Ergodic Theory

Jacob Denson

February 16, 2023

Chapter 1

Origins of Ergodic Theory

1.1 Poincare and Hamiltonian Dynamics

Ergodic theory's story begins in the 17th century, as does most of analysis, with the invention of Newton's calculus. With it began the 'quantitative age' of mathematics, which focused on techniques of integration and of the exact solution of differential equations arising in physics. But in the late 19th century, the quantitative techniques hit an impasse. It was shown that the differential equation defining the motion of three or more planets could not be solved numerically - a fact analogous to Galois' proof that one could not formulate general solutions to the quintic equation. Such a crisis could only be solved by one of the greats, and it is with a particular great, the Frenchman Henri Poincaré, that the theory of dynamics was saved, and ergodic theory began to take its form.

The classical method of looking at a differential equation is algebraic. A physical process leads us to an equation

$$\frac{d^2y}{dt^2} = t^2\frac{dy}{dt} - ty$$

the game is then to apply certain solution techniques to reduce the formula to a simpler statement, with the main goal of obtaining an expression for y as a function of t, unique after obtaining certain initial conditions, for instance, in this case, the position, velocity, and time at an initial point. A revolutionary view of differential equations emerges when we see a differential equation instead as an *operator* on the space of initial conditions of the equation, transforming the conditions into conditions in the future.

Poincare's methods are characterized by looking at a solution from a global point of view, rather than a local point of view. Instead of finding a solution for a particular initial condition, we try and find the action of the differential equation on the space of all initial conditions – solving all equations simultaneously. It is important to note that the initial conditions must describe the infinitisimal changes in the individual variables at present. For instance, in this equation we know that the change in position (velocity), the change in velocity ($t^2\dot{x}_0 - tx$), and the change in time (constant, since time passes uniformly). Therefore, a differential equation gives us a vector field on the 'phase space' of conditions, in this case

$$v(t, x, \dot{x}) = (1, \dot{x}, t^2\dot{x} - tx)$$

The theory of uniqueness and existence for differential equations tells us that, for each $(t, x, \dot{x})$, there is a unique function $g(t)$ solving the differential equation in a certain time interval. Such a function gives us a particular curve in phase space, taking the conditions of the function at a particular time point. By uniqueness, we can put all these curves together to obtain an 'evolution' function f, which takes a certain initial condition x_0, and a certain number t, and gives us the position $f_t(x_0)$ of the curve in phase space, t seconds after it begins to move. With this switch, it is much easier to describe 'qualitative facts about the differential equation', since we have now boxed all solutions in a single object to study.

Example. *It's best to see this method in action. Consider the spring equation*

$$y'' = -y$$

The phase-space for this equation is $\mathbf{R}^2$ *(one dimension for the position of the end of the spring, and one dimension for the velocity). We obtain the vector field*

$$v(x, \dot{x}) = (\dot{x}, -x)$$

If you draw out this vector field, it is easy to see that the action f_t *is just to rotate phase space clockwise by an angle t. Qualitatively, this tells us why the motion of the spring is periodic, and why the value*

$$x^2 + \dot{x}^2$$

is preserved in a particular trajectory. The beauty of this approach is that we need no formula to see why this is true, we just look at the qualitative effects of the vector field.

Example. *If a particle is pushed from left to right on the number line at a uniform velocity, then the differential equation is*

$$y' = C$$

with corresponding vector field $v(x) = C$, and whose propogation operators are translations $f_t(x) = x + Ct$.

Ergodic theory is especially concerned with the long-term qualitative effects of dynamical systems. That is, properties of the limiting values of $f_t(x)$ as $t \to \infty$. In this case, we rarely need t to take on all real values. We really need only consider a sequence $f_0(x), f_1(x), f_2(x), \ldots$. It only takes a little bit of thinking to convince yourself that $f_t \circ f_s = f_{t+s}$ (if we wait s seconds, and then wait t seconds, we are waiting $t + s$ seconds). In this case this means that $f_n = f_1^n$, so we can consider Ergodic theory as the study of a single map $T : X \to X$, and it's iterates $T, T^2, T^3, \ldots$. The real fun begins when T is invertible, so we can define T^n for all integers n, and so the map T possesses 'time symmetry'.

