Szemerédi’s Theorem via Ergodic Theory

Jonathan Hickman

January 25, 2016

Abstract

This essay investigates Furstenberg’s proof of Szemerédi’s Theorem. The necessary concepts and results from ergodic theory are introduced, including the Poincaré and Mean Ergodic Theorems which are proved in full. The Ergodic Decomposition Theorem is also discussed. Furstenberg’s Multiple Recurrence Theorem is then stated and shown to imply Szemerédi’s Theorem. The remainder of the essay concentrates on proving the Multiple Recurrence Theorem, following the method laid out in [10]. As part of this proof the notions of conditional expectation and measure are introduced and discussed in some detail.

1 Introduction

This essay is concerned with exploring whether certain subsets of the integers contain arithmetic progressions. Recall that an (integer) arithmetic progression is a set \(\{a + jb\}_{j=0}^{k-1} \subset \mathbb{Z}\) where \(k \in \mathbb{N}_{\geq 1}\) is the length of the progression; \(a \in \mathbb{Z}\) and \(b \in \mathbb{N}_{\geq 1}\). Usually we will not be concerned with the values of the starting point \(a\) and step-size \(b\), we merely want to know if given a set \(\Lambda \subseteq \mathbb{Z}\) and \(k \in \mathbb{N}\) there exists a progression of length \(k\) within \(\Lambda\). In 1927 van der Waerden proved the following fundamental theorem:

\[\text{Theorem 1.1 (van der Waerden). Suppose } A_1, \ldots, A_n \subseteq \mathbb{Z} \text{ are disjoint sets which partition the integers, } \mathbb{Z} = A_1 \cup A_2 \cup \cdots \cup A_n \text{. Then there exists some } l \in \{1, \ldots, n\} \text{ such that } A_l \text{ contains arithmetic progressions of arbitrary length.}\]

We will not present a proof of Theorem 1.1 here\(^1\), rather remark that when studying questions pertaining to the presence of arithmetic progressions, it transpires that the idea of partitioning the integers into a finite number of sets (or ‘colouring’ the integers, as it is often termed in such combinatorial contexts) is something of a red herring. A more general approach is to consider the ‘density’ of a set of integers.

\[\text{Definition 1. Let } \Lambda \subseteq \mathbb{Z}. \text{ We define the upper Banach density of } \Lambda \text{ to be } \]

\[d_B(\Lambda) := \limsup_{N \to \infty} \frac{|\Lambda \cap [M, N)|}{N - M}\]

In 1936 Erdős and Turán [6] conjectured the stronger statement that any subset of \(\mathbb{Z}\) with positive upper Banach density contains arithmetic progressions of arbitrary length. Proving this conjecture was no easy task: indeed, it was not until 1952 that the first non-trivial case, existence of arithmetic progressions of length \(k = 3\), was shown by Roth [20] who used techniques from harmonic analysis. Szemerédi [22] then proved the case \(k = 4\) in 1969 by applying combinatorial methods and finally the general case was proven, again by Szemerédi [23], in 1975. The result now bears his name.

\(^1\)A short proof of van der Waerden’s Theorem can be found in [5, 11]. Alternatively, it follows directly from Szemerédi’s Theorem, that is Theorem 1.2 below.

1
Theorem 1.2 (Szemerédi’s Theorem). Let $\Lambda \subseteq \mathbb{Z}$ be a set of positive upper Banach density. Then $\Lambda$ contains arithmetic progressions of arbitrary length.

Szemerédi’s proof of this theorem is long and difficult and we will not concern ourselves with it here. Instead, we investigate an alternative, more accessible proof given by Furstenberg. Soon after Szemerédi’s proof was published, Furstenberg noticed Theorem 1.2 would follow from a substantial generalisation of the classical Poincaré Recurrence Theorem from ergodic theory. He subsequently proved this generalisation, now known as Furstenberg’s Multiple Recurrence Theorem, and provided a scheme, the Correspondence Principle, for relating problems in combinatorial number theory to problems in ergodic theory [7].

The importance of Furstenberg’s work is arguably two-fold: first of all the Multiple Recurrence Theorem is itself a substantial result in ergodic theory. Secondly, the correspondence principle provides an effective tool for tackling problems in number theory. Some results, such as the multidimensional analogue of Szemerédi’s Theorem [9], have only been proven through the use of the ergodic theoretical methods laid out by Furstenberg.

This document presents the proof of Szemerédi’s Theorem as given in Furstenberg, Katznelson and Ornstein’s article [10] assuming the reader has no prior knowledge of ergodic theory (although some familiarity with elementary measure theory and functional analysis). In order to do this we will introduce the necessary definitions and results from ergodic theory and get a flavour of what the discipline entails. We then proceed to state Furstenberg’s Multiple Recurrence Theorem and show how this implies Szemerédi’s Theorem. The bulk of the essay then attempts to explain the proof of the Multiple Recurrence Theorem.

2 Upper Banach and Asymptotic Density

We begin by briefly discussing notions of density of a subset of the integers. The results proved here will not be required until Section 4 and they can thus be considered something of a prelude to the mathematics described later in the essay.

It is easy to deduce from Definition 1 that if a set $\Lambda \subseteq \mathbb{Z}$ has upper Banach density $\bar{d}_B(\Lambda)$ then there exists a sequence of intervals $[a_n, b_n)$ with $b_n - a_n \to \infty$ as $n \to \infty$ such that

$$\bar{d}_B(\Lambda) = \lim_{n \to \infty} \frac{|\Lambda \cap [a_n, b_n)|}{b_n - a_n} \tag{2.1}$$

If we enforce the stronger condition that (2.1) must hold for $[a_n, b_n) = [1, n+1)$ for all $n \in \mathbb{N}$ then we arrive at the stronger notion of asymptotic density.

Definition 2. For $\Lambda \subseteq \mathbb{N}$ define the lower and upper asymptotic densities as

$$d(\Lambda) = \liminf_{N \to \infty} \frac{|\Lambda \cap [1, N]|}{N}$$

and

$$\bar{d}(\Lambda) = \limsup_{N \to \infty} \frac{|\Lambda \cap [1, N]|}{N}$$

respectively. If $d(\Lambda) = \bar{d}(\Lambda)$ then we define the asymptotic density as

$$d(\Lambda) = \lim_{N \to \infty} \frac{|\Lambda \cap [1, N]|}{N}$$

so that $d(\Lambda) = d(\Lambda) = \bar{d}(\Lambda)$.

Obviously any subset $\Lambda \subseteq \mathbb{N}$ of positive upper asymptotic density is automatically of positive upper Banach density.

Example 1. 1. Trivially, any finite set $F \subseteq \mathbb{N}$ has $d(F) = 0$.

2. For any $a, k \in \mathbb{N}$ the arithmetic sequence $\Lambda = a + k\mathbb{N}$ has density $d(\Lambda) = \frac{1}{k}$.

2
An example which is only slightly less trivial, though important, is that of a syndetic set [5].

**Definition 3.** We say that a subset $K \subseteq \mathbb{N}$ is syndetic if there exists some $a \in \mathbb{N}$ such that

$$\bigcup_{i=0}^{a-1} (K - i) = \mathbb{N}$$

(2.2)

where $K - i = \{n \in \mathbb{N} : n + i \in K\}$. This is equivalent to the property that for every $k \in \mathbb{N}$ we have $\{k, k+1, \ldots, k+a-1\} \cap K \neq \emptyset$, so syndetic therefore are sets with ‘bounded gaps’.

It is easy to see a syndetic set $K \subseteq \mathbb{N}$ has positive lower density. For suppose $a \in \mathbb{N}$ satisfies (2.2). Then for each $k \in \mathbb{N}$ the interval $((k-1)a, ka]$ contains at least one element of $K$. Let $N \in \mathbb{N}$ with $N \geq a$. By writing $N = ka + r$ where $k \in \mathbb{N}$ and $0 \leq r < a$ observe

$$\frac{|K \cap [1, N]|}{N} = \sum_{j=1}^{k} \frac{|K \cap ((j-1)a, ja]|}{N} + \frac{|K \cap (ka, ka + r]|}{N} \geq \frac{k}{ka + r} > \frac{1}{2a} > 0$$

and so $d(K) > 0$ as required.

A significant result concerning asymptotic density is a theorem of Koopman and von Neumann, adapted here from ([5], Chapter 2; [1]), which relates asymptotic density and Cesàro summation of sequences. As we shall see, Cesàro sums play a dominant rôle in ergodic theory. The following lemma effectively tells us that the behaviour of a sequence on terms belonging to a set of zero density is ‘averaged out’ when taking the Cesàro mean.

**Theorem 2.1.** Let $(a_n)_{n \in \mathbb{N}} \in l^\infty(\mathbb{R}_{\geq 0})$ be a bounded sequence of non-negative real numbers. Then the following are equivalent

1. The Cesàro Mean of $(a_n)_{n \in \mathbb{N}}$ is 0:

   $$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} a_n = 0$$

2. There exists some set $Z \subset \mathbb{N}$ of zero density such that

   $$\lim_{n \to \infty} a_n = 0$$

Before presenting the proof of Theorem 2.1 we note the following corollary, which is an immediately consequence. This corollary will be repeatedly exploited in the proceeding discourse.

**Corollary 2.2.** Let $(a_n)_{n \in \mathbb{N}} \in l^\infty(\mathbb{R}_{\geq 0})$ be a bounded sequence of non-negative real numbers. Then

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} a_n = 0$$

if and only if

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} a_n^2 = 0$$

We now turn to the proof of Theorem 2.1.

**Proof (of Theorem 2.1).** We first assume 1. Define for each $k \in \mathbb{N}_{\geq 1}$ the set

$$Z_k := \{j \in \mathbb{N} : a_j \geq \frac{1}{k}\}$$

Note that the $Z_k$ form an increasing chain $Z_1 \subseteq Z_2 \subseteq \ldots$. Furthermore, fixing $k \in \mathbb{N},$

$$\frac{|Z_k \cap [1, N]|}{N} \leq \frac{k}{N} \sum_{n \in Z_k \cap [1, N]} a_n \leq \frac{k}{N} \sum_{n=1}^{N} a_n$$

3
Thus, by taking the limit a \( N \to \infty \) each \( Z_k \) is a set of zero density. Moreover, there exists an increasing sequence \( 1 < N_1 < N_2 < \ldots \) such that
\[
\frac{|Z_k \cap [1, N]|}{N} < \frac{1}{k} \quad \text{for all } N \geq N_k \text{ and } k \in \mathbb{N}_{\geq 1}
\]
Now define
\[
Z = \bigcup_{k=1}^{\infty} Z_k \cap [N_{k-1}, N_k)
\]
Note that if \( N \in \{N_{k_0-1}, \ldots, N_{k_0} - 1\} \) for some \( k_0 \in \mathbb{N} \) then, as the \( Z_k \) form an increasing chain, \( Z \cap [1, N] = Z_{k_0} \cap [1, N] \). Thus
\[
\frac{|Z \cap [1, N]|}{N} \leq \frac{|Z_{k_0} \cap [1, N]|}{N} < \frac{1}{k_0}
\]
By taking limits we conclude \( Z \) is a set of zero density. Furthermore, if \( n \in \mathbb{N} \setminus Z \) and \( n \geq N_k \) then by the definition of \( Z \) we must have \( n \notin Z_k \) and so \( a_n < \frac{1}{k} \). Hence it is easy to see that
\[
\lim_{n \to \infty} a_n = 0
\]
as required.

Conversely, suppose 2 holds. Then for any \( \epsilon > 0 \) there exists \( N_1 \in \mathbb{N} \) such that
\[
a_n < \frac{\epsilon}{3} \quad \text{for all } n \geq N_1, n \notin Z
\]
Let \( M > 0 \) be an upperbound on the sequence \( (a_n)_{n \in \mathbb{N}} \). As \( Z \) is a set of zero density there exists \( N_2 \in \mathbb{N} \) such that
\[
\frac{|Z \cap [1, N]|}{N} < \frac{\epsilon}{3M} \quad \text{for all } N \geq N_2
\]
Finally, let \( N_0 > \max\{N_1, N_2\} \) be such that \( \frac{M(N_1 - 1)}{N_0} < \frac{\epsilon}{3} \). Then for all \( N \geq N_0 \) we have
\[
\frac{1}{N} \sum_{n=1}^{N} a_n = \frac{1}{N} \left[ \sum_{n=1}^{N_1-1} a_n + \sum_{n=N_1 \atop n \in Z}^{N} a_n + \sum_{n=N_1 \atop n \notin Z}^{N} a_n \right] \leq \frac{M(N_1 - 1)}{N} + \frac{M|Z \cap [1, N]|}{N} + \frac{1}{N} \sum_{n=N_1 \atop n \notin Z}^{N} a_n < \epsilon
\]
and we are done.

Relating combinatorial concepts (such as asymptotic or upper Banach density) to notions of averaging (such as Cesàro summation or measure) is a major theme of this study and we will return to it in Section 4 when we discuss the connections between ergodic theory and Szemerédi’s theorem. Before then, however, we take some time to introduce the fundamentals of ergodic theory.

3 Elementary Ergodic Theory

This section acts as a brief introduction to ergodic theory. It is stressed that this survey is far from comprehensive\(^2\); indeed it was the author’s intention to minimise the number of prerequisites needed in proving Szemerédi’s Theorem and for the most part we focus only on introducing these prerequisites. The reader is encouraged to consult other sources\(^3\) for a broader introduction to this field.

\(^2\)For instance, one conspicuous omission is Birkhoff’s Pointwise Ergodic Theorem.
\(^3\)For example [5, 13, 17, 19, 21].
3.1 Basic Definitions

Ergodic theory belongs to the broader discipline of the qualitative study of dynamical systems. A dynamical system is essentially consists of a 'state space' \( X \) and transformation \( T : X \rightarrow X \) and one wishes to determine to some extent the behaviour of \( X \) under iterations of \( T \). In the case of ergodic theory the state space and transformation are endowed with measure theoretic structure; this is natural in the context of analysing the long-term average behaviour of systems [19]. The fundamental objects of study are measure preserving systems.

**Definition 4.** Let \((X, \mathcal{B}, \mu)\) be a measure space. Suppose \( T : X \rightarrow X \) is measurable and for all \( A \in \mathcal{B} \) we have \( \mu(T^{-1}A) = \mu(A) \). Then we say \( T \) is a measure preserving transformation on \((X, \mathcal{B}, \mu)\). Furthermore, the pair \((X, \mathcal{B}, \mu, T)\) consisting of the space and transformation is called a measure preserving system.

We say \((X, \mathcal{B}, \mu, T)\) is invertible if the transformation \( T \) is invertible with measurable inverse.

**Remark 1.** In some instances it is convenient to change the perspective somewhat: given a measurable space \((X, \mathcal{B})\) and measurable function \( T : X \rightarrow X \) we say a measure \( \mu \) on \((X, \mathcal{B})\) is \( T \)-invariant if \( T \) is a measure preserving transformation on \((X, \mathcal{B}, \mu)\).

The study of these objects arose from attempts to model certain physical systems in applied mathematics, particularly statistical mechanics (see, for instance [3]). Since then a rich and varied theory of measure preserving systems has been developed with applications in pure mathematics, not least combinatorial number theory and Furstenberg’s proof of Szemerédi’s Theorem.

3.1.1 Examples of Measure Preserving Systems

This discussion will only consider a few simple examples of measure preserving systems. Although practical in the context of this study, our examples do not reflect the true diversity of potential systems and are not representative of ergodic theory as a whole\(^4\).

**Example 2** (Rotations of the Circle). Consider the circle \( T := \mathbb{R}/\mathbb{Z} \) with the usual Lebesgue measure. Then for any \( \tau \in \mathbb{R} \) the rotation map \( R_\tau : T \rightarrow T \) defined by \( R_\tau : x \mapsto x + \tau \)

is clearly measurable with measurable inverse \( R_{-\tau} \). We claim this map is measure preserving. Indeed, by a simple application of Dynkin’s Lemma as \( \Lambda = \{ A \in \mathcal{B} : \mu(R_\tau^{-1}(A)) = \mu(A) \} \) is a \( \text{d}\)-system\(^5\) we only need check \( \mu(R_\tau^{-1}[a,b)) = \mu([a,b)) \) for all intervals \([a,b)) \in T \). But this is obvious.

In studying a measure preserving system \((X, \mathcal{B}, \mu, T)\) we wish to determine qualitatively how the space \((X, \mathcal{B}, \mu)\) behaves under repeated applications of the transformation \( T \). If we wish to investigate the behaviour of a single point \( x \in X \) we define its orbit \( \{ T^n x : n \in \mathbb{N} \} \) and study this set. Even in the simple case of circle rotation we can make non-trivial conclusions relating to the qualitative properties of the dynamics.

The following result is adapted from ([21], Chapter 3).

**Proposition 3.1.** 1. If \( \tau \) is rational then the orbit of any point \( x \in T \) under \( R_\tau \) is a finite subset of \( T \).

\(^4\)There is an abundance of examples of measure preserving systems to be found within the literature. For example, see [19] which give a much better idea of the scope of ergodic theory.

\(^5\)Recall, \( \Lambda \) is a \( \text{d}\)-system on \( T \) if:

1. \( T \in \Lambda \)
2. For all \( A, B \in \Lambda \) such that \( A \subseteq B \) we have \( B \setminus A \in \Lambda \)
3. If \((A_n)_{n \in \mathbb{N}} \subseteq \Lambda \) is an increasing sequence of sets, that is \( A_1 \subseteq A_2 \subseteq \ldots \), then \( \bigcup_{n \in \mathbb{N}} A_n \in \Lambda \)

All these properties are easily verified in this case.
2. If $\tau$ is irrational then the orbit of any point $x \in \mathbb{T}$ under $R_\tau$ is a dense subset of $\mathbb{T}$.

Proof. Let $x \in \mathbb{T}$ be given and suppose $\tau = \frac{p}{q}$ is rational. Then $R_\tau^q$ is the identity map on $\mathbb{T}$ and it follows immediately the orbit $\{T^nx : n \in \mathbb{N}\}$ is finite.

Now consider the case $\tau$ is irrational. First we show that the points $R_\tau^nx$ and $R_\tau^mx$ are distinct for $n \neq m$. For suppose $m,n \in \mathbb{N}$ are such that $R_\tau^nx = R_\tau^mx$. Then $R_\tau^{m-n}x = x$ and it follows that $x + (n-m)\tau = x \mod 1$ so $(n-m)\tau = k$ for some $k \in \mathbb{Z}$. Hence, as $\tau$ is irrational we must have $n = m$.

Let $d : \mathbb{T} \times \mathbb{T} \to [0,\infty)$ denote the metric $d(x,y) = \min\{|x-y|, 1 - |x-y|\}$

It is easy to verify that $d$ is an $R_\tau$-invariant metric and the quotient topology on $\mathbb{T}$ induced from $\mathbb{R}$ is precisely the metric topology on $\mathbb{T}$ with given by $d$. Furthermore, if $d(x,y) < \frac{1}{2}$ then $d(x,y) = |x-y|$.

Let $0 < \epsilon < \frac{1}{2}$ be given. By the Bolzano-Weierstrass property the sequence $(R_\tau^nx)_{n \in \mathbb{N}}$ must have some convergent subsequence. In particular, the must be some $l,k \in \mathbb{N}$ such that $|R_\tau^lx - R_\tau^kx| = d(R_\tau^{l-k}x,x) < \epsilon$. Consider the set $\{R_\tau^{(l-k)n}x : n \in \mathbb{N}\}$

We claim this set is $\epsilon$-dense in $\mathbb{T}$. Indeed, for consecutive points $R_\tau^{(l-k)n}x$ and $R_\tau^{(l-k)(n+1)}x$ we have $0 < d(R_\tau^{(l-k)n}x, R_\tau^{(l-k)(n+1)}x) = d(R_\tau^{l-k}x,x) < \epsilon$

and the claim follows immediately. As $0 < \epsilon < \frac{1}{2}$ was arbitrary it follows $\{R_\tau x : n \in \mathbb{N}\}$ is a dense subset of $\mathbb{T}$. \qed

Example 3 (Bernoulli Shifts). The reader is assumed some familiarity with this archetype probability space and thus we proceed rapidly. For further details see [5]. Consider the finite probability space on $[n] = \{1, \ldots, n\}$ where the elements are given probabilities $p_1, \ldots, p_n$, respectively where each $p_i \geq 0$ with $\sum_{i=1}^{n} p_i = 1$. Explicitly, we define a probability measure $\mu_0$ on the set $[n]$ by $\mu_0([i]) = p_i$ for all $i = 1, \ldots, n$

Consider the set $X = [n]^\mathbb{Z}$ of two-sided sequences in $[n]$, $X := \{ (\omega_l)_{l \in \mathbb{Z}} : \omega_l \in [n] \text{ for all } l \in \mathbb{N} \}$

Let $\mathcal{B}$ be the $\sigma$-algebra generated by the projections $\pi_i : \omega \mapsto \omega_i$ for each $i \in \mathbb{Z}$ and $\mu$ the product measure $\mu = \prod_{l \in \mathbb{N}} \mu_0$. Then $(X, \mathcal{B}, \mu)$ is a probability space.

Define a cylinder set to be any subset of $X$ containing precisely the sequences which agree on a finite number of fixed entries. Explicitly, $A \in \mathcal{B}$ is a cylinder set if there exists a finite co-ordinate set $I \subset \mathbb{Z}$ and $\mathbf{a} : I \to [n]$ such that $A = \{ (\omega_l)_{l \in \mathbb{Z}} \in X : \omega_i = \mathbf{a}(i) \text{ for all } i \in I \}$

Note that the measure of $A$ is given by $\mu(A) = \prod_{i \in I} p_{\mathbf{a}(i)}$

If we denote $\mathcal{G}$ the collection of all cylinder sets it is not difficult to deduce that $\mathcal{G}$ is an algebra\(^6\) and the $\sigma$-algebra $\mathcal{B}$ is generated by $\mathcal{G}$.

