of the results we present extend to such maps, but we will restrict ourselves for the most part to the invertible case. There is also an interesting theory of *nonsingular* maps ($T^{-1}\mathcal{B} \subset \mathcal{B}$, $\mu T^{-1} \ll \mu$) which we do not have space to discuss here.)

In the one-parameter case, we assume that T_t is a m.p.t. for all $t \in \mathbb{R}$, the map $(x, t) \to T_t x$ is jointly measurable from $X \times \mathbb{R}$ to X, T_0 is the identity, and $T_{s+t} = T_s T_t$ for all $s, t \in \mathbb{R}$.

If $T: X \to X$ is a m.p.t., the *orbit* $\{T^n x: n \in \mathbb{Z}\}$ of a point $x \in X$ represents a single complete history of the system, from the infinite past to the infinite future. The σ-algebra $\mathcal{B}$ is thought of as the family of observable events, with the T-invariant measure μ specifying the (time-independent) probabilities of their occurrences. A measurable function $f: X \to \mathbb{R}$ represents a measurement made on the system; $f(x), f(Tx), f(T^2x), \ldots$ may be thought of as the values of some physically interesting variable at successive instants of time, beginning with the world in initial state x. In statistical mechanics, information theory, and other areas of application it is interesting and sometimes necessary to consider the long-term time average

$$\frac{1}{N} \sum_{k=0}^{N-1} f(T^k x)$$

of a large number N of successive observations. (The average may not really be 'long-term', because the time unit involved may be very short:

in statistical mechanics, for example, molecular collisions may occur so often that we can actually observe *only* 'long-term' averages of any variable f.) A basic question of ergodic theory is that of the *convergence of these averages*: when does

$$\bar{f}(x) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} f(T^k x)$$

exist in some sense? If it exists, $\bar{f}(x)$ may be thought of as an equilibrium or central value of the variable f.

The convergence had been proved in special cases earlier (e.g. Borel's Strong Law of Large Numbers (1909) in case the $f\,T^k$ are independent and identically distributed), but the general convergence in the mean square (L^2) sense was proved by von Neumann and the almost everywhere convergence by Birkhoff, both in 1931. Their results are known as the Mean Ergodic Theorem and the Ergodic Theorem (or Pointwise Ergodic Theorem), respectively. Generalizations and improvements of these results have been appearing continually since 1932; we will have space for only a few in this book.

The question of the existence of averages occurred to mathematicians because of the physicists' concern with the 'ergodic hypothesis', which was formulated in an (erroneous and unsuccessful) attempt to bring about the conclusion that the time mean

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} f(T^k x)$$

and the space mean

$$\int_X f \, d\mu$$

coincide almost everywhere. It is desirable that the equilibrium or central value of a physical variable coincide with its weighted average over all possible states of the system. Boltzmann felt that if orbits penetrated all corners of the space, then this useful and symmetrical conclusion would follow. The investigation of the conditions under which this equality of time and space means, as well as even stronger conclusions, holds forms a second major topic in ergodic theory, the study of general *recurrence properties*, by which we mean the qualitative behavior of orbits.

If the time mean of every measurable function coincides almost everywhere (a.e.) with its space mean, the system (or just T itself) is called *ergodic*. It turns out that a system is ergodic if and only if the orbit of almost

every point visits each set of positive measure, or, equivalently, if $\mu(E) > 0$ and $\mu(F) > 0$ implies that $\mu(T^n E \cap F) > 0$ for some n. A recurrence property which implies this one is *strong mixing*: $\lim_{n\to\infty} \mu(T^n E \cap F) = \mu(E)\mu(F)$ for all $E, F \in \mathscr{B}$. Weak mixing lies between the two, and just plain *recurrence* – if $\mu(E) > 0$ then $\mu(T^n E \cap E) > 0$ for some n – always holds in a space of finite measure.