The historical development of the subject tells us which maps we study, as Poincare applied his new view to the study of differential equations in physics. The standard equation of motion in a Newtonian System is $m\ddot{x} = F(x)$. In 1833, the scientist William Hamilton, discovered that by a change in coordinates one could discover a much more beautiful representation of classical mechanics. If $p = m\dot{x}$ denotes the momentum of a particle at a certain time, and we rename the position coordinate x to q, then we may express the kinetic energy of the particle as

$$\frac{m\dot{q}^2}{2} = \frac{p^2}{2m}$$

Often, we find a scalar function $V(q)$ such that $\nabla V(q) = -F(q)$; such a function is known as the potential energy of the system. With these definitions in hand, we define the total energy of the system, known as the Hamiltonian, as

$$H(p,q) = \frac{p^2}{2m} + V(q)$$

H is known as the Hamiltonian of a physical system. Now in this form, Newton's laws take the pleasant form

$$\dot{p} = -\frac{\partial H}{\partial q} \qquad \dot{q} = \frac{\partial H}{\partial p}$$

Note that this *isn't* a partial differential equation to be solved, since H is a known quantity, and we are solving for p and q. The main reason to apply Hamilton's equations is that the approach generalizes much more simply to arbitrary coordinate systems, and it is often much easier to define H in terms of the energy of a particular configuration, rather than figuring out the forces between objects (try defining the forces for a system of pulleys attached to one another, and you'll get what I mean).

Now suppose we only know the initial position and momentum of an particle to a certain precision – then the initial conditions of our object in phase space lie in a certain area U. If we watch U evolve, we learn that the object must eventually lie in $f_t(U)$, for each t. But how does the precision of our measurements change over time? $f_t(U)$ might shrink in size, so that our measurements become more accurate, or increase in size, so they grow inaccurate.

Lemma 1.1. *If a vector field v has divergence zero,*

$$div\ v = \sum \frac{\partial v^i}{\partial x^i} = 0$$

Then the operators f_t preserve volume.

Proof. Let D be a region of space, define $w(0)$ to be the volume of D, and more generally, define $w(t)$ to be the volume of $f_t(D)$. Now f_t is a diffeomorphism of D, so by the change of variables formula

$$w(t) = \int_{f_t(D)} 1 = \int_D |\det(Df_t)|$$

and hence, provided that $(t,x) \mapsto f_t(x)$ is continuously differentiable (which occurs when H is twice continuously differentiable),

$$w'(t) = \int_D \frac{d|\mathrm{Det}(Df_t)|}{dt}$$

Now the Hamiltonian equation tells us exactly that

$$\frac{df_t(x)}{dt} = v(x)$$

This implies

$$\frac{d(Df_t)}{dt} = v(x)$$

If $M(t)$ is a family of matrices, with $M(0) = I$ and $M'(0)_{ij} = a_{ij}$, then

$$\frac{d\det(M(t))}{dt} = \mathrm{tr}(M)$$

because the determinant is a polynomial equation in the entries in the matrix, one for each row and each column, and for each monomial in the expansion, either the monomial is $M_{11}(t)M_{22}(t)\dots M_{nn}(t)$, or the monmial contains two coefficients off the diagonal. The product rule tells us that the derivative of the latter type of monomial vanishes, and the initial monomial has derivative $\mathrm{tr}(M)$. Hence

$$\frac{d\det(Df_t)}{dt} = \mathrm{tr}(Df_t) = \mathrm{div}(v) = 0$$

so it follows that $w'(t) = \int \mathrm{div}(v)$. □

Louiville observed that the vector field governing the motion in Hamiltonian mechanics has zero divergence (mixed partials are equal), which means volume is preserved by motion in momentum-position space under the action of the Hamiltonian equations. Thus we have a kind of 'Heisenberg uncertainty principle' in classical mechanics. We can never infer a more accurate measurement from an initial measurement, without losing accuracy in other measurements in the process – the 'area' of the measurements must remain the same.