\(^6\)Recall, a set system $\mathcal{G} \subseteq \mathcal{P}(X)$ is an algebra if

1. $X \in \mathcal{G}$
2. If $A \in \mathcal{G}$ then $A^c \in \mathcal{G}$
3. If $A,B \in \mathcal{G}$ then $A \cup B \in \mathcal{G}$
There is a simple and natural measure preserving transformation on \((X, \mathcal{B}, \mu)\). Define the left shift transformation \(T : X \to X\) by
\[
T : (\omega_i)_{i \in \mathbb{Z}} \mapsto (\omega_{i+1})_{i \in \mathbb{Z}}
\]
We claim \(T\) is measurable and measure preserving. As in the previous example it suffices only to check that \(T^{-1}(A) \in \mathcal{B}\) and \(\mu(T^{-1}A) = \mu(A)\) for all \(A \in \mathcal{B}\), and both these conditions are obvious.

### 3.1.2 Measure Preserving Transformations and \(L^2\)

Throughout this section, unless otherwise stated, we assume \((X, \mathcal{B}, \mu, T)\) is an arbitrary measure preserving system. It is often fruitful to study \((X, \mathcal{B}, \mu, T)\) by considering the associated Hilbert Space \(L^2(X, \mathcal{B}, \mu)\), exploiting its additional linear structure. We will briefly discuss some basic properties of \(T\) in relation to \(L^2(X, \mathcal{B}, \mu)\).

**Definition 5.** The measure preserving transformation \(T\) induces a linear map
\[
U_T : L^2(X, \mathcal{B}, \mu) \to L^2(X, \mathcal{B}, \mu) \quad U_T : f \mapsto f \circ T
\]
called the associated operator [5]. We will adopt the convention \(Tg = U_Tf\) and hence, depending on the context, \(T\) denotes a map on either \(X\) or \(L^2(X, \mathcal{B}, \mu)\).

**Proposition 3.2.** Let \(f \in L^1(X, \mathcal{B}, \mu)\). Then
\[
\int Tf \, d\mu = \int f \, d\mu
\]

**Proof.** As \(T\) is measure preserving the proposition holds for all characteristic functions. Explicitly, if \(A \in \mathcal{B}\) then
\[
\int T1_A \, d\mu = \int 1_{T^{-1}A} \, d\mu = \mu(T^{-1}A) = \mu(A) = \int 1_A \, d\mu
\]
Thus, by the linearity of the integral the proposition holds for all simple functions \(\phi \in L^1(X, \mathcal{B}, \mu)\).

For any non-negative \(g \in L^1(X, \mathcal{B}, \mu)\) there exists an increasing sequence of simple functions \((\phi_n)_{n \in \mathbb{N}}\) tending pointwise to \(g\). Note \((T\phi_n)_{n \in \mathbb{N}}\) is an increasing sequence of simple functions tending pointwise to \(Tg\). By monotone convergence
\[
\int Tg \, d\mu = \lim_{n \to \infty} \int T\phi_n \, d\mu = \lim_{n \to \infty} \int \phi_n \, d\mu = \int g \, d\mu.
\]
The general case \(f \in L^1(X, \mathcal{B}, \mu)\) is then obtained by writing \(f = f_+ - f_-\) where \(f_+, f_- \in L^1(X, \mathcal{B}, \mu)\) are non-negative. \(\square\)

An immediate implication of Proposition 3.2 is that the associated operator \(T : L^2(X, \mathcal{B}, \mu) \to L^2(X, \mathcal{B}, \mu)\) is an isometry. Furthermore, if \((X, \mathcal{B}, \mu, T)\) is an invertible system then \(T\) is unitary.

### 3.2 Poincaré Recurrence

Much of this essay is concerned with notions of ‘recurrence’ in measure preserving systems and here we introduce the simplest of these: that of Poincaré Recurrence. Recall, given a system \((X, \mathcal{B}, \mu, T)\) and \(x \in X\) we define the orbit of \(x \in X\) as the set \(\{T^n x : n \in \mathbb{N}\}\). Assuming the reader has some familiarity with the basic premise of the the study of dynamical systems, it is natural (or at least convenient for discursive purposes) to imagine \(x\) as the ‘initial state’ of a system and the \(T^n x\) as describing the evolution of the system over discrete units of time. An obvious question is whether an orbit will return to its initial state (or something close to it) at some point in the future. That is, given some ‘neighbourhood’ \(A\) of \(x\) does there exist some \(n \in \mathbb{N}\) such that \(T^n x \in A\)? In the measure-theoretic context it is natural that our neighbourhood should be a measurable set with \(\mu(A) > 0\). The Poincaré Recurrence Theorem states that for any finite\(^7\) measure preserving systems \((X, \mathcal{B}, \mu, T)\), given a measurable set \(A \in \mathcal{B}\) the orbits

\(^7\)We say the system \((X, \mathcal{B}, \mu, T)\) is finite if \(\mu(X) < \infty\).
of almost all the points \( x \in A \) return to \( A \) infinitely many times. The statement and proof of this theorem is adapted from ([5] Chapter 2; [19] Chapter 2).

**Theorem 3.3** (Poincaré Recurrence Theorem). Let \((X, \mathcal{B}, \mu, T)\) be a finite measure preserving system. For any \( A \in \mathcal{B} \), for almost every \( x \in A \) there exists a strictly increasing sequence \((n_k)_{k \in \mathbb{N}} \subseteq \mathbb{N}\) such that \( T^{n_k}x \in A \) for all \( k \in \mathbb{N} \).

**Proof.** Consider the set \( E := \{x \in A : T^n x \notin A \text{ for all } n \in \mathbb{N}_{\geq 1}\} \). Alternatively,

\[
E = \bigcap_{n=1}^{\infty} A \setminus T^{-n}A
\]

Clearly \( E \) is measurable. It is easy to see the sets \( E, T^{-1}E, T^{-2}E, \ldots \) are pairwise disjoint and, as \( T \) is measure preserving, they all have measure \( \mu(E) \). Note that

\[
\sum_{n=1}^{\infty} \mu(E) = \mu\left( \bigcup_{n=1}^{\infty} T^{-n}E \right) \leq \mu(X) < \infty
\]

Hence \( \mu(E) = 0 \) and if we define \( F_1 = A \setminus E \) then \( \mu(F_1) = \mu(A) \) and for all \( x \in F_1 \) there exists an \( n \in \mathbb{N}_{\geq 1} \) such that \( T^n x \in A \). Note that \( T^2, T^3, \ldots \) define measure preserving transformations on \((X, \mathcal{B}, \mu)\) and we may repeat the above construction to give \( F_2, F_3, \ldots \subseteq A \) each with measure \( \mu(A) \) and every point in \( F_k \) returning to \( A \) under the transformation \( T^k \). Finally, define

\[
F := \bigcap_{k=1}^{\infty} F_k
\]

then \( \mu(F) = \mu(A) \) and every \( x \in F \) returns to \( A \) under \( T \) infinitely many times. \( \qed \)

We now consider the orbit of an entire measurable set \( A \) with \( \mu(A) > 0 \) rather than just individual points.

**Corollary 3.4.** Let \((X, \mathcal{B}, \mu, T)\) be a finite measure preserving system and \( A \in \mathcal{B} \) with \( \mu(A) > 0 \). Then there exists some \( n \in \mathbb{N} \) such that

\[
\mu(T^{-n}A \cap A) > 0 \quad (3.1)
\]

**Proof.** For each \( j \in \mathbb{N}_{\geq 1} \) consider the set \( T^{-j}A \cap A = \{x \in A : T^j x \in A\} \). Then by the Poincaré Recurrence Theorem

\[
\mu\left( \bigcup_{j=1}^{\infty} T^{-j}A \cap A \right) = \mu(A) > 0
\]

and so there must exist some \( n \in \mathbb{N} \) such that \( \mu(T^{-n}A \cap A) > 0 \) as required. \( \qed \)

The main body of this essay is dedicated to proving a remarkable generalisation of the version of the Poincaré Recurrence Theorem stated in Corollary 3.4. Later we will investigate a much stronger notion of recurrence than that described in (3.1) known as multiple recurrence. The analogous result to Corollary 3.4 for multiple recurrence is precisely what is required in order to prove Szemerédi’s Theorem. This connection will be explained in detail in Section 4.

**Remark 2.** It is clear from the proof of the Poincaré Recurrence Theorem that the measure space in question must be finite\(^8\). As we will be considering a generalisation of this result, we will henceforth assume (tacitly) all measure preserving systems are finite, and in particular are probability systems.

\(^8\)It is not difficult to conceive counter-examples in an infinite setting; for instance, consider the shift map \( x \mapsto x+1 \) on \( \mathbb{R} \) with Lebesgue measure.
3.2.1 Ergodicity

A key notion in the study of measure preserving systems is that of ergodicity.

**Definition 6.** Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system. A set \(A \in \mathcal{B}\) is \(T\)-invariant if \(T^{-1}A = A\). The collection of all \(T\)-invariant sets forms a \(\sigma\)-algebra, which we denote \(\mathcal{E}_T\).

**Definition 7.** A measure preserving transformation \(T\) on \((X, \mathcal{B}, \mu)\) is ergodic if the \(\sigma\)-algebra \(\mathcal{E}_T\) contains only null and conull sets.

**Remark 3.** In some instances it is convenient to change the perspective: if \((X, \mathcal{B})\) is a measurable space and \(T : X \to X\) a measurable function we say that a \(T\)-invariant measure \(\mu\) is an ergodic measure for \(T\) if \(T\) is ergodic as a measure preserving transformation on \((X, \mathcal{B}, \mu)\).

We characterise ergodicity in two (equivalent) ways:

- Immediately from the definition we deduce that ergodicity is an indecomposability condition. If \(A \in \mathcal{B}\) is a \(T\)-invariant set with \(0 < \mu(A) < 1\), then so too is \(B = A^c\). For any \(x \in A\) (resp. \(B\)) the orbit of \(x\) never leaves \(A\) (resp. \(B\)) and hence, for all intents and purposes, \(A\) and \(B\) may be considered\(^9\) separate dynamical systems. When \(T\) is ergodic every \(T\)-invariant set has either measure 0 or measure 1: it is only possible to decompose the system into a portion which is essentially the whole system and a portion which is negligible.
- Ergodicity is precisely the condition that for almost all \(x \in X\) the orbit \(\{T^n x : n \in \mathbb{N}\}\) visits almost all of the space. This will be explained in detail below (Lemma 3.6). In particular, we will show that \(T\) is an ergodic transformation if and only if for all \(A, B \in \mathcal{B}\) with \(\mu(A), \mu(B) > 0\) there exists some \(n \in \mathbb{N}\) such that

\[
\mu(T^{-n}A \cap B) > 0
\]

Note that this condition is much stronger than the similar-looking statement of Poincaré Recurrence, given in (3.1).

In order to explain these characterisations fully, first we note that the definition of ergodicity is equivalent to a similar statement concerning ‘almost invariant’ rather than invariant sets. The proof of this lemma is adapted from ([5], Chapter 2).

**Lemma 3.5.** A system \((X, \mathcal{B}, \mu, T)\) is ergodic if and only if for all \(A \in \mathcal{B}\) such that \(\mu(A \Delta T^{-1}A) = 0\) either \(\mu(A) = 0\) or \(\mu(A) = 1\).

**Proof.** If the condition holds then it is obvious that the system is ergodic. For the converse, suppose \((X, \mathcal{B}, \mu, T)\) is ergodic and \(A \in \mathcal{B}\) satisfies \(\mu(A \Delta T^{-1}A) = 0\). It suffices to show that there exists a \(T\)-invariant set \(B\) such that \(\mu(B) = \mu(A)\). Define \(B\) as the measurable set

\[
B = \bigcap_{N=0}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}A
\]

First we show that \(B\) is \(T\)-invariant. Let \(B_N = \bigcup_{n=N}^{\infty} T^{-n}A\). Then \(B_0 \supseteq B_1 \supseteq B_2 \supseteq \ldots\) and so \(\bigcap_{N=0}^{\infty} B_{N_k} = \bigcap_{n=k}^{\infty} B_N\) for all \(k \in \mathbb{N}\). In particular,

\[
T^{-1}B = \bigcap_{N=0}^{\infty} \bigcup_{n=N}^{\infty} T^{-(n+1)}A = \bigcup_{N=0}^{\infty} \bigcap_{n=N}^{\infty} T^{-n}A = B
\]

Hence \(B\) is \(T\)-invariant. To see that \(\mu(B) = \mu(A)\) note that for all \(N \in \mathbb{N}\),

\[
\mu(A \Delta B_N) \leq \mu \left( \bigcup_{n=N}^{\infty} A \Delta T^{-n}A \right) \\
\leq \sum_{n=N}^{\infty} \mu(A \Delta T^{-n}A) = 0
\]

\(^9\)By restricting \(\mathcal{B}, \mu\) and \(T\) to these sets.
where we have used fact that \( \mu(A \triangle T^{-n}A) = 0 \) for all \( n \in \mathbb{N} \), which follows immediately from

\[
A \triangle T^{-n}A \subseteq \bigcup_{j=0}^{n-1} T^{-j}A \triangle T^{-(j+1)}A
\]

As \( \mu(A \triangle B_N) = 0 \) for all \( N \in \mathbb{N} \) and \( B = \bigcup_{N=0}^{\infty} B_N \), it follows \( \mu(A \triangle B) = 0 \) and so \( \mu(A) = \mu(B) \) as required.

We are now in a position to elucidate precisely the second characterisation of ergodicity.

**Proposition 3.6.** Let \((X,\mathscr{B},\mu,T)\) be a measure preserving system. The following are equivalent:

1. The system is ergodic.
2. For any \( A \in \mathscr{B} \) with \( \mu(A) > 0 \) we have
   \[
   \mu\left( \bigcup_{n=0}^{\infty} T^{-n}A \right) = 1 \tag{3.2}
   \]
3. For any \( A,B \in \mathscr{B} \) with \( \mu(A),\mu(B) > 0 \) there exists some \( n_0 \in \mathbb{N} \) such that
   \[
   \mu(T^{-n_0}A \cap B) > 0 \tag{3.3}
   \]

**Remark 4.** The behaviour of the orbits of an ergodic system can be understood thus: (3.2) tells us that for any given set \( A \) of positive measure, for almost every \( x \in X \) the orbit of \( x \) will at some point pass through \( A \). From this it is easy to deduce (3.3): given another set \( B \) of positive measure, at some point the orbits of a significant proportion of the elements of \( B \) will pass through \( A \). It is in this sense that ergodicity is the condition that ‘almost all of the orbits visit almost all of the space’.

**Proof (of Proposition 3.6).** (1 \( \Rightarrow \) 2). Let \( A \in \mathscr{B} \) with \( \mu(A) > 0 \) and consider the set

\[
B = \bigcup_{j=0}^{\infty} T^{-j}A
\]

Then \( T^{-1}B \subseteq B \) and \( \mu(T^{-1}B) = \mu(B) \) so we must have \( \mu(B \triangle T^{-1}B) = 0 \). By Lemma 3.5 and noting \( A \subseteq B \) it follows \( \mu(B) = 1 \).

(2 \( \Rightarrow \) 3). Let \( A,B \in \mathscr{B} \) with \( \mu(A),\mu(B) > 0 \) and note

\[
0 < \mu(B) = \mu\left( B \cap \bigcup_{j=1}^{\infty} T^{-j}A \right) \leq \mu \sum_{j=1}^{\infty} (B \cap T^{-j}A)
\]

Thus \( \mu(B \cap T^{-n}A) > 0 \) for some \( n \in \mathbb{N} \).

(3 \( \Rightarrow \) 1). Suppose \( A \in \mathscr{B} \) is \( T \)-invariant. Then for all \( j \in \mathbb{N} \) we have \( \mu(T^{-j}A \cap A^c) = 0 \) and thus either \( \mu(A) = 0 \) or \( \mu(A^c) = 0 \) as required.

Finally, we note an easily deduced equivalent definition of ergodicity, described in terms of invariant functions rather than invariant sets.

**Definition 8.** Let \((X,\mathscr{B},\mu,T)\) be a measure preserving space. We say \( f \in L^1(X,\mathscr{B},\mu) \) is \( T \)-invariant if \( f = Tf \). Explicitly, \( f \) is \( T \)-invariant if for almost all \( x \in X \) we have \( f(x) = f(Tx) \)

**Proposition 3.7.** Let \((X,\mathscr{B},\mu,T)\) be a measure preserving space. Then \( T \) is ergodic if and only if every \( T \)-invariant function is constant.
Proof. Suppose all \( T \)-invariant functions are constant. Given an invariant set \( A \in \mathcal{B} \) it is easy to see that the characteristic function \( 1_A \) is \( T \)-invariant. Hence either \( 1_A(x) = 0 \) or \( 1_A(x) = 1 \) almost everywhere and so \( T \) is ergodic.

Conversely, suppose \( T \) is ergodic and \( f \in L^1(X, \mathcal{B}, \mu) \) is \( T \)-invariant. Let \( a \in \mathbb{R} \) and define the set \( A(a) := \{ x \in X : f(x) < a \} \). Then \( A(a) \) is measurable and
\[
T^{-1}A(a) = \{ x \in X : Tf(x) < a \} = \{ x \in X : f(x) < a \} = A(a)
\]
For each \( a \in \mathbb{R} \) the set \( A(a) \) is \( T \)-invariant and so by ergodicity \( \mu(A(a)) = 0 \) or \( \mu(A(a)) = 1 \). Taking
\[
a_0 = \sup \{ a \in \mathbb{R} : \mu(A(a)) = 0 \}
\]
it is easy to see \( f(x) = a_0 \) for almost all \( x \in X \).

3.2.2 Examples of Ergodicity

In order to prove the ergodicity of certain systems we will utilise the following approximation argument.

Lemma 3.8. Let \((X, \mathcal{B}, \mu)\) be a probability space where \( \mathcal{B} = \sigma(\mathcal{F}) \) is generated by an algebra \( \mathcal{F} \). Then for any \( \epsilon > 0 \) and any \( A \in \mathcal{B} \) there exists an \( A_1 \in \mathcal{F} \) such that
\[
\mu(A \Delta A_1) < \epsilon \tag{3.4}
\]
In particular, we may take \( A_1 \in \mathcal{F} \) to be such that \( A \subseteq A_1 \) and \( A_1 = \bigcup_{i=1}^n B_i \) for some \( B_1, \ldots, B_n \in \mathcal{F} \) disjoint.

Proof. The proof relies on the notion of outer measure, as usually defined in the proof of Carathéodory’s Extension Theorem (see [14] Chapter 1; [4] Chapter 4). In particular, for \( A \in \mathcal{B} \) define
\[
U_A := \left\{ \mathcal{F} \subseteq \mathcal{F} : \text{is countable; } A \subseteq \bigcup_{F \in \mathcal{F}} F \right\}
\]
Then the outer measure of \( \mu^*(A) \) of \( A \) is the value
\[
\mu^*(A) := \inf_{\mathcal{F} \in U_A} \sum_{F \in \mathcal{F}} \mu(F)
\]
It is well known that the outer measure and measure of \( A \in \mathcal{B} \) coincide. Hence given \( \epsilon > 0 \) there exists a sequence \( (B_i)_{i \in \mathbb{N}} \subseteq \mathcal{F} \) such that
\[
\mu(A) \geq \sum_{i=1}^\infty \mu(B_i) - \frac{\epsilon}{2} \tag{3.5}
\]
As the left hand side of this inequality is finite, the sum on the right hand side must converge. Let \( n \in \mathbb{N} \) be such that
\[
\sum_{i=n+1}^\infty \mu(B_i) < \frac{\epsilon}{2} \tag{3.6}
\]
and define
\[
A_1 = \bigcup_{i=1}^n B_i \in \mathcal{F} \quad A_2 = \bigcup_{i=n+1}^\infty B_i \in \mathcal{B}
\]
Note that without loss of generality we may assume the \( B_1, \ldots, B_n \) are disjoint. To see the choice of \( A_1 \) satisfies (3.4), note that
\[
A \Delta A_1 \subseteq ((A_1 \setminus A) \cup (A \setminus (A_1 \cup A_2))) \cup A_2 \subseteq ((A_1 \cup A_2) \Delta A) \cup A_2
\]
Hence, using the bounds stated in (3.5) and (3.6) we have
\[
\mu(A \Delta A_1) \leq \mu((A_1 \cup A_2) \Delta A) + \mu(A_2) < \epsilon
\]
as required. \( \square \)
We now apply Lemma 3.8 to our two examples of measure preserving systems to see they are both ergodic. The proof of the next proposition is adapted from ([21], Chapter 3).

**Proposition 3.9.** Irrational rotations of the circle (c.f. Example 2) are ergodic.

**Proof.** We use the characterisation (3) of ergodicity from Proposition 3.6. Let $A, B \subseteq \mathbb{T}$ be measurable subsets with $\mu(A), \mu(B) > 0$.

**Claim.** There exist intervals $I = [a, b), J = [c, d) \subseteq \mathbb{T}$ such that

$$
\mu(A \cap I) > \frac{3}{4} \mu(I) \quad \mu(B \cap J) > \frac{3}{4} \mu(J) \quad (3.7)
$$

The collection of finite unions of (disjoint) intervals of the form $[x, y) \subseteq \mathbb{T}$ is an algebra which we denote $\mathcal{G}$. Note that $\mathcal{G}$ generates the Borel $\sigma$-algebra on $\mathbb{T}$. Thus, by Lemma 3.8 there exists a sequence of disjoint intervals $I_1, \ldots, I_N$ such that, by denoting $A_1 = \bigcup_{i=1}^N I_i$ we have

$$
A \subseteq A_1 \quad \text{and} \quad \mu(A_1 \setminus A) < \frac{1}{I} \mu(A)
$$

Now,

$$
\mu(A \cap A_1) = \mu(A_1) - \mu(A_1 \setminus A) > \frac{6}{I} \mu(A) \quad (3.8)
$$

On the other hand,

$$
\mu(A_1) = \mu(A_1 \setminus A) + \mu(A) < \frac{8}{I} \mu(A) \quad (3.9)
$$

We claim that for some $i_0 \in \{1, \ldots, N\}$ we have $\mu(A \cap I_{i_0}) > \frac{3}{4} \mu(I_{i_0})$. Once this is shown by setting $I = I_{i_0}$ and repeating the argument for the set $B$ to obtain $J$ the claim is proven. Suppose not; then for all $i \in \{1, \ldots, N\}$ we have

$$
\mu(A \cap I_i) \leq \frac{3}{4} \mu(I_i)
$$

Thus, by applying (3.9) we have $\mu(A \cap A_1) \leq \frac{3}{4} \mu(A_1) < \frac{6}{I} \mu(A)$. But this contradicts (3.8).