A third major question of ergodic theory is the *classification problem*. Let us say that two systems $(X, \mathscr{B}, \mu, T)$ and $(Y, \mathscr{C}, \nu, S)$ are *metrically isomorphic* if there are sets of measure zero $X_0 \subset X$ and $Y_0 \subset Y$ and a one-to-one onto map $\phi: X \backslash X_0 \to Y \backslash Y_0$ such that $\phi T = S\phi$ on $X \backslash X_0$ and $\mu(\phi^{-1} E) = \nu(E)$ for all measurable $E \subset Y \backslash Y_0$. (Such a map ϕ is sometimes called an *isomorphism mod 0*.) How can we tell whether two given systems are (metrically) isomorphic to one another? A classical way is by attaching isomorphism invariants to systems, of which ergodicity, weak mixing, and strong mixing are examples. These are, however, only *spectral invariants*. The map $T: X \to X$ determines a unitary operator $U = U_T$ on $L^2(X)$ by $U_T f(x) = f(Tx)$ (Koopman 1931). (We sometimes denote this map by U_T, sometimes by U, and sometimes even by T.) We can say that T and S are *spectrally isomorphic* if U_T and U_S are unitarily equivalent, in that $VU_T = U_S V$ for some unitary $V: L^2(X) \to L^2(Y)$. Then if S and T are spectrally isomorphic, either both or neither have any one of the recurrence properties mentioned above. An invariant which is sensitive to the nature of the action of T on individual points of X is the *entropy* $h(T)$ of T. The entropy can be used to distinguish some nonisomorphic systems (Kolmogorov and Sinai) and is in fact a complete isomorphism invariant within certain classes of systems (Ornstein).

In many parts of mathematics there is a construction problem as well as a classification problem. Such a question could be formulated in several ways in ergodic theory, one of which is the *realization problem*: which systems of type (3) above can be realized within type (2) (say with T preserving a unique Borel probability measure) or within type (1) (say with T preserving a measure determined by a smooth density)? The first of these questions is discussed below, and the second is the subject of current research. Still another question is that of *genericity*: which types of systems are 'typical', in various senses and in the several different settings?

Many of the important questions of ergodic theory receive scant or no attention here: the existence of invariant measures, the case of infinite measure spaces, operator ergodic theory, the actions of other groups, C^* dynamical systems, etc. And our discussion of the questions that we do treat is not alleged to be complete. The starred references in the bibliography are a good starting place for filling in any gaps that we leave.

1.2. The basic examples

Because the two major sources of ergodic theory are mathematical physics (especially statistical mechanics and Hamiltonian dynamics) and the theory of stationary stochastic processes, naturally these subjects provide a rich store of examples of measure-preserving transformations and flows. There are also several interesting and illustrative classes of abstract examples in an algebraic or geometric context. The following list will give some idea of the kinds of measure-preserving systems we will have in mind during the succeeding discussion. Obviously the various classes are not disjoint, and there are in fact inclusion relations among some of them.

A. Hamiltonian dynamics

The state at any time t of a physical system consisting of N particles can be specified by the three coordinates of position and the three of momentum of each particle, that is by a point in $\mathbb{R}^{6N}$, which is the *phase space* of the system. More generally (allowing for changes of variables and constraints on the system), let the state of the system be described by a pair of vectors (q, p), where $p = (p_1, \ldots, p_n)$ (the 'generalized momentum') and $q = (q_1, \ldots, q_n)$ (the 'generalized position') are in $\mathbb{R}^n$, in which case the phase space is $\mathbb{R}^{2n}$. There is given a ($\mathscr{C}^2$) *Hamiltonian function* $H(q, p)$, which we assume to be independent of time, and which is typically the sum of the kinetic energy $K(p)$ and potential energy $U(q)$ of the system. *Hamilton's equations* are

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \quad (i = 1, 2, \ldots, n).$$

These equations determine the state $T_t(q, p)$ at any time t if the system has initial state (q, p), by the theorem on the existence and uniqueness of solutions of first-order ordinary differential equations. We obtain in this way a one-parameter flow $\{T_t : -\infty < t < \infty\}$ on the phase space $\mathbb{R}^{2n}$.

Theorem 2.1 *Liouville's Theorem* The Hamiltonian flow $\{T_t\}$ preserves Lebesgue measure on $\mathbb{R}^{2n}$.

Sketch of Proof: Consider the vector field

$$V(q, p) = \left(\frac{\partial H}{\partial p_1}, \ldots \frac{\partial H}{\partial p_n}, -\frac{\partial H}{\partial q_1}, \ldots -\frac{\partial H}{\partial q_n}\right),$$

for which clearly div $(V) = 0$. Denoting the Jacobian at (q, p) of the map T_t by $JT_t(q, p)$, this implies through direct calculation that

$$\frac{\partial}{\partial t} JT_t(q, p) = 0.$$

If $E \subset \mathbb{R}^{2n}$ is measurable and μ denotes Lebesgue measure on $\mathbb{R}^{2n}$, then

$$\mu(T_t E) = \int_E JT_t(q, p)\mathrm{d}\mu(q, p).$$

Thus

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu(T_t E) = \int_E \frac{\partial}{\partial t}\left[JT_t(q, p)\right]\mathrm{d}\mu(q, p) = 0.$$

(For the details, see Khintchine 1949 or Plante 1976.)