Now the great part about measure preserving transformations is that they cannot 'stretch' or 'squish' any part of space, since the volume of that portion of space must stay the same. In spaces with only finite volume, this gives us something to work with. The first theorem taking advantage of this is a continuous form of the pidgeonhole principle.

Theorem 1.2. *Let $T : X \to X$ be a measure preserving map on a space of finite volume. Given a fixed $U \subset X$ of positive volume, almost every point of U returns to U after some time.*

Proof. Let $A = \{x \in U : x \text{ never returns to } U\}$. Then $T^{-n}(A)$ is disjoint from $T^{-m}(A)$, for each $n \neq m$, for if x arrives in A in both n iterations and m iterations, for $n < m$, then $T^n x$ arrives in A after $m - n$ iterations, a contradiction to the fact that $T^n x \in A$. Thus $v(A) = 0$, for otherwise

$$v(X) \geqslant v\left(\bigcup_{i=0}^{\infty} T^{-i}(U)\right) = \sum_{i=0}^{\infty} v(T^{-i}(U)) = \sum_{i=0}^{\infty} v(U) > \infty$$

So almost every point in U returns to U once. □

Example. *Consider the map on the Torus* $\mathbf{R}^2/\mathbf{Z}^2$ *defined by*

$$T(x,y) = (2x+y, x+y) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

T *is actually an invertible map, with inverse*

$$T^{-1}(x,y) = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

and is area preserving, since the determinant of the matrix is 1. Thus we can apply Poincare recurrence to see that almost every point returns to a neighbourhood of itself. T *is known as the cat map, since the map's properties were originally demonstrated on the image of a cat, which is torn apart, and then appears to return to it's original form hundreds of iterations later.*

In the case that T is injective, we can obtain explicit bounds on the time we will have to wait for a point in U to return to U. If $U, TU, \ldots T^k U$ are disjoint, then

$$v\left(\bigcup_{i=0}^{k} T^i U\right) = nv(U)$$

If $T^i U$ and $T^j U$ intersect, $T^{-j}T^i U = T^{i-j}U$ intersects U. Thus we find that, to avoid contradiction, some point in U must return to U in $\lceil v(X)/v(U) \rceil$ iterations.

Example. *Consider opening the lid on a bucket of air molecules in a sealed chamber. Poincare's theorem tells us that, paradoxically, to any desired margin of error, the air molecules will eventually return to their original position inside the bucket. Such a paradox is solved when we consider the 'curse of dimensionality' for the particles. Suppose we only know the initial positions of the particles to within an error of* $\pm\varepsilon$*. Then the 'volume of error' in phase space will be* $(2\varepsilon)^{3n}$*, and if the sealed chamber is a cube of length m, then the bound on the return time is on the order of*

$$\left(\frac{m}{2\varepsilon}\right)^{3n}$$

which grows exponentially as the number of particles increase. Thus to wait for 200,000 particles to return to their original position (a large underestimate on the number of particles, if we have a whole bucketful), we will likely have to wait until the end of the universe.

Example. *Consider the circle S^1, which has a canonical 'angular measure' placed on it. Let $T : z \mapsto wz$, for a fixed $w \in S^1$. Then T is a rotation of the circle, which preserves angular measure measure. If $w^n z = z$, for some x, then $w^n = 1$, and is a root of unity. In this case the system of translations $T, T^2, \ldots$, is cyclic, and points return to their initial positions exactly after a certain amount of time. If w is not a root of unity, then Poincare's recurrence theorem tells us something interesting – for any $\varepsilon > 0$, there is n such that $|w^n - 1| < \varepsilon$. Take the neighbourhood U around 1 which sweeps out an angle of $\varepsilon/2$ in each direction. Poincare's theorem tells us that there is $z \in U$ and n for which $w^n z \in U$, so $|w^n z - z| = |w^n - 1| < \varepsilon$. Since the group S^1 is isomorphic to $\mathbf{R}/\mathbf{Z}$, Poincare's theorem has interest arithmetically – if x is an irrational number, then we may always choose n such that the decimal values of nx are as close to zero as desired; a very elegant argument for what is normally a very messy proof.*

Now measure preserving maps cannot 'squish and hide' parts of their domain, but they can still keep parts of the domain too still.