Now we have the claim, the result follows easily. Let $I = [a, b)$ and $J = [c, d)$ satisfy (3.7). Without loss of generality $\mu(I) \geq \mu(J)$. Note that, we may also assume $\frac{1}{2} \mu(I) \leq \mu(J)$. Indeed, otherwise we can partition $I$ into two intervals $I_1 = [a, \frac{a+b}{2})$ and $I_2 = [\frac{a+b}{2}, b)$. At least one half $I_i$ must be such that $\mu(I_i \cap A) > \frac{3}{4} \mu(I_i)$ and so we rename the interval $I_i$ as $I$ and repeat the halving process until $\frac{1}{2} \mu(I) \leq \mu(J)$. By Proposition 3.1, the orbit of $d \in \mathbb{T}$ is dense in $\mathbb{T}$ and so there exists some $n \in \mathbb{N}$ such that

$$
\frac{b + a}{2} < T^n d < b
$$

It follows that $\mu(T^{-n} I \cap J) > \frac{1}{2} \mu(I)$. Finally

$$\begin{align*}
\mu(T^{-n} A \cap B) & \geq \mu(T^{-n} I \cap J) - \mu(I \setminus A) - \mu(J \setminus B) \\
& > \frac{1}{2} \mu(I) - \frac{3}{4} \mu(I) - \frac{1}{4} \mu(J) > 0
\end{align*}$$

as required. $\square$

The proof of the next proposition is adapted from ([5], Chapter 2).

**Proposition 3.10.** Bernoulli Shifts (c.f. Example 3) are ergodic.
Proof. Let \((X, \mathcal{B}, \mu, T)\) denote the Bernoulli system where \(X = [n]^\mathbb{Z}\). As in the previous proposition, we show that the characterisation (3) of ergodicity from Proposition 3.6 holds for this system. Moreover, we show the stronger property that for all \(A, B \in \mathcal{B}\) we have

\[
\lim_{n \to \infty} \left| \mu(T^{-n}A \cap B) - \mu(A)\mu(B) \right| = 0 \tag{3.10}
\]

Thus, considering the case \(\mu(A), \mu(B) > 0\), ergodicity follows directly.

Let \(A, B \in \mathcal{B}\) and \(\epsilon > 0\) be given. By Lemma 3.8 there exists cylinder sets \(A_1, B_1\) such that \(A \subseteq A_1, B \subseteq B_1\) and

\[
\mu(A_1 \triangle A) < \frac{\epsilon}{4} \quad \text{and} \quad \mu(B_1 \triangle B) < \frac{\epsilon}{4}
\]

Let \(I_{A_1}, I_{B_1} \subseteq \mathbb{Z}\) be the finite co-ordinate sets for \(A_1\) and \(B_1\) respectively and \(a : I_{A_1} \to [n]\) and \(b : I_{B_1} \to [n]\) be such that

\[
A_1 = \{ (\omega_l)_{l \in \mathbb{Z}} \in X : \omega_l = a(i) \text{ for all } i \in I_{A_1} \}
\]

\[
B_1 = \{ (\omega_l)_{l \in \mathbb{Z}} \in X : \omega_l = b(j) \text{ for all } j \in I_{B_1} \}
\]

Choose any \(n > \max_{i \in I_{A_1}} |i| + \max_{j \in I_{B_1}} |j|\) so that the co-ordinate sets \(T^{-n}A_1 = \{ i-n : i \in I_{A_1} \}\) and \(I_{B_1}\) are disjoint. Then

\[
T^{-n}A_1 = \{ (\omega_l)_{l \in \mathbb{Z}} \in X : \omega_l = a(i+n) \text{ for all } i \in I_{T^{-n}A_1} \}
\]

is a cylinder set and it follows that the sets \(T^{-n}A_1\) and \(B_1\) are independent. Thus

\[
\mu(T^{-n}A_1 \cap B_1) = \mu(A_1)\mu(B_1) \tag{3.11}
\]

Also note that

\[
\mu(T^{-n}A \cap B) \leq \mu(T^{-n}A_1 \cap B_1) \leq \mu(T^{-n}A \cap B) + \mu(A_1 \setminus A) + \mu(B_1 \setminus B) \tag{3.12}
\]

Hence, by applying the triangle inequality and (3.11) and (3.12) we have

\[
\left| \mu(T^{-n}A \cap B) - \mu(A)\mu(B) \right| \leq \left| \mu(T^{-n}A_1 \cap B_1) - \mu(T^{-n}A \cap B) \right| + \mu(A_1)|\mu(B_1) - \mu(B)| + |\mu(A_1) - \mu(A)||\mu(B)|
\]

\[
\leq 2\mu(A_1 \setminus A) + 2\mu(B_1 \setminus B) < \epsilon
\]

as required. \(\square\)

Remark 5. The property of the Bernoulli system described in (3.10) is known as mixing. This is an important concept and later in this discourse we will spend some time investigating the notion of mixing in its own right.

### 3.3 The Mean Ergodic Theorem

For our purposes, the Mean Ergodic Theorem will provide an invaluable tool in the study of measure preserving systems: in particular it gives a useful equivalent definition of ergodicity (see Corollaries 3.12 and 3.13) and also serves as a partial motivation for the study of mixing in Section 6. Having said that, the result is certainly of interest in its own right and it would be negligent to omit some explanation of the intuition behind it.

Given a (finite) measure preserving system \((X, \mathcal{B}, \mu, T)\) and function \(f \in L^2(X, \mathcal{B}, \mu)\) there are two canonical definitions of the average value of \(f\). The first is the ‘space’ average \(\bar{f}\), whereby we take the average value of \(f\) over all possible states:

\[
\bar{f} := \frac{1}{\mu(X)} \int f \, d\mu \tag{3.13}
\]
Of course, if $(X, \mathcal{B}, \mu)$ is a probability space this is merely the expected value of $f$. The second notion of averaging is the ‘time’ average. The time average (provided it exists) is a function $f \in L^1(X, \mathcal{B}, \mu)$ where $\hat{f}(x)$ is defined almost everywhere as the average value of $f$ over the orbit $\{T^n x : n \in \mathbb{N}\}$:

$$\hat{f} := \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} T^n f$$

(3.14)

An important question originally motivated by statistical mechanics [21] which we will now partially examine asks under what conditions the two averages agree. Perhaps it is reasonable to hypothesise that if the orbits of the system penetrated almost every facet of the state space, as in the case of $T$ ergodic, then (3.13) and (3.14) are equal. The classical Ergodic Theorems answer questions relating to the existence of (3.14) and detail precisely the relationship between the values (3.13) and (3.14). As the name suggests, the Mean Ergodic Theorem shows convergence of $\frac{1}{N} \sum_{n=0}^{N-1} T^n f$ in the $L^2(X, \mathcal{B}, \mu)$ sense. In particular, if the system is ergodic we have

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n f - f \right\|_{L^2(\mu)} = 0$$

providing an answer to our question.

The statement and proof of von Neumann’s Theorem are adapted from ([17], Chapter 2; [21], Chapter 5).

**Theorem 3.11** (von Neumann’s Mean Ergodic Theorem). Suppose $(X, \mathcal{B}, \mu, T)$ is a measure preserving system on a probability space and let $\pi_1$ be the orthogonal projection of $L^2(X, \mathcal{B}, \mu)$ onto the closed linear subspace $I$ of $T$-invariant $L^2(X, \mathcal{B}, \mu)$ functions

$$I := \{ g \in L^2(X, \mathcal{B}, \mu) : Tg = g \}$$

Then for any fixed $f \in L^2(X, \mathcal{B}, \mu)$

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n f - \pi_1(f) \right\|_{L^2(\mu)} = 0$$

**Proof.** Consider the set $J := \{ Tg - g : g \in L^2(X, \mathcal{B}, \mu) \}$. We claim that $J^\perp = I$. Indeed, given $f \in J^\perp$ we have

$$\langle f, Tg \rangle = \langle f, g \rangle \quad \text{for all} \ g \in L^2(X, \mathcal{B}, \mu)$$

Hence, $T^* f = f$ where $T^*$ denotes the adjoint of $T$. Note that

$$\| f - T f \|_{L^2(\mu)} = \| f \|_{L^2(\mu)}^2 - \langle f, T f \rangle - \langle T f, f \rangle + \| T f \|_{L^2(\mu)}^2 = 2 \| f \|_{L^2(\mu)}^2 - \langle f, T^* f \rangle - \langle f, T f \rangle = 0$$

and so $f = T f$ as required. Conversely, if $f \in I$ then $f = T f$ and thus

$$\langle f, T g - g \rangle = \langle T f, T g \rangle - \langle f, g \rangle = 0 \quad \text{for all} \ g \in L^2(X, \mathcal{B}, \mu)$$

Hence $I = J^\perp$ as claimed. Therefore, by orthogonal decomposition we have $L^2(\mu) = I \oplus \text{clos}(J)$ and so if we fix $f \in L^2(\mu)$ we may write

$$f - \pi_1(f) = h \quad \text{where} \ h \in \text{clos}(J)$$

To conclude the proof we show that

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n h \right\|_{L^2(\mu)} = 0$$

14
Indeed, given $\epsilon > 0$, as $h \in \text{clos}(J)$ there exists $h_0 \in J$ such that $\|h - h_0\|_{L^2(\mu)} < \frac{\epsilon}{2}$. Furthermore, there exists $g_0 \in L^2(X, \mathcal{B}, \mu)$ such that $h_0 = Tg_0 - g_0$ and we have

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n(Tg_0 - g_0) \right\|_{L^2(\mu)} = \frac{1}{N} \left\| T^N g_0 - g_0 \right\|_{L^2(\mu)} \leq \frac{2\|g_0\|_{L^2(\mu)}}{N}$$

and thus there exists $N_0 \in \mathbb{N}$ such that

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n h_0 \right\|_{L^2(\mu)} \leq \frac{\epsilon}{2} \quad \text{for all } N \geq N_0$$

Finally, it follows that for any $N \geq N_0$ we have

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n h \right\|_{L^2(\mu)} \leq \frac{1}{N} \sum_{n=0}^{N-1} \left\| T^n (h - h_0) \right\|_{L^2(\mu)} + \left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n h_0 \right\|_{L^2(\mu)} < \epsilon$$

\[\square\]

**Remark 6.** Considering the case $(X, \mathcal{B}, \mu, T)$ is an ergodic system, for any $f \in L^2(X, \mathcal{B}, \mu)$ the function $\pi_T(f)$ as defined above is $T$-invariant and thus, by Proposition 3.7 constant. It is then easy to deduce $\pi_T(f) = \int f \, d\mu$, as we will see in the proof of the following corollary.

A corollary to the Mean Ergodic Theorem is the following equivalent definition of ergodicity.

**Corollary 3.12.** A measure preserving system $(X, \mathcal{B}, \mu, T)$ is ergodic if and only if for every pair $f, g \in L^2(X, \mathcal{B}, \mu)$ the following holds

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} (T^n f, g) - \langle f, 1 \rangle \langle 1, g \rangle = 0 \quad (3.15)$$

**Proof.** Suppose $(X, \mathcal{B}, \mu, T)$ is ergodic and fix $f \in L^2(X, \mathcal{B}, \mu)$. By the Mean Ergodic Theorem

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} T^n f - \pi_T(f) \right\|_{L^2(\mu)} = 0$$

Where $\pi_T(f)$ is some $T$-invariant function. By ergocity $\pi_T(f) = c1$ for some constant $c$. As norm convergence implies weak convergence, it follows that for all $g \in L^2(X, \mathcal{B}, \mu)$

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} (T^n f, g) - c \langle 1, g \rangle \right\| = 0$$

In particular, taking $g = 1$ and noting $\langle T^n f, 1 \rangle = \langle T^n f, T^n 1 \rangle = \langle f, 1 \rangle$ gives $c = \langle f, 1 \rangle$.

Conversely, suppose (3.15) holds and $f \in L^2(X, \mathcal{B}, \mu)$ is $T$-invariant. Then for all $g \in L^2(X, \mathcal{B}, \mu)$ we have

$$\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} (T^n f, g) - \langle f, 1 \rangle \langle 1, g \rangle \right| = 0$$

On the other hand, for all $N \in \mathbb{N},$

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} (T^n f, g) - \langle f, 1 \rangle \langle 1, g \rangle \right| = \left| (f - \langle f, 1 \rangle 1, g) \right|$$

Thus $\langle f - \langle f, 1 \rangle 1, g \rangle = 0$ for all $g \in L^2(X, \mathcal{B}, \mu)$ and so by elementary properties of the inner product we must have $f \equiv \langle f, 1 \rangle 1$ is constant. \[\square\]
We can weaken the statement of Corollary 3.12.

**Corollary 3.13.** A measure preserving system $\langle X, \mathcal{B}, \mu, T \rangle$ is ergodic if and only there exists a dense subset $H \subseteq L^2(X, \mathcal{B}, \mu)$ such that for every pair $\phi, \psi \in H$ the following holds

$$\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} \langle T^n \phi, \psi \rangle - \langle \phi, 1 \rangle \langle 1, \psi \rangle \right| = 0 \tag{3.16}$$

**Proof.** By Corollary 3.12 it suffices to show that if (3.16) holds for all $\phi, \psi \in H$, then it holds for all $f, g \in L^2(X, \mathcal{B}, \mu)$.

Let $\epsilon > 0$ be given and $f, g \in L^2(X, \mathcal{B}, \mu)$ with $f, g \neq 0$. Then there exist $\phi, \psi \in H$ such that

$$\|g - \psi\|_2 < \frac{\epsilon}{4\|f\|_2}$$

$$\|f - \phi\|_2 < \frac{\epsilon}{8\|\psi\|_2}$$

Now, for each $N \in \mathbb{N}$

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} \langle T^n f, \psi \rangle - \langle f, 1 \rangle \langle 1, \psi \rangle \right| \leq \frac{1}{N} \sum_{n=0}^{N-1} \left| \langle T^n (f - \phi), \psi \rangle - \langle f - \phi, 1 \rangle \langle 1, \psi \rangle \right|$$

$$+ \left| \frac{1}{N} \sum_{n=0}^{N-1} \langle T^n \phi, \psi \rangle - \langle \phi, 1 \rangle \langle 1, \psi \rangle \right|$$

Note that as $(X, \mathcal{B}, \mu)$ is a probability space, $\|(f - \phi, 1)\| \leq \|f - \phi\|_2$. Hence, by the triangle and Cauchy-Schwarz inequalities,

$$\left| \langle T^n (f - \phi), \psi \rangle - \langle f - \phi, 1 \rangle \langle 1, \psi \rangle \right| \leq 2\|f - \phi\|_2 \|\psi\|_2 < \frac{\epsilon}{4} \tag{3.18}$$

By assumption (3.16) holds for $\phi, \psi \in H$ and thus by combining (3.17) and (3.18) there exists some $N_0 \in \mathbb{N}$ such that

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} \langle T^n f, \psi \rangle - \langle f, 1 \rangle \langle 1, \psi \rangle \right| < \frac{\epsilon}{2} \quad \text{for all } N > N_0 \tag{3.19}$$

By what amounts to the same procedure, this time estimating

$$\left| \langle T^n f, g \rangle - \langle f, 1 \rangle \langle 1, g \rangle \right| \leq \left| \langle T^n f, g - \psi \rangle - \langle f, 1 \rangle \langle 1, g - \psi \rangle \right|$$

$$+ \left| \langle T^n f, \psi \rangle - \langle f, 1 \rangle \langle 1, \psi \rangle \right|$$

and then applying (3.19), we conclude

$$\left| \frac{1}{N} \sum_{n=0}^{N-1} \langle T^n f, g \rangle - \langle f, 1 \rangle \langle 1, g \rangle \right| < \epsilon \quad \text{for all } N > N_1$$

for some $N_1 \in \mathbb{N}$, as required. \qed

### 3.4 Borel Probability Spaces

We conclude this lengthy introductory section with a brief discussion of Borel probability spaces and some of their interesting properties.

**Definition 9.** Suppose $(X, \mathcal{B}, \mu)$ is a measure space where $X$ is a compact metric space, $\mathcal{B}$ is the Borel $\sigma$-algebra generated by the metric topology on $X$ and $\mu$ a (Borel) probability measure. Then we say $(X, \mathcal{B}, \mu)$ is a Borel probability space.
Example 4. Note that all our previous examples of measure preserving systems are defined on Borel probability spaces. In particular, the state space $X = [n]^Z$ in the Bernoulli system $(X, \mathcal{B}, \mu, T)$ is a compact space by Tychonoff’s Theorem. Define the $\delta$ metric on $[n]$ by

$$\delta(a,b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{otherwise} \end{cases}$$

for all $a, b \in [n]$. Then let $d : X \times X \to [0,1]$ be given by

$$d(\alpha, \beta) = \frac{1}{3} \sum_{k \in \mathbb{Z}} \frac{\delta(\alpha_k, \beta_k)}{2^{\|k\|}}$$

for all $\alpha = (\alpha_t)_{t \in \mathbb{Z}}, \beta = (\beta_t)_{t \in \mathbb{Z}} \in X$. Then $d$ is a metric on $X$ and it is easy to see that the product topology on $X$ corresponds precisely to the metric topology induced by $d$.

There are numerous advantages in considering measure preserving systems on Borel probability space; they essentially constitute the class of ‘well behaved’ systems. Here we will highlight some of their favourable properties. We begin with a basic result from topology which will be used frequently throughout this study.

Proposition 3.14. Let $(X, d)$ be a compact metric space. Then the normed space $(C(X), \| \cdot \|_\infty)$ of real-valued continuous functions on $X$ with uniform norm is separable.

Proof. The rather elegant proof described here is adapted from ([5] Appendix B) and uses the Stone-Weierstrass Theorem (see [2]). Any compact metric space is separable so there exists some dense countable set $(x_n)_{n \in \mathbb{N}} \subset X$. Define the function $f_n : X \to \mathbb{R}$ by $f_n(x) = d(x_n, x)$ for each $n \in \mathbb{N}$. Each of these functions is continuous and by the density of $(x_n)_{n \in \mathbb{N}}$, the $f_n$ separate points. If $\mathcal{F}'$ denotes the algebra generated by $(f_n)_{n \in \mathbb{N}}$ then it follows that $\mathcal{F}'$ is dense in $C(X)$ by the Stone-Weierstrass Theorem. Let $\mathcal{F}$ denote the algebra generated by $(f_n)_{n \in \mathbb{N}}$ over $\mathbb{Q}$ (that is $\mathcal{F}$ is the smallest set containing $(f_n)_{n \in \mathbb{N}}$ which is closed under both (finite) $\mathbb{Q}$-linear combinations and multiplication). Then $\mathcal{F}$ is countable and dense in $\mathcal{F}'$. Thus $\mathcal{F}$ is a countable dense subset of $C(X)$. 

Definition 10. Let $X$ be a compact metric space and $\mathcal{B}$ the Borel $\sigma$-algebra on $X$. Denote by $\mathcal{M}(X)$ the set of Borel probability measures on $(X, \mathcal{B})$.

Proposition 3.15. If $(X, \mathcal{B}, \mu)$ is a Borel probability space then there exists a natural choice of topology on the set $\mathcal{M}(X)$ such that $\mathcal{M}(X)$ is a compact, metrizable topological space.

Proof. The details of the proof are sketched; for a more thorough discussion of the properties of $\mathcal{M}(X)$ see [18, 5].

By the Riesz Representation Theorem [18], the dual space $C(X)^*$ is identified with the space $\mathfrak{M}$ of finite signed measures on $(X, \mathcal{B})$. Explicitly, for all $\theta \in C(X)^*$ there exists a $\mu_\theta \in \mathfrak{M}$ such that

$$\theta(f) = \int f \, d\mu_\theta$$

for all $f \in C(X)$.

Each measure can be viewed as a linear functional and thus it makes sense to define the weak*-topology on $\mathfrak{M}$. This is the weakest topology such that for every $f \in C(X)$, the linear operator $\lambda \mapsto \int f \, d\lambda$ evaluating each functional at $f$ is continuous.

The set $\mathcal{M}(X)$ of Borel probability measures is then given the subspace topology induced by the weak*-topology on $\mathfrak{M}$. We claim that the topological space $\mathcal{M}(X)$ is compact and metrizable.

The proof that the space is metrizable follows by explicitly defining a metric on $\mathcal{M}(X)$. By the previous proposition $(C(X), \| \cdot \|_\infty)$ is separable and so there exists some dense sequence $(f_n)_{n \in \mathbb{N}} \subset C(X)$. Define a metric $d_{\mathcal{M}}$ on $\mathcal{M}(X)$ by

$$d_{\mathcal{M}}(\lambda, \kappa) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{1 + |\int f_n \, d\lambda - \int f_n \, d\kappa|}{1 + |\int f_n \, d\lambda - \int f_n \, d\kappa|}$$
The topology induced by the metric $d_M$ then corresponds precisely to the weak*-topology. Here we omit the details (for a full proof see [5], Appendix B).