Hamilton's Equations yield immediately that

$$\frac{\mathrm{d}H}{\mathrm{d}t} = 0.$$

Thus the system is not free to wander all over the phase space but is restricted to surfaces of constant total energy E. Usually most of these surfaces are compact manifolds. The flow restricted to any such surface also has an invariant measure. For the proof of this Proposition see Khintchine (1949).

Proposition 2.2 The Hamiltonian flow restricted to a surface $S = \{(q, p): H(q, p) = E\}$ of constant energy preserves the measure $\mathrm{d}\mu_S = \mathrm{d}S/\| \operatorname{grad} H \|$, where $\mathrm{d}S$ is the element of surface volume.

Through this formulation physical dynamical systems ranging from gas in a container to a cluster of galaxies enter the purview of ergodic theory.

B. Stationary stochastic processes

Let $(\Omega, \mathscr{F}, P)$ be a probability space and $\ldots f_{-1}, f_0, f_1, f_2, \ldots$ a sequence of measurable functions on Ω. Suppose that the sequence is *stationary*, in that for any $n_1, n_2, \ldots n_r$, any Borel subsets $B_1, B_2, \ldots B_r$ of $\mathbb{R}$, and any $k \in \mathbb{Z}$,

$$P\{\omega: f_{n_1}(\omega) \in B_1, \ldots f_{n_r}(\omega) \in B_r\} = P\{\omega: f_{n_1+k}(\omega) \in B_1, \ldots, f_{n_r+k}(\omega) \in B_r\}.$$

Such a stationary process corresponds to a measure-preserving system in a standard way.

Let $\mathbb{R}^{\mathbb{Z}} = \{(\ldots x_{-1}, x_0, x_1, \ldots): \text{each } x_i \in \mathbb{R}\}$, define $\phi: \Omega \to \mathbb{R}^{\mathbb{Z}}$ by $(\phi\omega)_n = f_n(\omega)$ for all $n \in \mathbb{Z}$, and define μ on the Borel subsets of $\mathbb{R}^{\mathbb{Z}}$ by

$$\mu(E) = P(\phi^{-1}E).$$

Extend μ to the completion $\mathscr{B}$ of the Borel field. Let $\sigma: \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ be the

shift transformation defined by

$$(\sigma x)_n = x_{n+1}.$$

Because of the stationarity of $\{f_n\}$, μ is shift-invariant on cylinder sets and hence on all of $\mathscr{B}$, so that we have constructed a measure-preserving system $(\mathbb{R}^{\mathbb{Z}}, \mathscr{B}, \mu, \sigma)$. Moreover, if $\pi_n : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ is the projection onto the nth coordinate ($\pi_n x = x_n$), then $\{\pi_n\}$ has the same joint distributions on $\mathbb{R}^{\mathbb{Z}}$ as $\{f_n\}$ on Ω. Thus every stationary stochastic process 'comes from' some shift-invariant measure on $\mathbb{R}^{\mathbb{Z}}$.

A similar construction applies in the case of a continuous-parameter stochastic process $\{T_t : -\infty < t < \infty\}$ and in more general situations as well (see Doob, 1953).

C. Bernoulli shifts

Let $\boldsymbol{n} = \{0, 1, \dots n-1\}$ be an alphabet of finitely many symbols with weights $p_0, p_1, \dots p_{n-1}$ such that all $p_i > 0$ and $\sum_{i=0}^{n-1} p_i = 1$. Form the product space $\boldsymbol{n}^{\mathbb{Z}}$ of all two-sided sequences of the symbols in $\boldsymbol{n}$, and give $\boldsymbol{n}^{\mathbb{Z}}$ the product measure μ determined by the given probability measure on $\boldsymbol{n}$. Thus for a typical cylinder set determined by a set of places $i_1, \dots i_k \in \mathbb{Z}$ and elements $j_1, \dots j_k \in \boldsymbol{n}$,

$$\mu\{x : x_{i_1} = j_1, \dots x_{i_k} = j_k\} = p_{j_1} p_{j_2}, \dots p_{j_k}.$$

Clearly the shift transformation $\sigma : \boldsymbol{n}^{\mathbb{Z}} \to \boldsymbol{n}^{\mathbb{Z}}$ preserves the measure μ. The resulting measure-preserving system is denoted by $\mathscr{B}(p_0, p_1, \dots p_{n-1})$ and represents a finite-valued stationary stochastic process with independent identically distributed terms (i.i.d.). $\mathscr{B}(\frac{1}{2}, \frac{1}{2})$ models an experimenter tossing a fair coin from the infinite past on into eternity.