Example. *Consider the numbers*

$$1, 2, 4, 8, 16, 32, \ldots, 2^k, \ldots$$

in particular, take the most significant digits of these numbers.

$$1, 2, 4, 8, 1, 3, 6, 1, 2, 5, \ldots$$

Do these digits follow some sort of regular pattern? Perhaps, but this pattern is certainly not obvious. Ergodic theory is not good at discovering 'pointwise' patterns, but can help us discover the behaviour of phenomena 'on average', or 'in the limit'. The trick to discovering the properties of the power sequence is to look at the numbers in scientific notation

$$1 \cdot 10^0, 2 \cdot 10^0, 4 \cdot 10^0, 8 \cdot 10^0, 1.6 \cdot 10^1, 3.2 \cdot 10^1, \ldots$$

In general, we may extract the first digit of a number of the form $z = x \cdot 10^y$, where $0 \leqslant x < 10$ in the following manner. We have

$$\log(z) = y + \log(x)$$

Since log *is an increasing function, the first digit of x is 1, if and only if $1 \leqslant x < 2$, so $\log(1) \leqslant x < \log(2)$. Similarily, the first digit of x is n if $\log(n) \leqslant x <$*

$\log(n+1)$. *Now the act of taking a power of* 2 *may stretch the area of the shape, but if we define an area function on* $[0,1]$ *by*

$$v([a,b]) = \log(b) - \log(a)$$

then we see that doubling preserves the area. The recurrence theorem then shows that every number from 1 *to* 9 *occur as the first digit of some power of 2. More advanced theorems of ergodic theory tell us that the powers of two are logarithmically 'equally distributed' on* $[0,1]$ *according to our new area measure: the probablity that a randomly selected power of 2 has first digit n is the same the logarithm lies between* $\log(n)$ *and* $\log(n+1)$*, so the probability is* $\log(n+1)-\log(n)$.

As a final example of our system, consider hitting a ball on a frictionless pool table. How will the angle of the shot impact the orbit of the ball? The recurrence theorem tells us that, if the slope of our shot is irrational, the points of contact of the ball on the side of the walls of the table are dense. Some more advanced ergodic theory tells us that the points of contact are equidistributed along the wall.

1.2 Exponential Sums

s

Chapter 2

The Ergodic Theorem

We now outline precisely the objects studied by Ergodic theory. We have discovered that maps which preserve the volume of a dynamical system have useful properties which deserve being explored in detail. Take a space X with measure μ. The maps of interest here are maps $T : X \to X$ which *preserve measure*, in the sense that $\mu(U) = \mu(T^{-1}U)$. These maps are called **measure-preserving transformations**. They need not be invertible, but if they are, and if the inverse of T is measure preserving, then we call T a **measure-preserving isomorphism**. The tuple (X, μ, T) is known as a measure-preserving system.

The first property of a measure-preserving transformation is that it preserves any meaningful 'measurement'. In physics, we can normally only measure some meaningful quantity of a system by averaging out a value over some measurement. A 'value' of a point x in a measure space X is the value of x under a measurable map $f : X \to \mathbf{C}$, and an average of the map under some region E is the integral

$$\frac{1}{\mu(E)} \int_E f(x) d\mu(x)$$

which we can write as

$$\int_X \frac{\chi_E(x)}{\mu(E)} f(x) d\mu(x)$$

Theorem 2.1. *If $T : X \to X$ is measure preserving, and $f : X \to \mathbf{C}$, then for any measurable E,*

$$\int_X f d\mu = \int_X (f \circ T) d\mu$$

Proof. If $f = \chi_E$ for some E, then $f \circ T = \chi_{T^{-1}(E)}$, and so

$$\int_X f d\mu = \mu(E) = \mu(T^{-1}(E)) = \int_X (f \circ T) d\mu$$

By linearity, the theorem also holds if f is a simple function. But then the theorem can be easily extended to positive functions by monotone convergence, and thus to all functions by breaking them into positive parts. □