To demonstrate compactness, recall the famous Alaoglu’s Theorem [2] states that the closed unit ball of the dual of a normed space is weak*-compact. Write

$$M(X) = \left\{ \lambda \in M : \int 1_X \, d\lambda = 1, \int f \, d\lambda \geq 0 \text{ for all } f \in C(X), f \geq 0 \right\}$$

It is then easy to see $M(X)$ is a subset of the closed unit ball of $M \cong C(X)^*$. Furthermore, if $(\lambda_n)_{n \in \mathbb{N}} \subseteq M(X)$ is a sequence of measures which converge to $\lambda \in M$ in the weak*-topology then

$$\int f \, d\lambda_n \to \int f \, d\lambda \quad \text{for all } f \in C(X)$$

Considering the case $f \geq 0$ and $f = 1_X$ it follows $\lambda \in M(X)$. Therefore $M(X)$ is a closed subspace of a compact space and so is itself compact. \hfill \square

The topological properties of $C(X)$ and $M(X)$ are of great benefit when studying measure preserving systems defined on Borel probability space. For this reason we will for the most part concern ourselves only with the theory of this specific class of systems. Exploiting their additional structure is crucial in our proof of Szemerédi’s Theorem. As an example of this, the following subsection describes the powerful Ergodic Decomposition Theorem which is a direct consequence of the geometric and topological structure of $M(X)$. The Ergodic Decomposition Theorem is important to us as it allows us to relate our study of ergodic systems to more general classes of measure preserving systems. Further, in Section 9 we use the properties of $C(X)$ and $M(X)$ to introduce the notions of conditional expectation and measure. These have a very significant role in our study.

### 3.5 Ergodic Decomposition

So far many of our major results are applicable only to ergodic systems. We shall briefly describe how the study of ergodic systems can be related to the study of systems in more generality.

As explained above, ergodicity is an indecomposability (or irreducibility) condition. An obvious question is whether it is possible to decompose an arbitrary system into irreducible components, akin to decomposing a positive integer into prime factors. A remarkable fact, and one of the major strengths of ergodic theory, is that for the class of systems defined on Borel probability space this is indeed possible. What we have described here is the content of the Ergodic Decomposition Theorem, which we will state (but refrain from proving) in this section.

We present a statement of the Ergodic Decomposition Theorem, adapted from [13].

**Theorem 3.16** (Ergodic Decomposition Theorem). Suppose $(X, \mathcal{B}, \mu, T)$ is a measure preserving system on a Borel probability space. Then there exists an almost everywhere defined function $x \mapsto \mu_x$ from $X$ to the space $M_T(X) \subseteq M(X)$ of $T$-invariant Borel probability measures on $(X, \mathcal{B})$ such that

1. $\mu_x$ is ergodic for almost every $x \in X$.
2. For any $f \in L^1(X, \mathcal{B}, \mu)$ we have $f \in L^1(X, \mathcal{B}, \mu_x)$ for almost every $x \in X$ and

$$\int f \, d\mu = \left( \int f \, d\mu_x \right) \, d\mu(x)$$

**Proof.** The details of the proof are omitted. See ([5] Chapter 4, [13] Chapter 2) for a more precise picture. The idea is to show that the space $M_T(X)$ is a compact, convex space and, moreover, that the ergodic measures are precisely the extreme points of $M_T(X)$. The decomposition theorem then follows by applying the Krein-Milman Theorem (see [2]) to show that any measure $\mu \in M_T(X)$ is a convex combination of ergodic measures. \hfill \square

18
Remark 7. Later in the discourse the idea of conditional measures is introduced. The construction of conditional measures generalises the Ergodic Decomposition Theorem and the latter result can be obtained from this construction (although it does not follow directly).

Essentially, Ergodic Decomposition tells us that any property which is preserved under integration and holds for ergodic systems, holds for systems in general. As we shall see later in our study, this is an important observation: it implies ergodicity is a far more reasonable condition to assume than it may first appear.

4 Furstenberg’s Correspondence Principle

In this section we will establish the connection between ergodic theory and Szemerédi’s Theorem. In particular we will show that a ‘generalised Poincaré Recurrence Theorem’ implies Szemerédi’s Theorem. In fact, the two results are equivalent but for the sake of brevity we will omit the proof of the converse (although it is not too difficult to establish).

Theorem 4.1. Let $(X, \mathcal{B}, \mu)$ be a probability space and $T$ a measure preserving transformation on $(X, \mathcal{B}, \mu)$. Then for any $A \in \mathcal{B}$ with $\mu(A) > 0$ and any $k \in \mathbb{N}$ there exists an $n \in \mathbb{N}$ such that

$$\mu\left(\bigcap_{j=0}^{k-1} T^{-jn} A\right) > 0$$

(4.1)

Remark 8. Note that the case $k = 2$ is given by the Poincaré Recurrence Theorem, specifically Corollary 3.4.

Proposition 4.2. Theorem 4.1 implies Szemerédi’s Theorem.

Remark 9. Before presenting a proof of Proposition 4.1 it is remarked that the connection described here is one instance of a more general scheme laid out by Furstenberg [7, 8] relating problems in combinatorial number theory to recurrence in ergodic theory. The idea, as explained in [7], is to exploit the similarities between asymptotic density in sets of integers and measure in a probability space. In particular, the shift map $S : \mathbb{Z} \to \mathbb{Z}$ given by $S : x \mapsto x + 1$ preserves the upper Banach density of sets and thus is analogous to a measure preserving transformation.

In short, in order to prove Proposition 4.2, following the argument given in [5, 10], we will take a set $\Lambda \subset \mathbb{Z}$ of positive upper Banach density and use the aforementioned analogies to construct a measure preserving system $(X, \mathcal{B}, \mu, T)$, the (hypothesised) dynamics of which reflect the structure of $\Lambda$.

Proof (of Proposition 4.2). Assuming Theorem 4.1, let $\Lambda \subset \mathbb{Z}$ be a subset of the integers with positive upper Banach density and fix some $k \in \mathbb{N}$. We will show $\Lambda$ contains an arithmetic progression of length $k$.

First consider the measure space $(X_0, \mathcal{B}_0)$ of Bernoulli trials on two letters. Explicitly, let $X_0 := \{0, 1\}^\mathbb{Z}$ be the space of all two-sided sequences in 0s and 1s given the product topology induced from the discrete topology on $\{0, 1\}$ and $\mathcal{B}_0$ the corresponding Borel $\sigma$-algebra. Equivalently, $\mathcal{B}_0$ is the $\sigma$-algebra generated by the collection of projections $\pi_i : \omega \mapsto \omega_i$ for all $i \in \mathbb{Z}$.

Let $T_0 : X_0 \to X_0$ be the left shift transform on $X_0$, that is

$$T_0(\omega_i)_{i \in \mathbb{Z}} = (\omega_{i+1})_{i \in \mathbb{Z}} \quad \text{for all } (\omega_i)_{i \in \mathbb{Z}} \in X_0$$

For each $i \in \mathbb{Z}$ let $A_i$ denote the cylinder set

$$A_i := \{\omega \in X_0 : \omega_i = 1\}$$

Let $\omega_\Lambda \in X_0$ be the characteristic function of $\Lambda$. Now for any $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ consider the length $k$ arithmetic progression $\{a + jb\}_{j=0}^{k-1}$ which starts at $a$ and has step size $b$. Note that

$$\bigcap_{j=0}^{k-1} T_0^{-(a+jb)} A_0 = \bigcap_{j=0}^{k-1} A_{a+jb}$$
and it follows that
\[ \{a + jb\}_{j=0}^{k-1} \subseteq \Lambda \iff \omega_{\Lambda} \in \bigcap_{j=0}^{k-1} T_0^{-(a+j)b} A_0 \]

We are interested only in showing there exists an arithmetic progression in \( \Lambda \) of length \( k \). In particular, the values of the starting point \( a \) and step size \( b \) are immaterial. Removing these conditions, it suffices to show
\[
\omega_{\Lambda} \in \bigcup_{a=-\infty}^{\infty} \bigcap_{b=1}^{\infty} \bigcap_{j=0}^{k-1} T_0^{-(a+j)b} A_0
\]
or, equivalently that there exists some \( a \in \mathbb{Z} \) such that
\[
T_0^a \omega_{\Lambda} \in \bigcup_{b=1}^{\infty} \bigcap_{j=0}^{k-1} T_0^{-jb} A_0 \tag{4.2}
\]

Prompted by (4.2) we restrict our attention to the (two-sided) orbit of \( \omega_{\Lambda} \) under \( T_0 \). Define
\[
X = \text{clos}\{T^a \omega_{\Lambda} : a \in \mathbb{Z}\}
\]
with the subspace topology induced from \( X_0 \) and let \( T \) be the restriction of \( T_0 \) to \( X_0 \). Consider the open set \( A = X \cap A_0 \). To see \( \Lambda \) contains an arithmetic progression of length \( k \) it suffices to show
\[
\bigcup_{b=1}^{\infty} \bigcap_{j=0}^{k-1} T_0^{-jb} A \neq \emptyset
\]
Indeed, if \( \omega \in \bigcap_{j=0}^{k-1} T_0^{-jb} A \) for some \( b \geq 1 \) then, as this set is open and \( \{T^a \omega_{\Lambda} : a \in \mathbb{Z}\} \) is dense in \( X \), there must be some \( a \in \mathbb{Z} \) for which (4.2) holds and we are done.

We claim, using the assumption \( \Lambda \) has positive upper Banach density, that we can construct a \( T \)-invariant measure \( \mu \) on \( X \) with the property \( \mu(A) > 0 \). Then by Theorem 4.1 there exists some \( b \in \mathbb{N} \) such that
\[
\mu\left( \bigcap_{j=0}^{k-1} T_0^{-jb} A \right) > 0
\]
and we are done.

We begin by constructing a sequence of ‘almost \( T \)-invariant’ measures on \( X \). By assumption
\[
\tilde{d}(\Lambda) = \limsup_{N-M \to \infty} \frac{\left| \Lambda \cap [M,N] \right|}{N-M} > 0
\]
Moreover, there exists a sequence of intervals \([a_n, b_n)\) with \( b_n - a_n \to \infty \) such that
\[
\lim_{n \to \infty} \frac{\left| \Lambda \cap [a_n, b_n) \right|}{b_n - a_n} = \tilde{d}(\Lambda) > 0 \tag{4.3}
\]
Define the measure
\[
\mu_n = \frac{1}{b_n - a_n} \sum_{j=a_n}^{b_n-1} \delta_{T^j \omega_{\Lambda}}
\]
where \( \delta_x \) denotes the delta measure
\[
\delta_x(\omega) = \begin{cases} 1 & \text{if } \omega = x \\ 0 & \text{otherwise} \end{cases}
\]
Then clearly each \( \mu_n \) is a probability measure. Moreover, it is easy to see
\[
\mu_n(A) = \frac{\left| \Lambda \cap [a_n, b_n) \right|}{b_n - a_n} \tag{4.4}
\]
Furthermore, the $\mu_n$ are ‘almost $T$-invariant’ in the sense that for all $B \in \mathcal{B}$
\[
|\mu_n(B) - \mu_n(T^{-1}B)| = \left| \frac{1}{b_n - a_n}(\delta_{T^{T_n} \omega_\Lambda} - \delta_{T^{T_n} \omega_\Lambda}) \right| 
\leq \frac{2}{b_n - a_n}
\]
so that $\lim_{n \to \infty} |\mu_n(B) - \mu_n(T^{-1}B)| = 0$.

In Proposition 3.15 we established that $\mathcal{M}(X)$ with the weak*-topology is a compact, metrizable space. Hence there must exist a weak*-subsequential limit $\mu$ of $(\mu_n)_{n \in \mathbb{N}}$. Clearly $\mu$ must be a Borel probability measure on $X$. By comparing (4.3) and (4.4) we conclude $\mu(A) > 0$. Finally, to see that $\mu$ is $T$-invariant, note by the definition of the weak*-limit for every $f \in C(X)$ we have
\[
\lim_{l \to \infty} \int T f - f \, d\mu_n = \int T f - f \, d\mu
\]
for some subsequence $(\mu_n)_l$. As the $\mu_n$ are almost $T$-invariant we conclude
\[
\int T f \, d\mu = \int f \, d\mu
\]

Given any $U \subseteq X$ open, there exists an increasing sequence of continuous functions $f_n \to 1_U$ pointwise. Then $(T f_n)_{n \in \mathbb{N}}$ is an increasing sequence of continuous functions such that $T f_n \to 1_{T^{-1} U}$ pointwise. Hence by the Monotone Convergence Theorem $\mu(T^{-1} U) = \mu(U)$. But the open sets form an algebra which generates $\mathcal{B}$, so $\mu$ is $T$-invariant.

5 Examples of Multiple Recurrence

In the previous section we deduced that in order to prove Szemerédi’s Theorem it suffices to show Theorem 4.1. Thus the remainder of this essay will concentrate on proving Theorem 4.1. We follow the argument described in Furstenberg, Katzenelson and Ornstein’s article [10] and first prove the result separately for the familiar examples of Bernoulli shifts and circle rotations. We will see that Theorem 4.1 holds in both cases, but for strikingly different reasons.

5.1 Multiple Recurrence for the Bernoulli and Circle Rotation Systems

Proposition 5.1. Let $(X, \mathcal{B}, \mu, T)$ be a Bernoulli system and $A_0, \ldots, A_{k-1} \in \mathcal{B}$. Then
\[
\lim_{n \to \infty} \left| \mu \left( \bigcap_{j=0}^{k-1} T^{-j n} A_j \right) - \prod_{j=0}^{k-1} \mu(A_j) \right| = 0
\]

Moreover, it follows that if $A \in \mathcal{B}$ with $\mu(A) > 0$ then
\[
\lim_{n \to \infty} \mu \left( \bigcap_{j=0}^{k-1} T^{-j n} A \right) = \mu(A)^{k-1} > 0
\]

Hence Theorem 4.1 holds for Bernoulli systems.

Proof. In the proof of Proposition 3.10, where we showed Bernoulli systems are ergodic, we demonstrated the case $k = 2$. It is easy to see that the general case can be obtained by a completely analogous method. First suppose $A_0, \ldots, A_{k-1} \in \mathcal{B}$ are all cylinder sets. Take

\[\text{[Footnote]} \text{ In fact we will prove the stronger Furstenberg Multiple Recurrence Theorem, which is introduced later in this section.}\]
n ∈ ℕ so large that the sets defining the co-ordinates for the \( T^jA_j \) are all disjoint. Then the \( T^jA_j \) are independent sets and

\[
\mu \left( \bigcap_{j=0}^{k-1} T^{-jn} A_j \right) = \prod_{j=0}^{k-1} \mu(A_j)
\]

The general case where \( A_0, \ldots, A_{k-1} \in \mathcal{B} \) are arbitrary measurable sets is now obtained by using approximating sequences of cylinder sets, as in Proposition 3.10.

The property of Bernoulli systems described in (5.1) is called ‘\( k \)-order mixing’. Here the asymptotic independence of the sets \( T^jA_j \) can be thought of as indicative of ‘random-like’ behaviour within the system. We will discuss this phenomenon in far more detail in the forthcoming section, but for now it is good to acknowledge that in this case Theorem 4.1 holds here due to the random-like behaviour of the system.

**Proposition 5.2.** Let \((X, \mathcal{B}, \mu, T)\) denote the system corresponding to rotations of the circle by \( \tau \). Then given any \( A \in \mathcal{B} \) with \( \mu(A) > 0 \) and any \( k \in \mathbb{N} \)

\[
\lim \inf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k-1} T^{-jn} A \right) > 0
\]

**Proof.** Fix \( A \in \mathcal{B} \) with \( \mu(A) > 0 \) and \( k \in \mathbb{N} \).

If \( \tau \) is rational then there exists some \( q \in \mathbb{Z} \) such that \( T^{-q}A = A \) and the result follows immediately.

For the case \( \tau \) irrational, by Lemma 3.8 there exists a finite sequence of (non-empty) disjoint intervals \( I_1, \ldots, I_N \) such that, writing \( H = \bigcup_{j=1}^{N} I_j \), we have \( A \subseteq H \) and \( \mu(H \setminus A) < \frac{\mu(A)}{4k} \).

Now suppose \( y \in \mathbb{T} \) with \( |y| < \delta \) where

\[
\delta = \min \left( \left\{ \mu(I_j) / (k-1) : j = 0, \ldots, N \right\} \cup \left\{ \mu(A) / 4(k-1) \right\} \right)
\]

Then

\[
\mu \left( \bigcap_{j=0}^{k-1} H - jy \right) \geq \mu(H \cap (H - (k-1)y)) > \mu(H) - \frac{\mu(A)}{4}
\]

Now for each \( 0 \leq j \leq k-1 \) we have \( A - jy \subseteq H - jy \). Note that

\[
\mu(A \cap (A - y)) \geq \mu(H \cap (H - y)) - \mu(H \setminus A) - \mu((H - y) \setminus (A - y))
\]

By a simple induction

\[
\mu \left( \bigcap_{j=0}^{k-1} A - jy \right) \geq \mu \left( \bigcap_{j=0}^{k-1} H - jy \right) - \sum_{l=0}^{k-1} \mu((H - jy) \setminus (A - jy))
\]

Furthermore, \( \mu((H - jy) \setminus (A - jy)) = \mu(H \setminus A) < \frac{\mu(A)}{4k} \) for each \( 0 \leq j \leq k-1 \) and so

\[
\mu \left( \bigcap_{j=0}^{k-1} A - jy \right) \geq \mu \left( \bigcap_{j=0}^{k-1} H - jy \right) - \frac{\mu(A)}{4}
\]

\[
\geq \frac{\mu(A)}{2} > 0
\]

To finish we claim that the set \( \Lambda = \{ n \in \mathbb{N} : \tau n \in (-\delta, \delta) \} \) has positive lower asymptotic density, that is

\[
\lim \inf_{N \to \infty} \frac{|\Lambda \cap [1, N]|}{N} > 0
\]
It then follows that
$$\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcup_{j=0}^{k-1} T^{-jn} A \right) \geq \liminf_{N \to \infty} \frac{\mu(A) |\Lambda \cap [1,N]|}{2} / N > 0$$

To prove the claim we show $\Lambda$ is syndetic: that there exists some $a \in \mathbb{N}$ such that $\{k, k+1, \ldots, k+a-1\} \cap \Lambda \neq \emptyset$ for all $k \in \mathbb{N}$. As $\tau$ is irrational, by Proposition 3.1 the orbit of 0 is dense in $\mathbb{T}$. Thus there exists some $a \in \mathbb{N}$ such that the set $\{0, \tau, \ldots, (a-1)\}$ is $2\delta$-dense in $\mathbb{T}$. Moreover, for any $k \in \mathbb{N}$ the translated set
$$\{k\tau, (k+1)\tau, \ldots, (k+a-1)\tau\}$$
is $2\delta$-dense in $\mathbb{T}$. It follows that for every $k \in \mathbb{N}$, the set (5.2) of $a$ consecutive points of the orbit of 0 contains at least one element in $(-\delta, \delta)$ and $\Lambda$ is syndetic.

The reason why Theorem 4.1 holds in the case of rotations is transparent: the rotations do not move nearby points apart. This is a markedly different situation than that of the Bernoulli shifts; here we have an example structured, as opposed to random-like behaviour.

### 5.2 Furstenberg’s Multiple Recurrence Theorem

In both Proposition 5.1 and 5.2 in order to verify Theorem 4.1 held in the respective cases we in fact demonstrated a stronger result. This suggests that it is more natural to attempt to prove a stronger statement.

**Theorem 5.3** (Furstenberg’s Multiple Recurrence Theorem). Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for any $k \in \mathbb{N}$
$$\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k-1} T^{-jn} A \right) > 0$$

The method by which we prove Theorem 5.3 is rather complicated; before we begin it is prudent to briefly sketch the first few steps of the argument. The proceeding section is occupied with determining the property of the Bernoulli system which accounts for the Multiple Recurrence Theorem holding in this instance. It transpires that Bernoulli systems belong a larger class of ‘weak mixing’ systems. In Section 6 we prove multiple recurrence holds for all weak mixing systems owing to their ‘random-like’ behaviour. On the other hand, rotations of the circle exemplifies what is known as a ‘compact system’. Compact, in contrast to weak mixing systems, exhibit ‘structured’ behaviour. In Section 7 we show multiple recurrence holds for all compact systems owing to their structured behaviour.

The notions of weak mixing and compactness are mutually exclusive and in terms of randomness and structure they represent two extremes in behaviour. To show that the Multiple Recurrence Theorem holds for these two cases is a significant step toward the full proof: at the very least, as noted in [5], the fact that the result should hold for these two opposite classes of systems could be interpreted as an indication that we are dealing with a more general phenomenon.

It is important to bear in mind that there are many examples of systems which are neither weak mixing nor compact. Importantly, however, we will see that there is a dichotomy between weak mixing and compact systems. Specifically, if a system is not weak mixing then it must contain some compact ‘subsystem’. The ‘subsystems’ in question are correctly known as factors. Thus the combined result of studying weak mixing and compact systems and demonstrating this dichotomy will be the following:

Given any measure preserving system, the Multiple Recurrence Theorem holds for some (non-trivial) factor.

This is the first significant step toward the full proof.
5.3 Reducing the Problem

Throughout the remainder of this discourse we will make a number of reasonable assumptions concerning the measure preserving systems we consider. The reader is reminded that all our measure preserving systems are defined on probability space: it is necessary that the systems are finite for multiple recurrence to hold.

In addition we will (tacitly) assume that all measure preserving systems we consider are both invertible and defined on Borel probability space. By careful consideration of the proof of the correspondence result Proposition 4.2, to deduce Szemerédi’s Theorem it suffices only to show the Multiple Recurrence Theorem for this class of systems.

Moreover, the general case follows fairly easily once Multiple Recurrence has been established for such systems, but here we omit the proof of this fact (for the details, see [5, 8]).

We claim that without loss of generality we may only consider ergodic systems. This is an application of the Ergodic Decomposition Theorem.

Proposition 5.4. Suppose multiple recurrence holds for all invertible, ergodic systems defined on Borel probability space. Then it holds for any invertible system \((X, \mathcal{B}, \mu, T)\) defined on a Borel probability space.