D. Markov shifts

Form the product space $\boldsymbol{n}^{\mathbb{Z}}$ and shift transformation as in (*C*). We will define a different invariant measure on $\boldsymbol{n}^{\mathbb{Z}}$, one for which the associated stochastic process is Markov rather than i.i.d.

Let $A = (a_{ij})$ be an $n \times n$ stochastic matrix, i.e. a matrix with nonnegative entries and each row sum equal to 1. Suppose also that $p = (p_0, p_1, \dots p_{n-1})$ is a row probability vector (all $p_i \geqslant 0$, $\Sigma p_i = 1$) which is fixed by A:

$$pA = p.$$

By the Perron–Frobenius Theorem (see Varga 1962) such a vector can always be found, and in some cases it is unique. Define the measure of a cylinder set determined by consecutive indices by

$$\mu_A\{x : x_i = j_0, x_{i+1} = j_1, \dots x_{i+k} = j_k\} = p_{j_0} a_{j_0 j_1} a_{j_1 j_2} \dots a_{j_{k-1} j_k}.$$

(Thus p gives the a priori probabilities of the symbols and A the transition probabilities from one symbol to another.) It can be verified that μ_A extends in a well-defined way to a countably additive measure on the algebra generated by the cylinder sets, and hence, by the Carathéodory–Hopf Theorem, μ_A extends to the Borel field of $n^{\mathbb{Z}}$ and its completion $\mathscr{B}$. The resulting measure-preserving system $(n^{\mathbb{Z}}, \mathscr{B}, \mu_A, \sigma)$ models a finite-state Markov chain.

E. Rotations of the circle

From the point of view of ergodic theory, the unit circle $\mathbb{K} = \{z \in \mathbb{C} : |z| = 1\}$ is the same as the unit interval $[0, 1)$, and both are versions of $\mathbb{R}/\mathbb{Z}$, the reals 'mod 1'. Given an $\alpha \in \mathbb{R}$, we consider the map $T_\alpha : [0, 1) \to [0, 1)$ defined by

$$T_\alpha x = x + \alpha \,(\text{mod } 1) = \langle x + \alpha \rangle = \text{fractional part of } x + \alpha$$
$$= x + \alpha - [x + \alpha].$$

Regarded as a map of $\mathbb{K}$,

$$T_\alpha e^{2\pi i\theta} = e^{2\pi i(\theta + \alpha)}.$$

It is clear that T_α preserves Lebesgue measure. If α is rational, then T_α is periodic, all orbits being finite and of the same cardinality. Thus T_α is most interesting when α is irrational.

F. Rotations of compact abelian groups

Let G be a compact abelian group and $g_0 \in G$. Define $T_{g_0} : G \to G$ by $T_{g_0} g = g + g_0$. Because Haar measure is translation-invariant, T_{g_0} is a m.p.t. Again the most interesting case is when G is monothetic and g_0 is a generator: $\{ng_0 : n \in \mathbb{Z}\}$ is dense in G.

G. Automorphisms of compact groups

Let G be a compact group and $T : G \to G$ a continuous automorphism. The uniqueness of normalized Haar measure implies that it is T-invariant.

A particular case of interest is when $G = \mathbb{K}^n = \mathbb{R}^n/\mathbb{Z}^n$ is the n-torus, when it can be shown that T is given by an $n \times n$ integer matrix with determinant ± 1.