Example. *If G is a locally compact group, then G possesses a left Haar measure μ. If $T : G \to G$ is given by left multiplication $x \mapsto yx$, for some fixed $y \in G$. Then $\mu(yU) = \mu(U)$, so T is a measure preserving transformation. This encapsulates many of the important examples of ergodic theory, such as shifts $n \mapsto n + m$ and $v \mapsto v + w$ on $\mathbf{Z}$ and $\mathbf{C}^n$.*

Example. *If $T : G \to G$ is a surjective group homomorphism on a locally compact group, then we may define a new measure ν on G by defining $\nu(U) = \mu(T^{-1}(U))$. Then one verifies that $\nu \neq 0$, and*

$$\nu(yU) = \mu(T^{-1}(yU)) = \mu(yT^{-1}U) = \mu(T^{-1}U) = \nu(U)$$

so ν must be a constant multiple of μ. Since $\nu(U) = \mu(T^{-1}(U)) = \mu(U)$, we must have $\nu = \mu$. This shows that T is measure preserving. This shows that the map $t \mapsto 2t$ on $\mathbf{R}/\mathbf{Z}$ is measure preserving, though it isn't injective.

Example. *If $X = [0, 1)$, and $T(x)$ is the fractional part of $1/x$ (the fraction part of a number is its image of the map obtained from the quotient map from $\mathbf{R}$ to $\mathbf{R}/\mathbf{Z}$, where we identify an equivalence class with the unique point in $[0, 1)$ representing it), then T preserves the measure $\frac{dx}{1+x}$. This is important to the theory of continued fractions in number theory.*

Example. *Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables, and consider the probability distribution induced on the countable product Ω^ω. Then the map $T(w_1, w_2, \dots) = (w_2, w_3, \dots)$ is measure preserving. This puts the theory of independant variables in the context of ergodic theory, and we will see the laws of large numbers is a special case of ergodic theory, that the averages*

$$\frac{1}{n} \sum_{k=0}^{n} X_k = \frac{1}{n} \sum_{k=0}^{n} f \circ T^k$$

converge almost everywhere to the mean of the X_k.

Example. *Consider a Markov chain on a set of states X induced by some stochastic operator P. Given any initial probability distribution μ. Then $P^n\mu$ is a probability distribution which gives the distribution 'n steps into the future'. Then we have an induced product measure on X^ω given by the infinite product $\left(\times_{n=0}^{\infty} P^n\mu\right)$, and the operator $T(w_1, w_2, \dots) = (w_2, w_3, \dots)$ is measure preserving if and only if μ is an invariant measure – that is, $P\mu = \mu$.*

Poincare showed that the trajectories of measure preserving systems almost always reoccur, but in statistical physics, more is desired. When we want to find a quantity of a system, we take values related to the averages

$$A_n(f)(x) = \frac{1}{n}\sum_{m=0}^{n-1} f(T^m x)$$

In the theory of probability, the strong law of large numbers tells us these averages almost always converge to the true average of the system. But in physics the systems are deterministic, so something is stronger is needed to guarantee the converges of the averages. The ergodic theorems characterize this study.

The first such theorem rises directly from the results of Hilbert space theory. A measure preserving system $T : X \to X$ induces an operator $S : f \mapsto f \circ T$ on $L^2(X)$, which satisfies $\|Sf\|_2 = \|f\|_2$. If T is an isomorphism, then S is in fact a unitary operator (it is certainly injective, but may not be surjective). Let $V \subset L^2(X)$ be the kernel of S, and let $P : L^2(X) \to V$ be the continuous projection onto the subspace of V.

Theorem 2.2. *The averages*

$$\frac{1}{n}\sum_{m=1}^{n-1} T^m$$

of a Hilbert space isometry converges weakly to the projection onto the set of vectors v such that $Tv = v$.