Proof. Consider \((X, \mathcal{B}, \mu, T)\) where \((X, \mathcal{B}, \mu)\) is a Borel probability space. By Ergodic Decomposition there exists some \(X' \in \mathcal{B}\) with \(\mu(X') = 1\) and a collection of probability measures \(\{\mu_x : x \in X'\}\) on \((X, \mathcal{B})\) such that each \(\mu_x\) is an ergodic measure for \(T\). Moreover, for all \(f \in L^1(X, \mathcal{B}, \mu)\) we have \(f \in L^1(X, \mathcal{B}, \mu_x)\) for almost all \(x\) and

\[
\int f \, d\mu = \int \left( \int f \, d\mu_x \right) \, d\mu(x)
\]

In particular, given \(A \in \mathcal{B}\) with \(\mu(A) > 0\) and \(k \in \mathbb{N}\) we have

\[
\int \mu_x(A) \, d\mu(x) = \mu(A) > 0
\]

and so \(\mu(\{x \in X : \mu_x(A) > 0\}) > 0\). By Fatou’s lemma,

\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu_x \left( \bigcap_{j=0}^{k-1} T^{-jn} A \right) = \liminf_{N \to \infty} \int \frac{1}{N} \sum_{n=1}^{N} \mu_x \left( \bigcap_{j=0}^{k-1} T^{-jn} A \right) \, d\mu(x) = \int \liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu_x \left( \bigcap_{j=0}^{k-1} T^{-jn} A \right) \, d\mu(x) > 0
\]

where we have applied the assumption that the Multiple Recurrence Theorem holds for ergodic systems.

6 Weak Mixing Systems

As discussed earlier, the multiple recurrence result holds for Bernoulli system owing to the ‘random-like’ behaviour the system exhibits. In this section we discuss and define precisely what is meant by ‘random-like’ behaviour, and conclude that in general the multiple recurrence result holds for the general class of systems with this property.

Definition 11. A measure preserving system \((X, \mathcal{B}, \mu, T)\) is mixing if for all \(A, B \in \mathcal{B}\) the following holds:

\[
\lim_{n \to \infty} \mu(A \cap T^{-n} B) = \mu(A) \mu(B) \quad (6.1)
\]
It is natural to think of a mixing system as one which exhibits random-like behaviour. To see why, assume \( \mu(B) > 0 \) and write (6.1) as
\[
\lim_{n \to \infty} \left| \frac{\mu(A \cap T^{-n}B)}{\mu(T^{-n}B)} - \frac{\mu(A)}{\mu(B)} \right| = 0
\]
where we have used the fact \( T \) is measure preserving. Then \( \frac{\mu(A \cap T^{-n}B)}{\mu(T^{-n}B)} \) is the conditional probability that \( x \in A \) given \( T^n(x) \in B \); namely, the probability of the event \( A \) happening now, given that \( B \) will occur \( n \) intervals of time into the future. Hence a mixing system is characterised by the property that for any pair of events \( A \) and \( B \), the events along the orbit of \( B \) asymptotically bear no relation to \( A \).

**Remark 10.** Whilst we are primarily concerned with studying mixing with regard to multiple recurrence, it is worthwhile to note that the consideration of mixing systems naturally emanates from the notion of ergodicity and the Mean Ergodic Theorem. As discussed in ([5] Chapter 2; [16]), by the Mean Ergodic Theorem (Corollary 3.12) a measure preserving system \( (X, \mathcal{B}, \mu, T) \) is ergodic if and only if for all \( f, g \in L^2(X, \mathcal{B}, \mu) \) we have
\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \int fT^n g \, d\mu - \int f \, d\mu \int g \, d\mu \right|^2 = 0
\]
By considering characteristic functions this is equivalent to the property that given any \( A, B \in \mathcal{B} \)
\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} \mu(A \cap T^{-n}B) - \mu(A)\mu(B) \right|^2 = 0
\]
This convergence may take place for a number of reasons, one of which being that the sets \( A \) and \( T^{-n}B \) are asymptotically independent, motivating the notion of mixing.

It transpires that the mixing condition is rather strong. A weaker and more malleable notion is that of weak-mixing, and it is with weak-mixing systems we will concern ourselves.

**Definition 12.** A measure preserving system \( (X, \mathcal{B}, \mu, T) \) is weak mixing if for all \( A, B \in \mathcal{B} \) the following holds:
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \mu(A \cap T^{-n}B) - \mu(A)\mu(B) \right]^2 = 0 \tag{6.2}
\]
Note that, trivially, weak-mixing systems are ergodic.

**Lemma 6.1.** A weak mixing system \( (X, \mathcal{B}, \mu, T) \) is ergodic.

**Proof.** Suppose \( B \in \mathcal{B} \) is \( T \)-invariant. Then by the weak mixing property
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \mu(B \cap T^{-n}B) - \mu(B) \right]^2 = 0
\]
For all \( n \in \mathbb{N} \) we have \( T^{-n}B = B \) and so it follows \( \mu(B) = \mu(B)^2 \). Hence either \( B \) or \( B^c \) is a null set.

We can define weak mixing in terms of functions.

**Lemma 6.2.** A system \( (X, \mathcal{B}, \mu, T) \) is weak mixing if and only if for all \( f, g \in L^2(X, \mathcal{B}, \mu) \)
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \int fT^n g \, d\mu - \int f \, d\mu \int g \, d\mu \right]^2 = 0 \tag{6.3}
\]
Proof. Clearly by taking \( f = 1_A \) and \( g = 1_B \) in (6.3) gives (6.2). Conversely, by assuming weak mixing we assume (6.3) holds for all characteristic functions. These functions span the subspace of simple functions, dense in \( L^2(X, \mathcal{B}, \mu) \). In general if (6.3) holds for all \( f, g \in S \) where \( S \) is a set linearly spanning a dense subspace \( V \), then (6.3) holds for all \( f, g \in L^2(X, \mathcal{B}, \mu) \). This fact follows a simple approximation argument, similar to that used in Corollary 3.13.

There are many equivalent formulations of the weak-mixing property\(^{11}\) which can been seen as an indication that the concept is a natural one. Presently we introduce some of the formulations which are necessary to our study.

**Proposition 6.3.** A system \((X, \mathcal{B}, \mu, T)\) is weak mixing if and only if the product system \((X \times X, \mathcal{B} \otimes \mathcal{B}, \mu \otimes \mu, T \otimes T)\) is ergodic.

**Proof.** Suppose \((X, \mathcal{B}, \mu, T)\) is weak mixing. By Lemma 6.1 it suffices to show the product system is weak mixing. Furthermore, by the proof of Lemma 6.2, it suffices to show that if \( f_1, f_2, g_1, g_2 \in L^2(X, \mathcal{B}, \mu) \) then the weak mixing property (6.3) holds for \( f_1 \otimes f_2 \) and \( g_1 \otimes g_2 \in L^2(X \times X) \) as functions of this type span a dense subspace of \( L^2(X \times X) \). Note that

\[
\int f_1 \otimes f_2 (T \times T)^n g_1 \otimes g_2 \, d(\mu \otimes \mu) = \int f_1 T^n g_1 \, d\mu \int f_2 T^n g_2 \, d\mu
\]

Thus the result is easily deduced through an application of the triangle inequality.

Conversely, suppose the product system is ergodic; it is easy to see that \((X, \mathcal{B}, \mu, T)\) must also be ergodic. For suppose \( B \in \mathcal{B} \) is a \( T \)-invariant set. Then \( B \times B \in \mathcal{B} \otimes \mathcal{B} \) is \( T \times T \)-invariant and hence by ergodicity \( \mu(B)^2 = \mu \otimes \mu(B \times B) = 0 \) or 1. Thus \( B \) is null or has null complement.

Let \( f, g \in L^2(X, \mathcal{B}, \mu) \). By the Mean Ergodic Theorem,

\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \int f T^n g \, d\mu - \int f \, d\mu \int g \, d\mu \right| = 0
\]

(6.4)

As the product system is ergodic we can carry out exactly the same proceedure on \((X \times X, \mathcal{B} \otimes \mathcal{B}, \mu \otimes \mu, T \otimes T)\) and the functions \( g \otimes g \) and \( f \otimes f \) to give

\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \left( \left( \int f T^n g \, d\mu \right)^2 - \left( \int f \, d\mu \right)^2 \left( \int g \, d\mu \right)^2 \right) \right| = 0
\]

(6.5)

We claim (6.4) and (6.5) together imply the weak mixing property, namely

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \int f T^n g \, d\mu - \int f \, d\mu \int g \, d\mu \right] = 0
\]

(6.6)

Let \( a_n = \int f T^n g \, d\mu \) and \( b = \int f \, d\mu \int g \, d\mu \). Then the sequence \((a_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}\) has the properties

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} a_n = b \quad \text{and} \quad \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} a_n^2 = b^2
\]

It follows that

\[
\frac{1}{N} \sum_{n=1}^{N} (a_n - b)^2 = \frac{1}{N} \sum_{n=1}^{N} a_n^2 - \frac{2b}{N} \sum_{n=1}^{N} a_n + b^2
\]

\(^{11}\)For lists and proofs of equivalences see [1, 5, 8].
Taking the limit as $N \to \infty$ we have
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} (a_n - b)^2 = 0
\]
which is precisely (6.6), as required.

**Definition 13.** A measure preserving system $(X, \mathcal{B}, \mu, T)$ is weak mixing of all orders if for any $k \in \mathbb{N}$, given a collection $A_0, \ldots, A_k \in \mathcal{B}$ of measurable sets
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \mu \left( \bigcap_{j=0}^{k} T^{-jn} A_j \right) - \prod_{j=0}^{k} \mu(A_j) \right]^2 = 0 \quad (6.7)
\]

We can instantly deduce multiple recurrence holds for any system which is weak mixing of all orders: for any $A \in \mathcal{B}$ of positive measure and any $k \in \mathbb{N}$ let $A_0 = A_1 = \cdots = A_k = A$ and apply (6.7) to obtain
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k} T^{-jn} A \right) = \mu(A)^{k+1} > 0
\]

Presently we demonstrate that the ostensibly stronger condition of weak mixing of all orders is in fact equivalent to the weak mixing condition. Thus we will have shown that for any weak mixing system multiple recurrence holds.

**Remark 11.** Analogous to weak mixing of all orders, there are several possible definitions of mixing of all orders. One such, given in ([10], §3), is that a system $(X, \mathcal{B}, \mu, T)$ is mixing of all orders if for any collection $A_0, \ldots, A_k \in \mathcal{B}$ we have
\[
\lim_{|n_j - n_i| \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_n + h, u_n \rangle \right| = \mu(A_0) \mu(A_1) \mu(A_2)
\]

Theorem 6.4 then follows by applying this result to characteristic functions. To prove (6.8) we use induction on $k \in \mathbb{N}$. As the system is weak mixing, applying Lemma 6.2 gives the case $k = 1$. In order to prove the inductive step we will need to introduce the van der Corput trick, the statement and proof of which are adapted from ([5] Chapter 7; [1]).

**Theorem 6.4.** Suppose $(X, \mathcal{B}, \mu, T)$ is a weak mixing system. Then it is weak mixing of all orders.

Assuming $(X, \mathcal{B}, \mu, T)$ is weak mixing, we will in fact show that for any finite collection $f_0, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu)$ the following holds:
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \int \prod_{l=0}^{k} T^{ln} f_l \, d\mu - \int f_0 \, d\mu \right]^2 = 0 \quad (6.8)
\]

Theorem 6.4 then follows by applying this result to characteristic functions. To prove (6.8) we use induction on $k \in \mathbb{N}$. As the system is weak mixing, applying Lemma 6.2 gives the case $k = 1$. In order to prove the inductive step we will need to introduce the van der Corput trick, the statement and proof of which are adapted from ([5] Chapter 7; [1]).

**Lemma 6.5** (Van der Corput Trick). Let $(u_n)_{n \in \mathbb{N}}$ be a sequence in some Hilbert Space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$. Define a sequence $(s_h)_{h \in \mathbb{N}_0} \subset \mathcal{H}$ by
\[
s_h = \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle \right|
\]

---

12For a brief discussion of this problem see [12].
If
\[
\lim_{H \to \infty} \frac{1}{H} \sum_{h=0}^{H-1} s_h = 0
\]  
(6.9)
then
\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} u_n \right\| = 0
\]

Proof. First note that for fixed \( H \), the sums \( \frac{1}{N} \sum_{n=1}^{N} u_n \) and \( \frac{1}{HN} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} \) are arbitrarily close for large \( N \). Indeed, the sums overlap: they agree except on the first and last \( H \) terms.

\[
\frac{1}{H N} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} = \frac{1}{N} \left( \sum_{n=1}^{N} u_n + \sum_{n=2}^{N+1} u_n + \cdots + \sum_{n=H+1}^{N+H-1} u_n \right)
\]

As \( H \) is fixed and the sequence \( (u_n)_{n \in \mathbb{N}} \) is bounded, the difference between the two sums is averaged out as \( N \to \infty \). Explicitly, if \( M > 0 \) is such that \( \|u_n\| \leq M \) for all \( n \in \mathbb{N} \) it follows

\[
\left\| \frac{1}{H N} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} - \frac{1}{N} \sum_{n=1}^{N} u_n \right\| \leq \frac{M(2H+1)}{N} = o(1)
\]  
(6.10)

Now it suffices to consider the asymptotic behaviour of the norm of the double average. This is preferable as one may estimate the limit using terms of \( (s_h)_{h \in \mathbb{N}} \). By the triangle inequality

\[
\limsup_{N \to \infty} \left\| \frac{1}{H N} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} \right\|^2 \leq \limsup_{N \to \infty} \left( \frac{1}{N} \sum_{n=1}^{N} \left\| \frac{1}{H} \sum_{h=0}^{H-1} u_{n+h} \right\|^2 \right)
\]

\[
= \limsup_{N \to \infty} \left( \frac{1}{N} \sum_{n=1}^{N} \frac{1}{H^2} \sum_{h,h'=0}^{H-1} \langle u_{n+h}, u_{n+h'} \rangle \right)
\]

\[
\leq \frac{1}{H^2} \sum_{h,h'=0}^{H-1} \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_{n+h'} \rangle \right|
\]

Notice that if \( h' \geq h \) then

\[
\limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_{n+h'} \rangle \right| = \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N+h} \langle u_n, u_{n+h'-h} \rangle \right|
\]

\[
= \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_{n+h'-h} \rangle \right| = s_{h'-h}
\]

By a symmetric argument if \( h' \geq h \) then the left hand side is equal to \( s_{n-h'} \). In short

\[
\limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_{n+h'} \rangle \right| = s_{|h-h'|}
\]

Thus

\[
\limsup_{N \to \infty} \left\| \frac{1}{H N} \sum_{n=1}^{N} \sum_{h=0}^{H-1} u_{n+h} \right\|^2 \leq \frac{1}{H^2} \sum_{h,h'=0}^{H-1} s_{|h-h'|}
\]

28
Finally we split the sum \( \frac{1}{H} \sum_{h,h'=0}^{H-1} s_{|h-h'|} \) into three parts and use the hypothesis (6.9) to estimate. Let \( \epsilon > 0 \) be given. Choose \( H_0 \in \mathbb{N} \) such that

\[
\frac{1}{H} \sum_{h=0}^{H-1} s_h < \frac{\epsilon}{4} \quad \text{for all } H \geq H_0.
\] (6.11)

For any \( H \geq H_0 \) we may write

\[
\frac{1}{H^2} \sum_{h,h'=0}^{H-1} s_{|h-h'|} = \frac{1}{H} \sum_{h=0}^{H-H_0} \frac{1}{H} \sum_{h'=h}^{H-1} s_{h'-h} + \frac{1}{H} \sum_{h=0}^{H-H_0} \frac{1}{H} \sum_{h'=h+1}^{H-1} s_{h-h'} + \frac{1}{H^2} \sum_{h,h'=H-H_0+1}^{H-1} s_{|h'-h|}
\] (6.12)

\[
+ \frac{1}{H} \sum_{h'=H-H_0+1}^{H-1} s_{|h'|} < \frac{\epsilon}{4}
\] (6.13)

\[
+ \frac{1}{H^2} \sum_{h,h'=H-H_0+1}^{H-1} s_{|h'-h|}
\] (6.14)

Applying (6.11) we deduce that for all \( H_0 \leq h, h' \leq H \)

\[
\frac{1}{H} \sum_{h'=h}^{H-1} s_{h'-h} < \frac{\epsilon}{4}
\]

so that the sum in (6.12) is bounded above by \( \frac{\epsilon}{4} \). Similarly the sum in (6.13) is also bounded above by \( \frac{\epsilon}{4} \). As the \( s_h \) are bounded we may fix \( H_0 \) large so that the third sum (6.14) is bounded above by \( \frac{\epsilon}{4} \). Now if we choose \( N_0 \in \mathbb{N} \) so large that \( \frac{M(2H+1)}{N_0} < \frac{\epsilon}{4} \) then by (6.10) we conclude

\[
\| \frac{1}{N} \sum_{n=1}^{N} u_n \| < \epsilon \quad \text{for all } N \geq N_0
\]

As the choice of \( \epsilon > 0 \) was arbitrary, the result is established.

The Van der Corput trick is the central ingredient in the inductive step in the proof of Theorem 6.4 which we now turn to. Here we follow the method outlined in ([10], §3).

**Proof (of Theorem 6.4).** We shall show that for all \( f_0, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu) \) the following two statements hold:

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \prod_{l=0}^{k} T^{ln} f_l \right] = 0
\] (6.15)

and

\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \left[ \prod_{l=1}^{k} T^{ln} f_l \right] \right\|_{L^2(\mu)} = 0
\] (6.16)

Here we use an inductive argument. Recall that (6.15) case \( k = 1 \) holds by the hypothesis that the system is weak mixing. We show that for \( k \geq 2 \)

- **Claim (1):** (6.15) case \( k - 1 \Rightarrow (6.16) \) case \( k \)
- **Claim (2):** (6.16) case \( k \Rightarrow (6.15) \) case \( k \)

and conclude that both results hold for all \( k \in \mathbb{N} \).

**Claim (1).** Suppose every weak mixing system \( (X, \mathcal{B}, \mu, T) \) has the property that for all \( g_0, \ldots, g_{k-1} \in L^\infty(X, \mathcal{B}, \mu) \) the following holds:

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \prod_{j=0}^{k-1} T^{jn} g_j \right] = 0
\] (6.17)
Then for a fixed weak mixing system \((X, \mathcal{B}, \mu, T)\), given \(f_1, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu)\) we have

\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{l_n} f_l - \prod_{l=1}^{k} \int f_l \, d\mu \right\|_{L^2(\mu)} = 0
\]

**Proof.** First consider the case that for some \(j_0 \in 1, \ldots, k\) the function \(f_{j_0}\) has zero expectation: \(\int f_{j_0} \, d\mu = 0\). In this case we are required to prove

\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{l_n} f_l \right\|_{L^2(\mu)} = 0 \tag{6.18}
\]

To do this we use the van der Corput trick: let \((u_n)_{n \in \mathbb{N}} \subset L^\infty(X, \mathcal{B}, \mu)\) be the sequence defined by

\[
u_n = \prod_{l=1}^{k} T^{l_n} f_l
\]

Note that (6.18) can then be written as

\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \nu_n \right\|_{L^2(\mu)} = 0 \tag{6.19}
\]

and by van der Corput, to prove (6.19) it suffices to show

\[
\lim_{H \to \infty} \frac{1}{H} \sum_{h=0}^{H-1} s_h = 0 \tag{6.20}
\]

where

\[
s_h = \lim \sup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle \right|
\]

Now, using the fact \(T\) is measure preserving, we have

\[
\langle u_{n+h}, u_n \rangle = \int \prod_{l=1}^{k} T^{(n+h)} f_l \prod_{l=1}^{k} T^{l_n} f_l \, d\mu = \int \prod_{l=1}^{k} T^{l_n} (f_l T^{lh} f_l) \, d\mu
\]

\[
= \int T^n \prod_{l=0}^{k-1} T^{l_n} (f_{l+1} T^{(l+1)h} f_{l+1}) \, d\mu
\]

\[
= \int \prod_{l=0}^{k-1} T^{l_n} (f_{l+1} T^{(l+1)h} f_{l+1}) \, d\mu
\]

For each \(h \in \mathbb{N}\) and each \(l = 0, \ldots, k-1\) define \(g_{h,l} = f_{l+1} T^{(l+1)h} f_{l+1} \in L^\infty(X, \mathcal{B}, \mu)\). By applying the hypothesis (6.17) of the claim to the functions \(g_{h,l}\) for a fixed \(h\),

\[
s_h = \lim \sup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle \right|
\]

\[
= \lim \sup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \int \prod_{l=0}^{k-1} g_{h,l} \, d\mu \right|
\]

\[
= \prod_{l=0}^{k-1} \int g_{h,l} \, d\mu
\]
Now we show that
\[
\lim_{h \to \infty, h \notin Z} s_h = 0 \quad (6.21)
\]
where \( Z \subset \mathbb{N} \) is a set of zero density and hence, recalling Theorem 2.1, establish (6.20). Indeed, by the weak mixing property applied to the function \( f_{j_0} \),
\[
\lim_{H \to \infty} \frac{1}{H} \sum_{h=1}^{H} \left[ \int f_{j_0} T^h f_{j_0} \, d\mu - \left( \int f_{j_0} \, d\mu \right)^2 \right] = 0
\]
Recall by hypothesis \( \int f_{j_0} \, d\mu = 0 \). Summing over a subsequence gives
\[
\lim_{H \to \infty} \frac{1}{H} \sum_{h=1}^{H} \left[ \int f_{j_0} T^h \frac{f_{j_0}}{a_{j_0}} \, d\mu \right] = 0
\]
Once again recalling Theorem 2.1, there exists some set \( Z \subset \mathbb{N} \) of zero density such that
\[
\lim_{h \to \infty, h \notin Z} \int g_{h,j_0-1} \, d\mu = 0
\]
Noting that \( \|g_{h,l}\|_{\infty} \leq \max_{j=1,...,k} \|f_j\|_{\infty} \) is bounded uniformly in both \( h \) and \( l \),
\[
\lim_{h \to \infty, h \notin Z} \prod_{l=0}^{k-1} \int g_{h,l} \, d\mu = 0
\]
Thus we have (6.21) and consequently (6.18).