H. Gaussian systems

Consider a stochastic process $\ldots f_{-1}, f_0, f_1, \ldots$ on a probability space $(\Omega, \mathscr{F}, P)$ which is *Gaussian* in that the joint distribution of any finite number of the f_i is Gaussian: given $i_1 < i_2 < \ldots < i_n$, there are $m_1, \ldots, m_n \in \mathbb{R}$ and a symmetric positive-definite $n \times n$ matrix $A = (A_{ij})$

such that for each Borel $E \subset \mathbb{R}^n$,

$$P\{\omega:(f_{i_1}(\omega),f_{i_2}(\omega),\dots,f_{i_n}(\omega))\in E\}$$
$$=\frac{1}{2\pi^{n/2}\sqrt{\det A}}\int_E \exp\left[-\tfrac{1}{2}(x-m)^{tr}A^{-1}(x-m)\right]dx_1\dots dx_n.$$

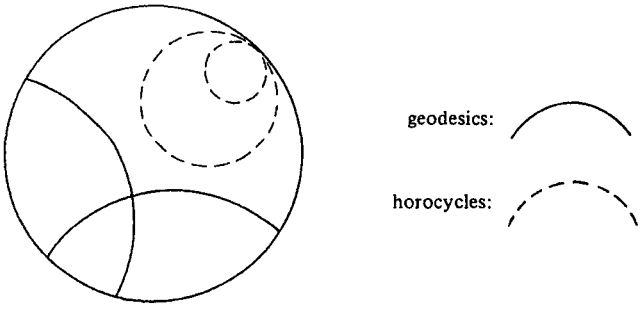
If $\int_\Omega f_i dp$ is a constant, m_0, and $A_{ij}=\int_\Omega (f_i-m_i)(f_j-m_j)dP$ depends only on $i-j$, then the process is stationary, and as in (*B*) determines a measure-preserving system (see Totoki 1970).

I. Geodesic flows

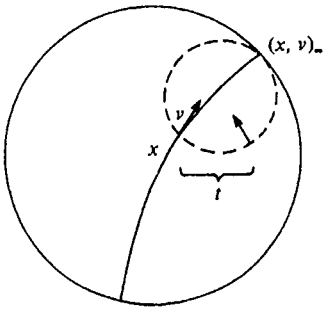
Let M be a compact Riemannian manifold and $UT(M)$ the unit tangent bundle of M, i.e. the collection of all (x, v), where $x\in M$ and v is a unit tangent vector to M at x. The geodesic flow $\{T_t\}$ on $UT(M)$ is defined as follows. Given (x, v), find the geodesic $\gamma(t)$ which at time 0 passes through x and is tangent to v. Flow along the geodesic at unit speed for a time t, and take for $T_t(x, v)$ the point and unit tangent vector that you finish with. The geodesic flow preserves the measure on the manifold that is determined by the Riemannian metric (see Gottschalk–Hedlund 1955).

J. Horocycle flows

Let M be a compact oriented surface of constant negative curvature. Then M is a quotient of the Poincaré disk by a discrete subgroup of isometries. Geodesics in the Poincaré disk are circular arcs that are perpendicular to the boundary circle:



By a *horocycle* we mean a circle which is interior to the disk (except for the point of tangency) and tangent to the boundary. The *horocycle flow* on $UT(M)$ is defined as follows. Given $(x, v)\in UT(M)$, find the geodesic through x in the direction of v. Find the point $(x, v)_\infty$



at which this geodesic intersects the boundary. Construct the horocycle tangent to this point, through x, and orthogonal to v. Flow along this horocycle (in a pre-selected sense) at unit speed for a time t, and take for $T_t(x, v)$ the equivalence class of the point and unit normal vector that you finish with. Again the natural volume is preserved.

It is important to note that the geodesic and horocycle flows on the full Poincaré disk have no recurrence whatsoever, but even strong mixing does arise when we pass to a compact quotient by a discrete group of isometries (see Gottschalk and Hedlund 1955).

K. Flows and automorphisms on homogeneous spaces

Let G be a unimodular Lie group, Γ a discrete subgroup such that G/Γ has finite volume (as determined by the Haar measure on G), and $\{g_t : t \in \mathbb{R}\}$ a one-parameter subgroup of G. Then $\{g_t\}$ determines a volume-preserving flow by left multiplication on the set of right cosets. In fact, the classical geodesic and horocycle flows arise in this way (see Auslander, Green and Hahn, 1963 and Brezin and Moore, 1981).

Alternatively, let $T: G \to G$ be a continuous automorphism such that $T\Gamma = \Gamma$. Then the map induced on G/Γ by T is a m.p.t.

By combining translations and automorphisms, one can produce affine maps on homogeneous spaces (see Parry 1969*c*, 1971).

1.3. The basic constructions

In any subject it is valuable to have ways to modify or combine old objects to make new ones. In this section we list some of the techniques that are available in ergodic theory.

A. Factors

Let $(X, \mathcal{B}, \mu, T)$ and $(Y, \mathcal{C}, \nu, S)$ be measure-preserving systems