Proof. Set

$$V = \{v \in H : Tv = v\} \qquad W = \{w \in H : T^*w = w\}$$

These subspaces are closed, since they are the kernels of the operators $(T-1)$ and (T^*-1). Since T is an isometry, the polarization identity tells us that $\langle T^*Tv, v\rangle = \langle v, v\rangle$ for all v, and so that $T^*T = 1$. Thus if $T^*v = v$,

$$\langle Tv, v\rangle = \langle v, T^*v\rangle = \|v\|^2$$

This is a case of equality in the Cauchy-Schwartz inequality, since $\|Tv\| = \|v\|$, so $Tv = \lambda v$, and the above relation gives us $\lambda = 1$. Conversely, if $Tv = v$, then because $T^*T = 1$, we obtain $v = T^*v$. Thus $V = W$.

Now take $K = \{v - Tv : v \in H\}$. Unlike V and W, K is not necessarily closed. However, note that if

$$\langle w, v - Tv \rangle = \langle Tw, (1 - T)v \rangle = 0$$

for all $v \in H$, then $\langle w - T^*w, v \rangle = 0$ for all $v \in H$, hence $w = T^*w$. Thus $V = W$ is the orthogonal complement of $\overline{K}$.

Now given any $v \in H$, decompose $v = v_0 + v_1$, with $v_0 \in V$, $v_1 \in \overline{K}$. Then

$$A_n(f) = \frac{1}{n} \sum_{m=0}^{n-1} T^m v = v_0 + \frac{1}{n} \sum_{m=0}^{n-1} T^m v_1$$

Fix $\varepsilon > 0$, and find w such that $\|v_1 - (w - Tw)\| < \varepsilon$. Then

$$\begin{aligned} \|A_n(v_1)\| &= \|A_n(v_1 - (w - Tw))\| + \|A_n(w - Tw)\| \\ &= \varepsilon + \left\| \frac{1}{n} \sum_{m=0}^{n-1} T^m (w - Tw) \right\| \\ &= \varepsilon + \frac{1}{n} \|w - T^m w\| \\ &\leqslant \varepsilon + \frac{2\|w\|}{n} \end{aligned}$$

For n large enough, we find $\|A_n(f_1)\| \leqslant 2\varepsilon$, and thus $A_n(f) \to f_0$. □

We apply this theorem to any measure preserving system (X, μ, T) and we conclude that for any $f \in L^2(X)$, the functions

$$A_n(f) = \frac{1}{n} \sum_{m=0}^{n-1} f \circ T^m$$

converge on average to a map g for which $g \circ T = g$, where $f - g$ can be approximated in $L^2(X)$ by functions of the form $h - h \circ T$.

The mean ergodic theorem is a nice result, but it is analogous to the weak law of large numbers, and like with the weak law, we can do much better. We wish to show that we have pointwise convergence almost everywhere, which implies the

mean law, but is a stronger result. The standard approach to establishing pointwise convergence from average convergence is found by finding estimates to a corresponding 'maximal functions', which bounds the pointwise convergence at each point. The maximal function in this case is the ergodic function

$$f^*(x) = \sup_{1 \leqslant m \leqslant \infty} \frac{1}{m} \sum_{k=0}^{m-1} |f(T^k x)|$$

which is certainly measurable for any measurable f.

Theorem 2.3 (The Maximal Ergodic Theorem)**.** *If $f \in L^1(X)$, then f^* is finite for almost all x, and there is a universal constant such that*

$$\mu(\{x : f^*(x) > \alpha\}) \leqslant \frac{A}{\alpha} \|f\|_1$$

Proof. To be continued... □

Pointwise convergence is much easier to obtain if we assume that the measure space is finite, and we can always normalize so that we are working over a probability space.

Theorem 2.4 (Pointwise Ergodic Theorem)**.** *Suppose that (X, μ, T) is a measure preserving system in a probability space. If $f \in L^1(X)$, then the averages $A_m(f)$ converge pointwise to a limiting function g.*