The general case reduces to the above case by noting the following lemma:

**Lemma 6.6.** Let \( (a_i)_{i=1}^{k}, (b_i)_{i=1}^{k} \subset \mathbb{R} \) be finite sequences of real numbers of length \( k \). Then
\[
\prod_{i=1}^{k} a_i \prod_{i=1}^{k} b_i = \sum_{j=1}^{k} \left( \prod_{l=1}^{j-1} a_l \right) (a_j - b_j) \left( \prod_{l=j+1}^{k} b_l \right)
\]

**Proof.** We induct on \( k \). The case \( k = 1 \) is trivial and for \( c_1, c_2, d_1, d_2 \in \mathbb{R} \) we have
\[
c_1c_2 - d_1d_2 = (c_1 - d_1)d_2 + c_1(c_2 - d_2) \quad (6.22)
\]
Assume the result holds for the case \( k - 1 \) and let \( (a_i)_{i=1}^{k}, (b_i)_{i=1}^{k} \subset \mathbb{R} \) be finite sequences of real numbers of length \( k \). Then
\[
\sum_{j=1}^{k} \left( \prod_{l=1}^{j-1} a_l \right) (a_j - b_j) \left( \prod_{l=j+1}^{k} b_l \right) = \sum_{j=1}^{k} \left( \prod_{l=1}^{j-1} a_l \right) (a_j - b_j) \left( \prod_{l=j+1}^{k-1} b_l \right) b_k + \prod_{l=1}^{k-1} a_l (a_k - b_k)
\]
By applying the induction hypothesis and setting \( c_1 = \prod_{i=1}^{k-1} a_i, c_2 = a_k \) and \( d_1 = \prod_{i=1}^{k-1} b_i, d_2 = b_k \) we reduce the problem to the case \( k = 2 \) and thus we are done by (6.22).

Given \( f_1, \ldots, f_k \in L^{\infty}(X, \mathcal{B}, \mu) \) with arbitrary expectation, pointwise application of the above lemma gives:
\[
\prod_{l=1}^{k} T^{jn} f_l - \prod_{l=1}^{k} \int f_l \, d\mu = \sum_{j=1}^{k} \left( \prod_{l=1}^{j-1} T^{jn} f_l \right) \left( T^{jn} f_j - \int f_j \, d\mu \right) \left( \prod_{l=j+1}^{k} \int f_l \, d\mu \right) \quad (6.23)
\]
Now for each \( j = 1, \ldots, k \) define
\[
g_{j,l} := \begin{cases} f_l & \text{for } l = 1, \ldots, j - 1 \\ f_j - \int f_j \, d\mu & \text{for } l = j \\ \int f_l \, d\mu & \text{for } l = j + 1, \ldots, k \end{cases}
\]
31
Then for a fixed weak mixing system $(X, \mathcal{B}, \mu, T)$ the function $g_{j,x}$ has zero expectation and so we may apply the previous case to each set of functions $g_{j,1}, \ldots, g_{j,k} \in L^\infty(X, \mathcal{B}, \mu)$. Moreover,

$$ \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{ln} f_l - \prod_{l=1}^{k} \int f_l \, d\mu \right\|_{L^2(\mu)} = \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{ln} g_{j,l} \right\|_{L^2(\mu)} \leq \sum_{j=1}^{k} \left\| \frac{1}{N} \sum_{n=1}^{N} T^{ln} g_{j,l} \right\|_{L^2(\mu)}$$

And so the first claim is established.

The second claim is somewhat easier to deduce.

**Claim (2).** Suppose that every weak mixing system $(X, \mathcal{B}, \mu, T)$ has the property that for all $f_1, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu)$ the following holds:

$$ \lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{ln} f_l - \prod_{l=1}^{k} \int f_l \, d\mu \right\|_{L^2(\mu)} = 0 \quad (6.24)$$

Then for a fixed weak mixing system $(X, \mathcal{B}, \mu, T)$, given $g_0, \ldots, g_k \in L^\infty(X, \mathcal{B}, \mu)$ we have

$$ \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left[ \int \prod_{l=0}^{k} T^{ln} f_l \, d\mu - \int \prod_{l=0}^{k} f_l \, d\mu \right]^2 = 0 $$

Indeed, given $g_0, \ldots, g_k \in L^\infty(X, \mathcal{B}, \mu)$, by the hypothesis (6.24) of the claim

$$ \lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{ln} g_l - \prod_{l=1}^{k} \int g_l \, d\mu \right\|_{L^2(\mu)} = 0 $$

Now norm convergence implies weak convergence and so by considering the linear functional $f \mapsto \int f g_0 \, d\mu$ on $L^\infty(X, \mathcal{B}, \mu)$ we have

$$ \lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \int \prod_{l=0}^{k} T^{ln} g_l \, d\mu - \int \prod_{l=0}^{k} g_l \, d\mu \right\| = 0 $$

In the proof of Proposition 6.3 we demonstrated that the product system $(X \times X, \mathcal{B} \otimes \mathcal{B}, \mu \otimes \mu, T \otimes T)$ is also weak mixing and hence we may applying the foregoing reasoning to this system and the functions $g_l \otimes g_l$ for $l = 0, \ldots, k$. The result of this is that

$$ \lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \left( \int \prod_{l=0}^{k} T^{ln} g_l \, d\mu \right)^2 - \left( \int \prod_{l=0}^{k} g_l \, d\mu \right)^2 \right\| = 0 $$

By precisely the same argument as used in the end of Proposition 6.3, taking this time $a_n = \int \prod_{l=0}^{k} T^{ln} g_l \, d\mu$ and $b = \prod_{l=0}^{k} \int g_l \, d\mu$ we obtain the desired result.

Thus we have obtained our first general result toward the proof of the Multiple Recurrence Theorem.

**Theorem 6.7.** The Multiple Recurrence Theorem holds for any weak mixing system. More precisely, if $(X, \mathcal{B}, \mu, T)$ is a weak mixing system, $A \in \mathcal{B}$ a set of positive measure and $k \in \mathbb{N}$ then

$$ \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k} T^{-jn} A \right) = \mu(A)^{k+1} > 0 $$

32
7 Compact Systems

We now turn our attention to the class of systems which exhibit ordered behaviour, exemplified by rotations of the circle. As in the previous section, we will prove the Multiple Recurrence Theorem holds for this particular class of systems. Thus we will have shown the result holds in the two extreme cases: when the system behaves randomly and when the behaviour of the system is structured. The work in this section is adapted from ([10], §4).

Definition 14. Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system and \(f \in L^2(X, \mathcal{B}, \mu)\). We say \(f\) is almost periodic (AP) if the orbit of \(f\) under \(T\) is relatively compact. In other words

\[\text{clos}\{T^n f : n \in \mathbb{N}\}\]

is a compact set in the norm topology.

A subset of a complete metric space is relative compact if and only if it is totally bounded. Hence a function \(f\) is AP if and only if for every \(\epsilon > 0\) there exists some finite set \(g_1, \ldots, g_m \in L^2(X, \mathcal{B}, \mu)\) such that

\[
\min_{1 \leq j \leq m} \|T^n f - g_j\|_{L^2(\mu)} < \epsilon \quad \text{for all } n \in \mathbb{N}
\]

Definition 15. A system \((X, \mathcal{B}, \mu, T)\) is compact if every \(f \in L^2(X, \mathcal{B}, \mu)\) is almost periodic.

Proposition 7.1. Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system and \(f \in L^2(X, \mathcal{B}, \mu)\) an AP function. Then for every \(\epsilon > 0\) there exists a syndetic set \(K \subseteq \mathbb{N}\) such that

\[
\|T^n f - f\|_{L^2(\mu)} < \epsilon \quad \text{for all } n \in K
\]

Proof. Let \(\epsilon > 0\) be given. Then as \(\text{clos}\{T^n f : n \in \mathbb{N}_0\}\) is compact there exists a (finite) maximum \(\epsilon\)-separated set, and by density we may assume this set has the form

\[
\{T^{t_1} f, \ldots, T^{t_r} f\}
\]

for some \(t_i \in \mathbb{N}\) numbered so that \(t_1 < t_2 < \cdots < t_r\). For each \(n \in \mathbb{N}\) consider the set

\[
\{T^{n+t_1} f, \ldots, T^{n+t_r} f\}
\]

(7.1)

Note that as \(T\) is measure preserving, for any pair \(1 \leq i < j \leq r\)

\[
\|T^{n+t_i} f - T^{n+t_j} f\|_{L^2(\mu)} = \|T^n(T^{t_i} f - T^{t_j} f)\|_{L^2(\mu)} = \|T^{t_i} f - T^{t_j} f\|_{L^2(\mu)} \geq \epsilon
\]

Thus, for each \(n \in \mathbb{N}\) the set (7.1) is also a maximum \(\epsilon\)-separated set.

It follow for each \(n \in \mathbb{N}\) there exists \(1 \leq i(n) \leq n\) such that

\[
\|T^{n+t_{i(n)}} f - f\|_{L^2(\mu)} < \epsilon
\]

as otherwise one may add \(f = T^0 f\) to the set (7.1) to obtain a larger \(\epsilon\)-separated set, contradicting maximality of (7.1).

Finally take \(K := \{n + t_{i(n)} : n \in \mathbb{N}\}\). It is easy to see \(\mathbb{N} = \bigcup_{a=1}^{k} K - a\) and so \(K\) is indeed syndetic.

\[\square\]

Theorem 7.2. Let \((X, \mathcal{B}, \mu, T)\) be a compact measure preserving system, \(f \in L^\infty(X, \mathcal{B}, \mu, T)\) such that \(f \geq 0\) and \(f\) is not almost everywhere 0. Then for all \(k \in \mathbb{N}\)

\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int \prod_{i=1}^{k} T^{i_n} f \, d\mu > 0
\]

(7.2)
Proof. By scaling we may assume without loss of generality $0 \leq f \leq 1$. To prove the theorem it suffices to show there exists a set $K \subseteq \mathbb{N}$ of positive lower density and constant $a > 0$ such that for all $n \in \mathbb{N}$

$$\int \prod_{l=1}^{k} T^{ln} f \, d\mu \geq a \quad (7.3)$$

For once this is established it then follows that

$$\frac{1}{N} \sum_{n=1}^{N} \int \prod_{l=1}^{k} T^{ln} f \, d\mu \geq \frac{1}{N} \sum_{n \in F \cap [1, N]} \int \prod_{l=1}^{k} T^{ln} f \, d\mu \geq a \frac{|F \cap [1, N]|}{N}$$

and as $d(F) > 0$, taking the limit infimum gives (7.2).

The proof proceeds in two simple stages:

- We show Proposition 7.1 implies given $0 < \epsilon < 1$ there exists a syndetic set $K \subseteq \mathbb{N}$ such that for all $n \in K$ and for all $l = 0, \ldots, k$ one has $\|T^{ln} f - f\|_{L^2(\mu)} < \epsilon$.

- Given $n \in K$ we demonstrate that as the $T^{ln} f$ are $\epsilon$-close in norm to $f$, it follows that the difference between $\int \prod_{l=1}^{k} T^{ln} f \, d\mu$ and $\int f^{k+1} \, d\mu$ is small. As the latter integral is strictly positive, for appropriate choice of $\epsilon$ so too is the former.

By the proposition, for any $\epsilon > 0$ there exists a syndetic set $K \subseteq \mathbb{N}$ such that for all $n \in K$ $\|T^{ln} f - f\|_{L^2(\mu)} < \frac{\epsilon}{k}$

Now, for all $n \in K$ and all $l = 0, \ldots, k$, by the triangle inequality and $T$ measure preserving

$$\|T^{ln} f - f\|_{L^2(\mu)} \leq \sum_{j=1}^{l} \|T^{(l-j+1)n} f - T^{(l-j)n} f\|_{L^2(\mu)} = \sum_{j=1}^{l} \|T^{ln} f - f\|_{L^2(\mu)} < \epsilon$$

For the second stage, note that for each $n \in K$

$$\left| \int \prod_{l=0}^{k} T^{ln} f \, d\mu - \int f^{k+1} \, d\mu \right| = \left| \int \sum_{j=0}^{k} \prod_{l=0}^{j-1} T^{ln} f(T^{jn} f - f)f^{k-j} \, d\mu \right| \leq \sum_{j=0}^{k} \int \prod_{l=0}^{j-1} T^{ln} f|T^{jn} f - f|f^{k-j} \, d\mu \leq \sum_{j=0}^{k} \int |T^{jn} f - f| \, d\mu$$

Where we have used our assumption $0 \leq f \leq 1$. By the Cauchy Schwarz inequality $\int |T^{jn} f - f| \, d\mu \leq \|T^{jn} f - f\|_{L^2(\mu)} < \epsilon$. Hence

$$\left| \int \prod_{l=0}^{k} T^{ln} f \, d\mu - \int f^{k+1} \, d\mu \right| < (k + 1)\epsilon \quad (7.4)$$

So far the choice of $\epsilon$ has been arbitrary. Now if we take $0 < \epsilon < \frac{1}{k+1} \int f^{k+1} \, d\mu$, then by (7.4) we conclude

$$\int \prod_{l=0}^{k} T^{ln} f \, d\mu \geq \int f^{k+1} \, d\mu - (k + 1)\epsilon > 0$$

Thus setting $a = \int f^{k+1} \, d\mu - (k + 1)\epsilon$ we have (7.3) as required. \qed
Theorem 7.3. Multiple recurrence holds for compact systems. Let \((X, \mathcal{B}, \mu, T)\) be a compact measure preserving system, \(A \in \mathcal{B}\) a set of positive measure and \(k \in \mathbb{N}\). Then
\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu\left(\bigcap_{i=0}^{k-1} T^{-i} A\right) > 0
\]

Proof. Simply apply Theorem 7.2 to the characteristic function \(1_A\). \qed

8 The Dichotomy Between Weak Mixing and Compact Systems

The purpose of the previous sections has been to prove the multiple recurrence result for certain types of measure preserving systems, namely weak mixing and compact systems. Whilst these classes are large enough to include important and interesting examples, by no means do they classify all measure preserving systems.

In order to arrive at a comprehensive multiple recurrence result we shall have to delve deeper and use a more subtle approach. Recalling the intuitive notion of weak mixing systems exhibiting random and compact systems exhibiting structured behaviour, it is reasonable to suppose if a measure preserving system fails to behave completely randomly then some significant ‘sub-system’ must behave in a structured manner.

Definition 16. Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system and suppose \(\mathcal{A} \subseteq \mathcal{B}\) is a sub-\(\sigma\)-algebra which is \(T\)-invariant: i.e. \(T^{-1} \mathcal{A} \subseteq \mathcal{A}\). Then \((X, \mathcal{A}, \mu, T)\) is a measure preserving system and we say this system is a factor of \((X, \mathcal{B}, \mu, T)\). If \(\mathcal{A}\) only contains null and conull sets, we say the factor \((X, \mathcal{A}, \mu, T)\) is trivial.

Factors of measure preserving systems will henceforth dominate our study. Rather than attempt demonstrate multiple recurrence for the system as a whole, we shall concentrate on demonstrating it for factors.

Definition 17. Suppose \((X, \mathcal{A}, \mu, T)\) is a factor of a system \((X, \mathcal{B}, \mu, T)\) and that for all \(A \in \mathcal{A}\) with \(\mu(A) > 0\) and all \(k \in \mathbb{N}\) we have
\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu\left(\bigcap_{j=0}^{k-1} T^{-jn} A\right) = 0
\]
then we say the action of \(T\) on \((X, \mathcal{A}, \mu)\) is Szemerédi (SZ).

The bulk of this section will be occupied with proving the following Theorem:

Theorem 8.1. A system \((X, \mathcal{B}, \mu, T)\) is not weak mixing if and only if there exists some non-trivial compact factor.

An intuitive interpretation of the content of this theorem might be: ‘a system exhibits completely random behaviour if and only if no non-trivial factor exhibits structure’.

An important consequence of this is the following:

Corollary 8.2. Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system. Then there exists a non-trivial factor \((X, \mathcal{A}, \mu, T)\) for which the action of \(T\) on \((X, \mathcal{A}, \mu)\) is SZ.

Proof. Assuming Theorem 8.1 the result follows directly from the multiple recurrence results for weak mixing and compact systems, as explained by the following flow chart:
In order to prove Theorem 8.1 we use the following characterisation of compact factors:

**Proposition 8.3.** Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system. Then there exists a non-trivial compact factor if and only if there exists some non-constant \(AP\) function.

**Proof.** Assume that every \(AP\) function \(f \in L^2(X, \mathcal{B}, \mu)\) is constant and \((X, \mathcal{A}, \mu, T)\) is some compact factor of \((X, \mathcal{B}, \mu, T)\). Then every \(g \in L^2(X, \mathcal{A}, \mu)\) is an \(AP\) function and thus also \(AP\) in the sense of \(L^2(X, \mathcal{B}, \mu)\). Thus the space \(L^2(X, \mathcal{A}, \mu)\) is precisely the set of constant functions on \(X\) and therefore \(\mathcal{A}\) must only contain sets of measure 1 and their complements.

For the converse, assume there exists some non-constant \(F \in L^2(X, \mathcal{B}, \mu)\) which is \(AP\). Let \(A\) denote the \(\sigma\)-algebra generated by \(\{T^n F : n \in \mathbb{N}\}\). It is easy to check \(A\) is a \(T\)-invariant sub-\(\sigma\)-algebra. Indeed, if \(A \in A\) then \(A = (T^n F)^{-1}(B)\) for some Borel set \(B \subseteq \mathbb{R}\) and

\[
T^{-1}A = T^{-1}\{x \in X : T^n F(x) \in B\} = \{x \in X : T^{n+1} F(x) \in B\} = (T^{n+1} F)^{-1}(B) \in \mathcal{A}
\]

So \(T^{-1}A \subseteq \mathcal{A}\) as required.

It remains to show the factor is compact: that every \(L^2(X, \mathcal{A}, \mu)\) function is \(AP\). To begin we show the subset \(V \subseteq L^\infty(X, \mathcal{A}, \mu)\) of bounded \(AP\) functions is a closed vector subspace. Suppose \(f, g \in V\) and \(a, b \in \mathbb{R} \setminus \{0\}\) and let \(\epsilon > 0\) be given. Then there exist functions \(f_1, \ldots, f_l\) and \(g_1, \ldots, g_m \in L^2(X, \mathcal{A}, \mu)\) such that

\[
\min_{1 \leq i \leq l} \|T^nf - f_i\|_{L^2(\mu)} < \frac{\epsilon}{2a} \quad \text{for all } n \in \mathbb{N}
\]

\[
\min_{1 \leq j \leq m} \|T^ng - g_j\|_{L^2(\mu)} < \frac{\epsilon}{2b} \quad \text{for all } n \in \mathbb{N}
\]

Applying the triangle inequality gives

\[
\min_{1 \leq i \leq l} \min_{1 \leq j \leq m} \|T^n(af + bg) - (af_i + bg_j)\| < \epsilon \quad \text{for all } n \in \mathbb{N}
\]

Hence the function \(af + bg \in V\) and \(V\) is a vector space. By a method similar to the proof that \(V\) is a vector space, we can show that the limit in \(L^2(X, \mathcal{A}, \mu)\) of any convergent sequence of \(AP\) functions is \(AP\) and hence the space \(V\) is topologically closed.

To complete the proof we claim that for every \(A \in \mathcal{A}\) the function \(1_A\) is \(AP\). It then follows by linearity that all simple functions are \(AP\) and thus, by the density of the space of simple functions, \(V = L^2(X, \mathcal{A}, \mu)\) as required.
Consider the case $A = F^{-1}(a, \infty)$ for some $a \in \mathbb{R}$. It is easy to show that for any pair $f, g \in V$ the functions $\min(f, g), \max(f, g) \in V$. In particular the function $G = \max(F - a, 0)$ and functions $H_n = \min(1, nG)$ defined for all $n \in \mathbb{N}$ are AP. Now

$$\|1_A - H_n\|_{L^2(\mu)} = \int_{\{x \in A, nG(x) < 1\}} |1 - nG|^2 \, d\mu \leq \mu(\{x \in A : nG(x) < 1\})$$

and thus

$$\lim_{n \to \infty} \|1_A - H_n\|_{L^2(\mu)} = 0$$

Hence $1_A$ is the $L^2(X, \omega, \mu)$-limit of AP functions and is therefore AP. As the collection of functions of the form $F^{-1}(a, \infty)$ together with $\{0, X\}$ in a generating algebra $\mathcal{G}$ for $\mathcal{B}$, the general case $A \in \mathcal{B}$ follows by approximating $A$ by a sequence of sets $(A_n)_{n \in \mathbb{N}} \subset \mathcal{G}$. \hfill \square

Applying Proposition 8.3 we restate Theorem 8.1 as the following:

**Proposition 8.4.** A system $(X, \mathcal{B}, \mu, T)$ is not weak mixing if and only if there exists a non-constant A.P. function $f \in L^2(X, \mathcal{B}, \mu)$.

**Proof.** Recall by Proposition 3.7 a system $(X, \mathcal{B}, \mu, T)$ is ergodic if and only all $T$-invariant functions are constant. Hence if $(X, \mathcal{B}, \mu, T)$ is not ergodic the result is trivial: simply take $f \in L^1(X, \mathcal{B}, \mu)$ a non-constant $T$-invariant function. Clearly $f$ is A.P. as the closure of the orbit $\text{clos}(\{T^n f : n \in \mathbb{N}\})$ is just the (compact) singleton $\{f\}$.

Otherwise we consider $(X, \mathcal{B}, \mu, T)$ an ergodic system which is not weak mixing. In this case, by Proposition 6.3 the product system $X \times X$ is not ergodic. It follows that there must exist some non-constant $T \times T$-invariant function $H \in L^1(X \times X)$.

Consider the convolution operator $H * : L^2(X, \mathcal{B}, \mu) \to L^2(X, \mathcal{B}, \mu)$ given by

$$H * \phi(x) = \int H(x, x') \phi(x') \, d\mu(x') \quad \text{for all } \phi \in L^2(X, \mathcal{B}, \mu) \quad (8.1)$$

We will show that every function in the image of this operator is AP. It will provide a wide selection of AP functions from which it is straightforward to find one which is non-constant.

It is an elementary fact from linear analysis that (8.1) is a (Hilbert-Schmidt and therefore) compact operator (see [2, 24]). Thus for any $R \geq 0$, the set

$$\text{clos}\{H * \theta : \|\theta\|_{L^2(\mu)} \leq R\}$$

is compact in $L^2(X, \mathcal{B}, \mu)$. We claim the following:

$$T^n(H * \phi) = H * T^n \phi \quad \text{for all } \phi \in L^2(X, \mathcal{B}, \mu) \text{ and all } n \in \mathbb{N}$$

This is easily deduced using the fact $T$ is measure preserving and $H$ is $T \times T$-invariant:

$$T^n(H * \phi)(x) = \int H(T^n x, x') \phi(x') \, d\mu(x') = \int H(x, T^n x') \phi(T^n x') \, d\mu(x') = \int H(x, x') \phi(T^n x') \, d\mu(x') = H * T^n \phi(x)$$

Now fix $\phi \in L^2(X, \mathcal{B}, \mu)$, let $R = \|\phi\|_{L^2(\mu)}$ and note that $\|T^n \phi\|_{L^2(\mu)} = R$ for all $n \in \mathbb{N}$. It follows that

$$\text{clos}\{T^n H * \phi : n \in \mathbb{N}\} = \text{clos}\{H * T^n \phi : n \in \mathbb{N}\} \subset \text{clos}\{H * \theta : \|\theta\| \leq R\}$$

Thus $\text{clos}\{T^n H * \phi : n \in \mathbb{N}\}$ is a closed subset of a compact set, and hence compact. Moreover for each $\phi \in L^2(X, \mathcal{B}, \mu)$ the function $H * \phi \in L^2(X, \mathcal{B}, \mu)$ is AP.
It remains to find a function \( \tilde{\phi} \in L^2(X, \mathcal{B}, \mu) \) such that \( H \ast \tilde{\phi} \) is non-constant. Consider the function \( H_1 \in L^1(X, \mathcal{B}, \mu) \) given by

\[
H_1(x') = \int H(x, x') \, d\mu(x)
\]

It is easy to see \( H_1 \) is \( T \)-invariant:

\[
H_1(Tx') = \int H(Tx, Tx') \, d\mu(x) = \int H(x, x') \, d\mu(x) = H_1(x')
\]

Recalling our assumption \( (X, \mathcal{B}, \mu, T) \) is ergodic, this function must therefore be constant, \( H_1 \equiv c \), say.

Now \( H - c \) is a non-constant \( T \times T \)-invariant function, so all our previous reasoning holds if we replace \( H \) with \( H - c \). Thus we may assume without loss of generality that \( H \) has the property \( H_1 \equiv 0 \).

By definition \( H \) itself is non-constant and in particular \( H \) is not almost everywhere 0. Thus there exists some \( \tilde{\phi} \in L^2(X, \mathcal{B}, \mu) \) such that

\[
H \ast \tilde{\phi} (x') = \int H(x, x') \phi(x') \, d\mu(x') \neq 0
\]

It follows from our assumption \( H_1 \equiv 0 \) and Fubini’s Theorem that

\[
\int H \ast \tilde{\phi} (x) \, d\mu(x) = \int \left( \int H(x, x') \, d\mu(x) \right) \tilde{\phi}(x') \, d\mu(x') = 0
\]

As the function \( H \ast \tilde{\phi} \in L^2(X, \mathcal{B}, \mu) \) is non-zero but has zero expectation, it must be non-constant as required.

For the converse, suppose \( (X, \mathcal{B}, \mu, T) \) is a weak mixing system. We will show that any AP function is constant.

Let \( \epsilon > 0 \) be given and suppose \( f \in L^2(X, \mathcal{B}, \mu) \) is AP. If \( f \equiv 0 \) we are done, so we may assume \( \|f\|_{L^2(\mu)} \neq 0 \). By Proposition 7.1 there exists a syndetic set \( G \subseteq \mathbb{N} \) such that

\[
\|T^n f - f\|_{L^2(\mu)} < \frac{\epsilon}{2\|f\|_{L^2(\mu)}} \text{ for all } n \in G \tag{8.2}
\]

On the other hand, if we recall the definition of weak mixing we have

\[
\lim_{N \to \infty} \sum_{n=1}^{N} \left[ \int f T^n f \, d\mu - \left( \int f \, d\mu \right)^2 \right] = 0
\]

Using the alternative characterisation of convergence in the mean, stated in Theorem 2.1, there exists a set \( Z \subseteq \mathbb{N} \) of zero density such that

\[
\left| \int f T^n f \, d\mu - \left( \int f \, d\mu \right)^2 \right| < \frac{\epsilon}{2} \text{ for all } n \in \mathbb{N} \setminus Z \tag{8.3}
\]

Obviously \( G \cap \mathbb{N} \setminus Z \neq \emptyset \) as otherwise \( G \subseteq Z \) and \( G \) is a set of positive lower density, a contradiction. Thus there exists some \( n_0 \in \mathbb{N} \) such that the both equalities stated in (8.2) and (8.3) hold for \( n = n_0 \).
Finally, by the triangle and Cauchy-Schwarz inequalities

\[
\left| \int f^2 \, d\mu - \left( \int f \, d\mu \right)^2 \right| \leq \left| \int f(T^{n_0} f - f) \, d\mu \right| + \left| \int T^{n_0} f \, d\mu - \left( \int f \, d\mu \right)^2 \right|
\]

\[
< \|f\|_{L^2(\mu)} \|T^{n_0} f - f\|_{L^2(\mu)} + \frac{\epsilon}{2} < \epsilon
\]

As \( \epsilon \) is arbitrary we must have \( \int f^2 \, d\mu = (\int f \, d\mu)^2 \). The function \( x \mapsto x^2 \) on \( \mathbb{R} \) is convex and we conclude for this equality to hold \( f \) must be constant.

\[ \square \]

8.1 Proof of the Multiple Recurrence Theorem: The Strategy

We are now in a position to sketch the details of the strategy used to prove the Multiple Recurrence Theorem. The language used here is somewhat vague and naïve, but it is hoped that this brief description will give the reader some sense of what we are aiming to achieve.

Essentially the proof is a standard Zorn’s Lemma argument. Using our new terminology, we are required to prove that given any system \((X, \mathcal{B}, \mu, T)\) the action of \( T \) on \((X, \mathcal{B}, \mu, T)\) is SZ. Let \( \mathcal{F} \) denote the collection of factors \( \mathcal{A} \subseteq \mathcal{B} \) for which the action of \( T \) on \((X, \mathcal{A}, \mu, T)\) is SZ, ordered by inclusion. By our previous work \( \mathcal{F} \) is non-empty and in particular contains some non-trivial factor. In Section 10 we use Zorn’s Lemma to show \( \mathcal{F} \) contains a maximal element.

Sections 11 to 13 then concentrate on showing that no proper factor can be maximal and hence this maximal SZ factor must be the whole of \( \mathcal{B} \).

In order to prove the latter claim we embark on a programme of ‘relativising’ the notions introduced in the previous sections. In particular, in Section 11 we concentrate on defining ‘weak mixing relative to \((X, \mathcal{A}, \mu, T)\)’. The most significant result of this will be:

If the action of \( T \) on \((X, \mathcal{A}, \mu, T)\) is SZ and \((X, \mathcal{B}, \mu, T)\) is a weak mixing relative to \((X, \mathcal{A}, \mu, T)\) then the action of \( T \) on \((X, \mathcal{B}, \mu, T)\) is also SZ.

Similarly, in Section 12 we define what it is meant for a system to be ‘relatively compact’ and derive an entirely complementary result for the case \((X, \mathcal{B}, \mu, T)\) compact relative to \((X, \mathcal{A}, \mu, T)\).

If the action of \( T \) on \((X, \mathcal{A}, \mu, T)\) is SZ and \((X, \mathcal{B}, \mu, T)\) is compact relative to \((X, \mathcal{A}, \mu, T)\) then the action of \( T \) on \((X, \mathcal{B}, \mu, T)\) is also SZ.

The final step in proving the Multiple Recurrence Theorem will be to prove a dichotomy between relative weak mixing and relative compact systems. This result is very similar in flavour to Theorem 8.1. In particular we show

For any factor \((X, \mathcal{A}, \mu, T)\) either:

- \((X, \mathcal{B}, \mu, T)\) is a weak mixing relative to \((X, \mathcal{A}, \mu, T)\), or
- There exists some intermediate factor \( \mathcal{A} \subseteq B^* \subseteq B \) such that \( \mathcal{A} \) is a proper factor of \( B^* \) and \((X, B^*, \mu, T)\) is compact relative to \((X, \mathcal{A}, \mu, T)\).

Combining these three results we conclude that no proper factor can be a maximal SZ factor and the proof of the Multiple Recurrence Theorem is complete.

Although the terminology used here may be as yet unclear to the reader, it should be obvious that the remainder of the proof mirrors to some extent the work we have done so far. In short, we prove a result for two opposite cases and then demonstrate and exploit a relationship between these extremes.

The definitions of ‘relatively compact’ and ‘relatively weak mixing’ will arise from considering conditional expectation and measure. In the following section we develop important measure theoretical techniques, generalising the familiar idea of conditional expectation of a discrete random variable from elementary probability theory. This will allow us to define a system of probability measures \( \mu_y \) which are ‘conditioned on the factor \( \mathcal{A} \)’. The definitions of weak mixing relative to \((X, \mathcal{A}, \mu, T)\) and compact relative to \((X, \mathcal{A}, \mu, T)\) are then simply analogues of the familiar definitions of weak mixing and compact systems which use these conditional measures.
9 Factors of Measure Preserving Systems

In the previous section we made significant progress toward the proof of the Multiple Recurrence Theorem. We demonstrated that for any measure preserving system \((X, \mathcal{B}, \mu, T)\) there exists some non-trivial factor \((X, \mathcal{A}, \mu, T)\) for which the action of \(T\) is SZ. Now we develop the theory necessary in order to make a serious study of factors, introducing certain tools which will allow us to relate the properties of the factors to the properties of an entire system.

9.1 Conditional Expectation

9.1.1 Motivation

This example serves as a reminder of the simple notion of conditional expectation assumed familiar to the reader\(^{13}\).

Example 5. Suppose \((\Omega, \mathcal{F}, \mathbb{P})\) is a probability space and \(f, g\) are random variables, each with a finite number of outcomes

\[
f : \Omega \rightarrow \{a_1, \ldots, a_n\}
\]

\[
g : \Omega \rightarrow \{b_1, \ldots, b_m\}
\]

Then the conditional probability that \(f = a_i\) given that \(g = b_j\) is defined as

\[
\mathbb{P}(f = a_i \mid g = b_j) = \frac{\mathbb{P}(f = a_i, g = b_j)}{\mathbb{P}(g = b_j)}
\]

This leads to the definition of the conditional expectation of \(f\), given \(g = b_j\)

\[
E(f \mid g = b_j) = \sum_{i=1}^{n} a_i \mathbb{P}(f = a_i \mid g = b_j)
\]

The conditional expectation of \(f\) given \(g\) is then defined as a random variable \(E(f \mid g)\)

\[
E(f \mid g)(\omega) = E(f \mid g = b_j) \quad \text{for } \omega \in g^{-1}(b_j)
\]

Implicit in this definition is the fact that the sets \(D_j = g^{-1}\{b_j\}\) for \(j = 1, \ldots, m\) are disjoint and partition \(\Omega\) and that the conditional expectation is constant on each of the \(D_j\). An important idea to grasp is that by defining \(\mathcal{G} := \sigma\{D_j : j = 1, \ldots, m\} = \sigma(g)\), we may think of \(E(f \mid g)\) as a \(\mathcal{G}\)-measurable function. Furthermore, if we give the set \(\Omega' = \{b_j\}_{j=1}^{m}\) the pull-back measure \(\mathbb{P}' = g^*\mathbb{P}\) induced by \(g\), that is

\[
\mathbb{P}'(\{b_j\}) = \mathbb{P}(g^{-1}\{b_j\}) \quad \text{for } j = 1, \ldots, m
\]

then

\[
g : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\Omega', \mathcal{P}(\Omega'), \mathbb{P}')
\]

is a measure preserving map and we may equally consider \(E(f \mid g)\) a \(\mathbb{P}'\)-measurable function on \(\Omega'\).

Our aim will be to generalise this notion of conditional expectation so that \(f, g\) may be taken to be continuous random variables. This is rather a non-trivial task and in some cases no meaningful analogies are available (although in many reasonable cases, and in particular when we are dealing with Borel probability space, such an analogy does exist).

\(^{13}\)Although it is stressed in advance that we will not, in fact, be adopting any decidedly probabilistic viewpoint in any part of this study.
Our approach will be to consider $\theta: (X, \mathcal{B}, \mu) \to (Y, \mathcal{D}, \nu)$ a measure preserving map (c.f. $g$ in (9.1)) and define for all $f \in L^1(X, \mathcal{B}, \mu)$ the conditional expectation as a function $E(f \mid Y) \in L^1(Y, \mathcal{D}, \nu)$. Intuitively, we think of $E(f \mid Y)$ as a direct analogy of the simple notion of conditional expectation from Example 5: that is, for every $y \in Y$ we have

$$E(f \mid Y)(y) = E(f \mid \theta = y)$$

This, however, is not a precise way of thinking; it is clear that much more subtlety is required\(^{14}\).

The sophomore part of this section will relate the theory of conditional expectation to the study of factors by introducing the notion of an extension of measure preserving systems.

The probabilistic context is not vital for this study: we shall introduce condition expectation as an object of linear algebra and not probability theory\(^{15}\). However, understanding the major result of this section, Theorem 9.4, is greatly aided by some probabilistic intuition.

### 9.1.2 Definition and Basic Properties

We follow the construction of the conditional expectation as given in ([8] Chapter 5). Let $\theta: (X, \mathcal{B}, \mu) \to (Y, \mathcal{D}, \nu)$ denote a measure preserving map between measure spaces. There is a naturally defined lift

$$\tilde{\theta}: L^2(Y, \mathcal{D}, \nu) \to L^2(X, \mathcal{B}, \mu)$$

given by

$$\tilde{\theta}(f) = f^\theta \text{ where } f^\theta(x) = f(\theta(x)) \text{ for all } x \in X.$$

Now, it is easy to see $L^2(Y, \mathcal{D}, \nu)^\theta \subseteq L^2(X, \mathcal{B}, \mu)$ is a closed, linear subspace. We may use orthogonal projects to relate the whole space to this subspace, and then map back to $L^2(Y, \mathcal{D}, \nu)$ in order to give a map $L^2(X, \mathcal{B}, \mu) \to L^2(Y, \mathcal{D}, \nu)$.

**Definition 18.** Define the conditional expectation operator\(^{16}\)

$$E(\cdot \mid Y): L^2(X, \mathcal{B}, \mu) \to L^2(Y, \mathcal{D}, \nu)$$

by

$$E(f \mid Y)^\theta = P f \quad \text{for all } f \in L^2(X, \mathcal{B}, \mu)$$

where $P$ is the orthogonal projection of $L^2(X, \mathcal{B}, \mu)$ onto the closed subspace $L^2(Y, \mathcal{D}, \nu)^\theta$.

It is clear that $E(\cdot \mid Y)$ is a linear operator and furthermore it is a contraction. Note also that if $f \in L^2(Y, \mathcal{D}, \nu)$ then $E(f^\theta \mid Y) = f$. Using elementary Hilbert space theory we establish the following:

**Proposition 9.1.** Let $f \in L^2(X, \mathcal{B}, \mu)$ and $h \in L^2(Y, \mathcal{D}, \nu)$. Then

$$\int f h^\theta \, d\mu = \int E(f \mid Y) h \, d\nu \quad (9.2)$$

In particular, taking $h = 1_A$ for any $A \in \mathcal{D}$ we have

$$\int_{\theta^{-1}(A)} f \, d\mu = \int_A E(f \mid Y) \, d\nu \quad \text{for all } f \in L^2(X, \mathcal{B}, \mu) \quad (9.3)$$

Finally, if $g \in L^\infty(Y)$ then

$$E(g^\theta f \mid Y) = gE(f \mid Y) \quad (9.4)$$

\(^{14}\)For instance, are all the singletons $\{y\}$ measurable; are the functions $E(f \mid Y)$ necessarily everywhere defined?

\(^{15}\)Of course a probabilistic definition does exists, see for instance ([5] Chapter 5; [25], Chapter 9).

\(^{16}\)Provided there is no risk of ambiguity, throughout the remainder of this discourse $X$ will denote the measure space $(X, \mathcal{B}, \mu)$ and $Y$ the measure space $(Y, \mathcal{D}, \nu)$. 

41
Proof. Note that \( L^2(Y, \mathcal{S}, \nu)^0 \) is a closed subspace of the Hilbert space \( L^2(X, \mathcal{B}, \mu) \) and so the projection \( P \) satisfies \( \langle Pf, k \rangle = \langle f, k \rangle \) for all \( k \in L^2(Y, \mathcal{S}, \nu)^0 \). Thus, as \( \theta \) is measure preserving,
\[
\int f h^\theta \, d\mu = \int Pf h^\theta \, d\mu = \int E(f \mid Y)^\theta h^\theta \, d\mu = \int E(f \mid Y) h \, d\nu
\]
giving (9.2) and (9.3). To show (9.4), note for any \( g \in L^\infty(Y) \) we have \( gh \in L^2(Y) \) and \( fg^\theta \in L^2(X) \). Applying (9.3) to \( gh \) gives
\[
\int f(gh)^\theta \, d\mu = \int E(f \mid Y) gh \, d\nu
\]
On the other hand,
\[
\int f(gh)^\theta \, d\mu = \int (fg^\theta)h^\theta \, d\mu = \int E(fg^\theta \mid Y) h \, d\nu
\]
Thus \( \int E(f \mid Y) gh \, d\nu = \int E(fg^\theta \mid Y) h \, d\nu \) for all \( h \in L^2(Y) \) so we must have (9.4) by the elementary properties of the inner product. \( \square \)

**Proposition 9.2.** For any \( f \in L^2(X, \mathcal{B}, \mu) \) with \( f(x) \geq 0 \) for almost every \( x \in X \) we have \( E(f \mid Y)(y) \geq 0 \) for almost every \( y \in Y \).

*Proof.* Suppose \( f \in L^2(X, \mathcal{B}, \mu) \) with \( f \geq 0 \), that is \( f(x) \geq 0 \) for almost every \( x \in X \). The subspace \( L^2(Y, \mathcal{S}, \nu)^+ \subseteq L^2(X, \mathcal{B}, \mu) \) is closed under \( f_1, f_2 \mapsto \max(f_1, f_2) \) so that \( g = \max(E(f \mid Y), 0) \in L^2(Y, \mathcal{S}, \nu)^+ \). Now suppose \( E(f \mid Y) \geq 0 \). The \( g \neq E(f \mid Y)_{L^2(\mu)} \) and \( \|f - g\|_{L^2(\mu)} < \|f - E(f \mid Y)\|_{L^2(\mu)} \). But this contradicts the closest point property of orthogonal projections onto closed linear subspaces. Hence \( E(f \mid Y) \geq 0 \). \( \square \)

**Proposition 9.3.** The conditional expectation operator extends to an operator \( E(\cdot \mid Y) : L^1(X, \mathcal{B}, \mu) \to L^1(Y, \mathcal{S}, \nu) \) satisfying all previously established properties mutatis mutandis.

*Proof.* The result follows by the density of \( L^2(X, \mathcal{B}, \mu) \) in \( L^1(X, \mathcal{B}, \mu) \) and applying standard Banach space theory. The details are omitted; see [5] or [8] for the full proof. \( \square \)

### 9.1.3 Disintegration of Measure

We are now in a position to present the main result of this section which makes explicit the sense by which the conditional expectation operator generalises the notion of conditional expectation as detailed in Example 5. The statement and subsequent proof of the following theorem are adapted from ([5] Chapter 5; [8] Chapter 5).

**Theorem 9.4** (Disintegration Theorem). Let \( \theta : (X, \mathcal{B}, \mu) \to (Y, \mathcal{S}, \nu) \) be a measure preserving map between probability spaces, where we assume \((X, \mathcal{B}, \mu)\) is a Borel probability space. Then there exists a \( \mathcal{S} \)-measurable set \( Y' \subseteq Y \) of measure 1 and system of probability measures \( \{\mu_y : y \in Y'\} \) on \((X, \mathcal{B})\) such that

1. For all \( f \in L^1(X, \mathcal{B}, \mu) \) and \( y \in Y' \), \( f \in L^1(X, \mathcal{B}, \mu_y) \) and
\[
\int f \, d\mu_y = E(f \mid Y)(y) \tag{9.5}
\]
2. For all \( f \in L^1(X, \mathcal{B}, \mu) \) the \( \nu \)-almost everywhere defined function
\[
y \mapsto \int f \, d\mu_y
\]
is \( \nu \)-measurable. Furthermore,
\[
\int \left( \int f \, d\mu_y \right) \, d\nu(y) = \int f \, d\mu \tag{9.6}
\]

42
Example 7. Thus the conditional measures and expectations generalise the simple case of Example 5.

Let \( B_\theta = \theta^{-1}\left(\{b_i\}\right) \) for each \( i \in \{1, \ldots, m\} \). If \( y = b_i \) then for any \( f \in L^1(X, \mathcal{B}, \mu) \) observe

\[
\int f \, d\mu_y = \frac{1}{\mu(B_\theta)} \int_{B_\theta} f \, d\mu
\]

Hence on each of the \( B_i \) the value of \( \int f \, d\mu_y \) is the average of the value of \( f \) over the region \( B_i \). In particular, if \( A \subset \mathcal{B} \) then

\[
\mu_y(A) = \frac{\mu(B_i \cap A)}{\mu(B_i)}
\]

Thus the conditional measures and expectations generalise the simple case of Example 5.

As the above example is somewhat degenerate we now turn to a more appropriate example.