Proof. Since $\mu(X) = 1$, $L^2(X) \subset L^1(X)$, and $L^2(X)$ is dense in $L^1(X)$. For any $\varepsilon > 0$, write $f = f' + h$, where $f' \in L^2(X)$, and $\|h\|_1 < \varepsilon$. As in the mean ergodic theorem, we can write $f' = f_0 + (1 - S)f_1 + h$, where $Tf_0 = f_0$, and $\|h\|_1 \leqslant \|h\|_2 < \varepsilon$. We conclude that for every ε, a function $f \in L^2(X)$ can be written $f_0 + (1 - S)f_1 + h$, where $f_0, f_1 \in L^2(X)$, $Sf_0 = f_0$, and $\|h\|_1 \leqslant \varepsilon$. Then

$$A_n(f) = f_0 + \frac{1}{n} - \frac{S^n(f_1)}{n} + A_n(h)$$

For almost any x,

$$\frac{(S^n f_1)(x)}{n} = \frac{(f_1 \circ T^n)(x)}{n} \to 0$$

because the series

$$g(x) = \sum_{n=1}^{\infty} \frac{|(f_1 \circ T^n)(x)|^2(x)}{n^2}$$

converges for almost all x – by the monotone convergence theorem,

$$\int_X g d\mu = \sum_{m=1}^{\infty} \frac{1}{m^2} \|f_1 \circ T^m\|_2^2 = \|f_1\|_2^2 \sum_{m=1}^{\infty} \frac{1}{m^2} < \infty$$

Thus $A_n(f_0 + (1-S)f_1)$ converges for almost all x. To establish that f converges, set

$$E_\alpha = \{x : \lim_{N\to\infty} \sup_{n,m\geqslant N} |A_n(f)(x) - A_m(f)(x)| > \alpha\}$$

It suffices to prove that $\mu(E_\alpha) = 0$ for each $\alpha > 0$. If we define

$$E'_\alpha = \{x : \lim_{N\to\infty} \sup_{n,m\geqslant N} |A_n(h)(x) - A_m(h)(x)| > \varepsilon\}$$

Then we see that almost every point of E_α is in E'_α, and thus

$$\mu(E_\alpha) \leqslant \mu(E'_\alpha) \leqslant \mu(\{x : \lim_{N\to\infty} \sup_{n\geqslant N} 2|A_n(h)(x)| > \alpha\})$$

We can now apply the maximal ergodic thoerem to conclude the value on the left is bounded by

$$\frac{2A}{\alpha}\|h\|_1 \leqslant \frac{2A\varepsilon}{\alpha}$$

since ε was arbitrary, $\mu(E_\alpha) = 0$, and the theorem is proved. □

If $f \in L^2(X)$, then $A_n(f) \to g$ in $L^2(X)$. If we take a subsequence which converges almost everywhere, we obtain that g must be the function found in the pointwise ergodic theorem.

Chapter 3

Translation Invariant Operators

There is a close connection between the ergodic theorem and certain facts about operators on the real line. Let X be a σ-finite measure space, and h_t a one parameter group of measure preserving transformation of X. We assume the map $(t,x) \mapsto f(h_t(x))$ is measurable in the product space $\mathbf{R} \times X$. Let T be a sublinear map from the space $L^1_{\mathrm{loc}}(\mathbf{R})$ of locally integrable functions on the real line to $C(X)$, which commutes with all translation operators, and for which these is a universal $\varepsilon > 0$ for which the support of Tf is in an ε neighbourhood of the support of f. We shall associate an operator $T^{\#}$, now defined on X-measurable functions. Let $f : X \to \mathbf{R}$ be a measurable map, such that $F_x(t) = f(h_t(x))$ is also measurable on $X \times \mathbf{R}$. Then F_x is a measurable function of t for almost all t, and is in fact locally integrable for almost all x, since by Fubini's theorem,

$$\int_X \int_A F_x(t)dt\, dx = \int_A \int_X f(h_t(x))dx\, dt = \int_A \int_X f(x)dx\, dt = \mu(A) \int_X f(x)dx$$

It follows that the function $g(x) = T(F_x)$ is well defined for almost all x, and we define $T^{\#}f = g$.

Theorem 3.1. *If T_n is a sequence of operators defined as above, and the operator $Sf = \sup |T_n f|$ is bounded as a map with L_p norm, for $1 \leqslant p \leqslant \infty$. Then the same holds for the operator $S^{\#}f = \sup |T_n^{\#} f|$, and $\|S^{\#}\| < \|S\|$.*

Proof. Suppose first of all that the sequence of operators is a finite set $T_1, T_2, \ldots, T_n$. Then S is an operator with the same properties that T possesses. We note that $F_{h_s(x)}(t) = F_x(t+s)$, which means that F_x is always integrable, and $\int F_x(t)dt = \int F_x(t')dx$ □

Lemma 3.2. *Let $T_n f = k_n * f$, where k_n is a bounded function with bounded support. If $S f = \sup |T_n f|$ has weak type in L_p, $\int k_n$ converges, and $k_n * \phi$ converges in L_1 for each infinitely differentiable ϕ with compact support and vanishing integral. Then $T_n^{\#} f$ converges almost everywhere in X for each f in $L^p(X)$.*

Proof. s □

Chapter 4

Equidistribution of Sequences

In this chapter, we discuss the distribution of the decimal parts of a sequence of integers. Given such a sequence $x_1, x_2, \ldots$, we can consider the images $\langle x_1 \rangle, \langle x_2 \rangle, \ldots$ in $\mathbf{R}/\mathbf{Z}$ in $\mathbf{R}/\mathbf{Z}$. We will identify $\mathbf{R}/\mathbf{Z}$ with it's fundamental domain $[0, 1)$, which is precisely the correspondence that enables us to discuss the distribution of the decimals. When discussing elements of $\mathbf{R}/\mathbf{Z}$, we let $[a, b]$ denote the set of all $\langle x \rangle$, for $x \in [a, b]$. If we have a sequence $x_1, x_2, \ldots$, we say it is equidistributed if for any interval $[a, b] \subset \mathbf{R}/\mathbf{Z}$,

$$\frac{\{1 \leqslant n \leqslant N : \langle x_n \rangle \in [a, b]\}}{N} \to b - a$$

Thus the number of points in $[a, b]$ is, in the limit, proportional to the length $[a, b]$. The observation that

$$\frac{\{1 \leqslant n \leqslant N : x_n \in [a, b]\}}{N} = \frac{1}{N} \sum_{n=1}^{N} \chi_{[a,b]}(x)$$

leads to an important, alternate classification of equidistribution.

Theorem 4.1. *The sequence $x_1, x_2, \ldots$ is uniformly distributed if and only if for every Riemann integrable (or continuous) 1-periodic function f,*

$$\frac{1}{N} \sum f(x) \to \int_0^1 f(x)\, dx$$

Proof. Any continuous function on $\mathbf{R}/\mathbf{Z}$ can be uniformly approximated by simple functions (sums of characteristic functions of intervals), and conversely, the characteristic function of an interval can be uniformly approximated by continuous functions. □

Lemma 4.2. *If $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ are sequences such that $\alpha = \lim y_n - x_n$ exists, and $\{x_n\}$ is equidistributed, then $\{y_n\}$ is also equidistrbuted.*

Proof. TODO □

Given an arbitrary partition $\Delta = 0 = t_0 < t_1 < \cdots \to \infty$, for $t_n \leqslant x < t_{n+1}$, we let $[x] = t_n$, and $\langle x \rangle = (x - t_n)/(t_{n+1} - t_n)$. Then $\langle x \rangle$ lies in $[0, 1)$, and we can consider the distributions in this domain, leading to the definition that x_n is uniformly distributed relative to Δ.

4.1 Speed of Convergence

Given a finite sequence $x_1, \ldots, x_N \in \mathbf{T}$, we let the **discrepancy** of the sequence

$$D(x_1, \ldots, x_N) = \sup_{[a,b) \in \mathbf{T}} \left| \frac{\#\{x_n \in [a, b)\}}{N} - (b - a) \right|$$

For an infinite sequence $x_1, x_2, \cdots \in \mathbf{T}$, we can consider a sequence of discrepancy $D_1, D_2, \ldots$ with $D_N = D(x_1, \ldots, x_N)$, measuring the failure for the sequence to be uniformly distributed.

Theorem 4.3. *The sequence is uniformly distributed if and only if $D_N \to 0$.*