Example 6. Inspired by Example 5, let \((X, \mathcal{B}, \mu)\) be any probability space satisfying the hypotheses of Theorem 9.4 and \( \theta \in L^1(X, \mathcal{B}, \mu) \) be any simple function. Then there exists some (minimal) finite subset \( \{b_1, \ldots, b_m\} \subset \mathbb{R} \) such that

\[
\theta(x) \in \{b_1, \ldots, b_m\} \quad \text{for almost every } x \in X
\]

Let \( B_i = \theta^{-1}(\{b_i\}) \) for each \( i \in \{1, \ldots, m\} \). If \( y = b_i \) then for any \( f \in L^1(X, \mathcal{B}, \mu) \) observe

\[
\int f \, d\mu_y = \frac{1}{\mu(B_i)} \int_{B_i} f \, d\mu
\]

Then \( \mu_y \) is a measure preserving map. For any \( f \in L^1(X, \mathcal{B}, \mu) \) the conditional expectation is given by

\[
E(f | Y)(x_1) = \int f(x_1, x_2) \, d\nu(x_2)
\]

For any \( y \in Y \) the set \( A = \{a\} \times [0, 1] \subset \mathcal{B} \) is \( \mu \)-null. However,

\[
\mu_y(A) = \int 1_A(a, x_2) \, d\nu(x_2) = \int \, d\nu(x_2) = 1
\]

Hence the measure \( \mu_y \) is supported on a \( \mu \)-null set.

In proving Theorem 9.4 the idea is to fix \( y \in Y \) and use the Riesz Representation Theorem to associate a measure to the (restricted) functional \( E(\cdot | Y)(y) : C(X) \rightarrow \mathbb{R} \). This approach, however, is naive and flawed: each \( E(f | Y) \in L^1(Y, \mathcal{B}, \nu) \) and thus only defined up to null-functions. For any given \( f \in C(X) \), the existence of \( E(f | Y)(y) \) is only guaranteed \( \nu \)-almost everywhere.

Happily, one can show the set on which the definition can fail is a countable union of null sets. The measures \( \mu_y \) can then be defined for \( \nu \)-almost every \( y \in Y \) as we will presently demonstrate.

Proof (of Theorem 9.4). First note that once Part 1 of the theorem holds for some \( f \in L^1(X, \mathcal{B}, \mu) \), Part 2 holds for \( f \) by virtue of the properties of the conditional expectation operator. Hence we are only required to prove Part 1.

Recall for the compact metric space \( X \), the space \( (C(X), \| \cdot \|_\infty) \) is separable. Therefore there exists a countable, dense set of functions \( \{b_0 = 1_X, b_1, b_2, \ldots\} \subset C(X) \). The set of linear combinations of these functions over \( \mathbb{Q} \) is countable dense vector space lying within \( C(X) \). We write this space as

\[
\mathcal{F} = \{f_0 = 1_X, f_1, f_2, \ldots\} \subset C(X)
\]

Recall, \( E(\cdot | Y) : L^1(X, \mathcal{B}, \mu) \rightarrow L^1(Y, \mathcal{B}, \nu) \) is a linear, bounded (with operator norm 1) and positive operator. Thus, fixing \( n \in \mathbb{N} \)

1. If \( f_n = pf_i + qf_j \) for some \( p, q \in \mathbb{Q} \) and \( i, j \in \mathbb{N} \) then

\[
E(f_n | Y)(y) = pf_i | Y(y) + qE(f_j | Y)(y) \quad \nu \text{-almost everywhere.}
\]
2. \(|E(fn|Y)|(y)| \leq \|fn\|_\infty \) \(\nu\)-almost everywhere.
3. If \(fn \geq 0\) then \(E(fn|Y)(y) \geq 0\) \(\nu\)-almost everywhere.

Thus, the set \(A_n \subseteq Y\) on which conditions 1, 2 and 3 hold for the function \(fn\) is of measure 1. The set \(Y'\) on which the conditions hold for all \(f \in \mathcal{F}\) is also measure 1: it is given by the countable intersection \(\bigcap_{n \in \mathbb{N}} A_n\).

For each \(n \in \mathbb{N}\) let \(g_n \in L^1(Y, \mathcal{F}, \nu)\) be an (everywhere defined) function representing the equivalence class \(E(fn|Y) \in L^1(Y, \mathcal{F}, \nu)\). For each \(y \in Y'\) we may then define a functional \(E_y : \mathcal{F} \rightarrow \mathbb{R}\) by \(E_y(fn) = g_n(x)\). By the construction of the set \(Y'\) this functional is \(\mathbb{Q}\)-linear, bounded and positive.

As the set \(\mathcal{F}\) is dense in \(C(X)\), by elementary theory of linear operators we may extend \(E_y\) to a continuous \(\mathbb{Q}\)-linear functional
\[
E_y : C(X) \rightarrow \mathbb{R}
\]

It then follows by continuity \(E_y\) is \(\mathbb{R}\)-linear. By the Riesz Representation Theorem, for every \(y \in Y'\) as \(E_y \in C(X)^*\) it can be represented by a signed measure \(\mu_y\) on \((X, \mathcal{B})\) with
\[
E_y(f) = \int f \, d\mu_y \quad \forall f \in C(X)
\]
Moreover, \(E_y\) is a positive functional and by Proposition 9.1 we have \(E_y(fn) = E(1_X|Y) = 1\). It follows for all \(y \in Y'\), the measure \(\mu_y\) is a probability measure on \((X, \mathcal{B})\).

We have obtained our system of measures \(\{\mu_y : y \in Y'\}\). It remains to show the properties enumerated in the statement of the theorem.

By definition for all \(f_n \in \mathcal{F}\) we have
\[
\int f_n \, d\mu_y = E(fn|Y)(y) \quad \text{almost everywhere } y \in Y'
\] (9.7)

Thus the theorem holds for all \(f_n \in \mathcal{F}\). We will use approximations to extend the result to all \(f \in L^1(X, \mathcal{B}, \mu)\).

By the density of \(\mathcal{F}\), given any \(g \in C(X)\) there exists a sequence \((f_n)_{n \in \mathbb{N}} \subseteq \mathcal{F}\) such that \(\lim_{n \to \infty} \|g - f_n\|_\infty = 0\). Therefore, by (9.7) we immediately deduce the theorem holds for all continuous functions.

As the continuous functions are dense in \(L^1(X, \mathcal{B}, \mu)\) for any \(f \in L^1(X, \mathcal{B}, \mu)\) there exists a sequence
\[
(g_n)_{n \in \mathbb{N}} \subseteq C(X) \quad \text{with } \sum_{n=1}^{\infty} \|g_n\|_{L^1(\mu)} < \infty \quad \text{such that } \sum_{n=1}^{\infty} g_n - f\|_{L^1(\mu)} = 0.
\]

Note that by the preceding argument, for almost every \(y \in Y\) and each \(n \in \mathbb{N}\) we have \(E(g_n|Y) = \int g_n \, d\mu_y\).

The conditional expectation operator is a contraction. As \(\|E(|g_n||Y)\|_{L^1(\nu)} \leq \|g_n\|_{L^1(\mu)}\) it follows \(\sum_{n=1}^{\infty} \|E(|g_n||Y)\|_{L^1(\nu)} < \infty\). In particular
\[
\sum_{n=1}^{\infty} E(|g_n||Y)(y) < \infty \quad \text{almost everywhere } y \in Y
\]
At a point \(y \in Y\) such that the above sum converges we have
\[
\sum_{n=1}^{\infty} |g_n| \in L^1(X, \mathcal{B}, \mu_y)
\] (9.8)
For each $N \in \mathbb{N}$ and almost every $y \in Y$ the function $\sum_{n=1}^{N} g_n \in L^1(X, \mathcal{B}, \mu_y)$ is dominated by (9.8) and hence by dominated convergence
\[
\int \sum_{n=1}^{\infty} g_n \, d\mu_y = \sum_{n=1}^{\infty} \int g_n \, d\mu_y \quad \text{almost everywhere } y \in Y
\]
\[
= \sum_{n=1}^{\infty} E(g_n \mid Y)(y) \quad \text{almost everywhere } y \in Y
\]
Equivalently, thinking of $y \mapsto \int \sum_{n=1}^{N} g_n \, d\mu_y$ as an $L^1(Y, \mathcal{P}, \nu)$ function,
\[
\lim_{N \to \infty} \left\| \sum_{n=1}^{N} E(g_n \mid Y) - \int \sum_{n=1}^{N} g_n \, d\mu(y) \right\|_{L^1(\nu)} = 0 \quad (9.9)
\]
On the other hand,
\[
\left\| \sum_{n=1}^{N} |E(g_n \mid Y)| \right\|_{L^1(\nu)} \leq \sum_{n=1}^{N} \|E(g_n \mid Y)\|_{L^1(\nu)} \leq \sum_{n=1}^{N} \|g_n\|_{L^1(\nu)} < \infty
\]
and so $\left\| \sum_{n=1}^{\infty} |E(g_n \mid Y)| \right\|_{L^1(\nu)} < \infty$. Hence, recalling (9.9) we conclude
\[
\int \sum_{n=1}^{\infty} g_n \, d\mu_y = E\left( \sum_{n=1}^{\infty} g_n \mid Y \right)(y) \quad \text{almost everywhere } y \in Y
\]
\[
= E(f \mid Y)(y)
\]
It remains to show
\[
\int f \, d\mu_y = \int \sum_{n=1}^{\infty} g_n \, d\mu_y \quad \text{almost everywhere } y \in Y \quad (9.10)
\]
In particular, suppose $A \in \mathcal{B}$ is such that $\sum_{n=1}^{\infty} g_n - f \neq 0$ for all $x \in A$. Then $\mu(A) = 0$ so it is clear that (9.10) follows once we have shown $A$ is also $\mu_y$-null for almost all $y \in Y$.

The open subsets of $X$ generate $\mathcal{B}$ and so there exists a sequence $(A_n)_{n \in \mathbb{N}}$ of open sets such that $A \subseteq A_n$ and $\mu(A_n) \to 0$ as $n \to \infty$. Let $h_n \in C(X)$ be supported on $\text{clos}(A_n)$ with $0 \leq h_n \leq 1$ for all $n \in \mathbb{N}$. Then for each $\epsilon > 0$ the functions $h_n$ are continuous and
\[
\int h_n \, d\mu_y \, d\nu(y) = \int h_n \, d\mu \leq \mu(A_n)
\]
Letting $\epsilon \to 0$ we have
\[
\int \mu_y(A_n) \, d\nu(y) \leq \mu(A_n)
\]
and so, by letting $n \to \infty$
\[
\int \mu_y(A) \, d\nu(y) = 0
\]
Thus $\mu_y(A) = 0$ for almost every $y \in Y$ as required. \hfill \Box

### 9.2 Extensions of Measure Preserving Systems

So far we have only studied maps $\theta$ between measure spaces, rather than measure preserving systems. In order to apply the preceding theory to the ergodic theory setting we make the following definition.

**Definition 19.** Let $\alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{P}, \nu, S)$ be a measurable map between measure preserving systems. If $\alpha$ preserves the structure of the systems in the sense $\alpha \circ T = S \circ \alpha$ then $\alpha$ is called an extension of measure preserving systems.
Proposition 9.5. Let \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) be an extension of invertible measure preserving systems. The conditional expectation operator \( E(\cdot \mid Y) : L^1(X, \mathcal{B}, \mu) \to L^1(Y, \mathcal{D}, \nu) \) satisfies
\[
SE(f \mid Y) = E(Tf \mid Y) \quad \text{for all } f \in L^1(X, \mathcal{B}, \mu)
\]

Proof. Let \( f \in L^1(X, \mathcal{B}, \mu) \). By the Proposition 9.1 we have
\[
\int E(Tf \mid Y) h \, d\nu = \int Tf h^\alpha \, d\mu \quad \text{for all } h \in L^1(Y, \mathcal{D}, \nu)
\]
As \( T \) is measure preserving so is \( T^{-1} \) and hence
\[
\int Tf h^\alpha \, d\mu = \int f T^{-1} h^\alpha \, d\mu
\]
Now, by the property of extensions \( T^{-1} h^\alpha = (S^{-1} h)^\alpha \). Putting everything together
\[
\int E(Tf \mid Y) h \, d\nu = \int f(S^{-1} h)^\alpha \, d\mu = \int E(f \mid Y) S^{-1} h \, d\mu = \int SE(f \mid Y) h \, d\mu
\]
As this holds for all \( h \in L^1(Y, \mathcal{D}, \nu) \), by the elementary properties of the inner product \( SE(f \mid Y) = E(Tf \mid Y) \).

Corollary 9.6. With the same hypotheses as the previous proposition,
\[
\mu_y(T^{-1} A) = \mu_{Sy}(A) \quad \text{for all } A \in \mathcal{B}
\]
for all \( y \in Y \) be such that the conditional measure \( \mu_y \) is defined.

Proof. Let \( A \in \mathcal{B} \) and \( y \in Y \) such that \( \mu_y \) is defined in accordance with Theorem 9.4. Note that \( T1_A = 1_{T^{-1} A} \) so by Proposition 9.5
\[
SE(1_A \mid Y)(y) = E(1_{T^{-1} A} \mid Y)(y) = \int 1_{T^{-1} A} \, d\mu_y = \mu_y(T^{-1} A)
\]
On the other hand,
\[
SE(1_A \mid Y)(y) = E(1_A \mid Y)(Sy) = \int 1_A \, d\mu_{Sy} = \mu_{Sy}(A)
\]
Thus \( \mu_y(T^{-1} A) = \mu_{Sy}(A) \) as required.

9.3 Factors and Extensions of Measure Preserving Systems

In this section we will demonstrate a correspondence between the factors of a system and extensions which will allow us to apply the notion of conditional measures to the study of factors. Our work here is adapted from ([5] Chapter 5). It is easy to see that any extension of measure preserving space induces a factor.

Lemma 9.7. Let \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) be an extension of measure preserving systems. Then \( \mathcal{A} = \alpha^{-1} \mathcal{D} \) is \( T \)-invariant sub-\( \sigma \)-algebra of \( \mathcal{B} \) and hence defines factor of \( (X, \mathcal{B}, \mu, T) \).
Proof. It is clear that \( \mathcal{A} \) is a \( \sigma \)-algebra on \( X \) and, as \( \alpha \) is measurable, \( \mathcal{A} \subseteq \mathcal{B} \). All that remains is to show \( \mathcal{A} \) is \( T \)-invariant.

Let \( B \in \mathcal{A} \) then \( B = \alpha^{-1}D \) for some set \( D \in \mathcal{D} \). Using the property of an extension,

\[
T^{-1}B = \{ x \in X : \alpha \circ T(x) \in D \} = \{ x \in X : S \circ \alpha(x) \in D \} = \alpha^{-1}(S^{-1}D)
\]

As \( S \) is \( \mathcal{D} \)-measurable, \( S^{-1}D \in \mathcal{D} \) and so \( T^{-1}B \in \mathcal{A} \) as required.

Our aim will be to prove a converse. Before we do this we note an important fact about Borel probability systems \((X, \mathcal{B}, \mu)\). We show that any sub-\( \sigma \)-algebra \( \mathcal{A} \subseteq \mathcal{B} \) is ‘almost generated by a countable collection of sets’.

**Definition 20.** A \( \sigma \)-algebra \( \mathcal{A} \) on a set \( X \) is countably generated if there exists a countable collection of sets \( \{A_n : n \in \mathbb{N}\} \subseteq \mathcal{P}(X) \) such that \( \mathcal{A} = \sigma(A_1, A_2, \ldots) \).

**Definition 21.** Let \( \mathcal{B} \) be a \( \sigma \)-algebra and \( \mathcal{A}, \mathcal{A}' \subseteq \mathcal{B} \) sub-\( \sigma \)-algebras. We write \( \mathcal{A} \leq_{\mathcal{B}} \mathcal{A}' \) to denote the relation: for all \( A \in \mathcal{A} \) there exists some \( A' \in \mathcal{A}' \) such that \( \mu(A \Delta A') = 0 \).

Furthermore, we write \( \mathcal{A} \preceq_{\mathcal{B}} \mathcal{A}' \) whenever \( \mathcal{A} \leq_{\mathcal{B}} \mathcal{A}' \) and \( \mathcal{A}' \leq_{\mathcal{B}} \mathcal{A} \).

**Proposition 9.8.** If \((X, \mathcal{B}, \mu)\) is a Borel probability space and \( \mathcal{A} \subseteq \mathcal{B} \) is a sub-\( \sigma \)-algebra then there exists a countably generated sub-\( \sigma \)-algebra \( \mathcal{A}' = \sigma(A_1, A_2, \ldots) \) such that \( \mathcal{A} \preceq_{\mathcal{B}} \mathcal{A}' \).

**Proof.** As \( C(X) \) is a separable metric space and dense in \( L^1(X, \mathcal{B}, \mu) \) it follows \( L^1(X, \mathcal{B}, \mu) \) is separable. Furthermore, the collection of characteristic functions

\[
\{ 1_A : A \in \mathcal{A} \} \subseteq L^1(X, \mathcal{A}, \mu) \subseteq L^1(X, \mathcal{B}, \mu)
\]

is separable. Let \( A_1, A_2, \ldots \) be such that \( \{1_{A_n} : n \in \mathbb{N}\} \) is a dense subset of \( \{1_A : A \in \mathcal{A}'\} \) and denote \( \mathcal{A}' = \sigma(A_1, A_2, \ldots) \). Then \( \mathcal{A}' \subseteq \mathcal{A} \). Moreover, by density for any \( A \in \mathcal{A} \) there exists a sequence \( (n_k)_{k \in \mathbb{N}} \) of positive integers such that

\[
\mu(A \Delta A_{n_k}) = \| 1_A - 1_{A_{n_k}} \|_{L^1(\mu)} < \frac{1}{k}
\]

Denote \( A' = \bigcap_{k \in \mathbb{N}} A_{n_k} \). Then \( A' \in \mathcal{A}' \) and \( \mu(A \Delta A') = 0 \) as required.

We now turn to the converse of Lemma 9.7.

**Theorem 9.9.** Let \((X, \mathcal{B}, \mu, T)\) be a measure preserving system defined on a Borel probability space and \( \mathcal{A} \) a \( T \)-invariant sub-\( \sigma \)-algebra of \( \mathcal{B} \). Then there exists a system \((Y, \mathcal{D}, \nu, S)\) defined on a Borel probability space and an extension \( \alpha : (X, \mathcal{B}, \mu, T) \rightarrow (Y, \mathcal{D}, \nu, S) \) such that \( \mathcal{A} \preceq_{\mathcal{B}} \alpha^{-1}\mathcal{D} \).

At a first glance the theorem may appear somewhat trivial: why not take \((Y, \mathcal{D}, \nu, S) = (X, \mathcal{A}, \mu, T)\) and \( \alpha \) the identity map? Of course, the key to this result is that \((Y, \mathcal{D}, \mu)\) is a Borel probability space: we cannot guarantee that this holds in general for \((X, \mathcal{A}, \mu, T)\). Our entire theory of conditional measure is reliant on the spaces in question being Borel probability spaces and thus this theorem is necessary when dealing with important situations involving intermediate extensions.

We will take \( Y \) in the statement of Theorem 9.9 to be the (compact, metrizable) space \( \mathcal{M}(X) \) of Borel probability measures on \((X, \mathcal{B})\). First we show that a canonical Borel probability system may be constructed on this space; once this is accomplished, Theorem 9.9 follows with ease.

**Proof (of Theorem 9.9).** As we demonstrated in Section 3.4, the set \( Y = \mathcal{M}(X) \) with the weak* topology is a compact, metrizable space. Let \( \mathcal{D} \) be the Borel \( \sigma \)-algebra generated by the weak* topology. In order to find a natural choice of measure on \((Y, \mathcal{D})\) we use the Disintegration Theorem. Note that the identity on \( X \) induces a map

\[
\iota : (X, \mathcal{B}, \mu) \rightarrow (X, \mathcal{A}, \mu) ; \quad x \mapsto x
\]
from a Borel probability space and is (trivially) measure preserving. By disintegration we have an almost everywhere defined measurable map

$$\alpha : (X, \mathcal{A}, \mu) \rightarrow (Y, \mathcal{D}) ; \quad y \mapsto \mu_y$$

(9.11)

taking \( y \in X \) to the conditional measure \( \mu_y \). We give \((Y, \mathcal{D})\) the pull-back measure \( \nu = \alpha^* \mu \) induced by this map, so that

\[ \nu(D) = \mu(\alpha^{-1}D) \quad \text{for all } D \in \mathcal{D} \]

Finally we wish to define a natural measure preserving transformation on the Borel probability space \((Y, \mathcal{D}, \nu)\). This is marginally trickier than the previous constructions.

Claim. The pull-back map \( S = T^* : (Y, \mathcal{D}, \nu) \rightarrow (Y, \mathcal{D}, \nu) \) defined by

\[ S\lambda(A) = \lambda(T^{-1}A) \quad \text{for all } \lambda \in \mathcal{M}(X) \text{ and all } A \in \mathcal{A} \]

is measurable. Moreover, it preserves the measure \( \nu \).

Proof (of Claim). We follow the argument given in ([5] Chapter 5). In order to prove \( S \) is measurable we must first make some observations concerning the topology that generates \( \mathcal{D} \). In particular, it is easy to see by the definition of the weak*-topology that a set \( U = \{ \lambda \in \mathcal{M}(X) : |\int f \, d\lambda - \int f \, d\nu| < \epsilon \}, \text{ for all } i = 1, \ldots, r \} \subseteq U \)

Recall \( C(X) \) is separable. Let \((g_n)_{n \in \mathbb{N}} \subseteq C(X)\) be a dense subset of \( C(X) \) and \( \epsilon > 0 \) be given. Then for any \( f \in C(X) \) there exists an \( n_0 \in \mathbb{N} \) such that \( ||f - g_{n_0}||_{\infty} < \frac{\epsilon}{3} \). If \( \lambda, \kappa \in \mathcal{M}(X) \) are such that

\[ |\int f \, d\lambda - \int f \, d\kappa| < \frac{\epsilon}{3} \]

then by the triangle inequality

\[ |\int g_{n_0} \, d\lambda - \int g_{n_0} \, d\kappa| < \epsilon \]

It follows that the collection of all finite intersections of sets of the form

\[ \left\{ \lambda \in \mathcal{M}(X) : |\int g_n \, d\lambda - \int g_n \, d\kappa| < \epsilon \right\} \]

for some \( \kappa \in \mathcal{M}(X) \), \( \epsilon > 0 \) and \( n \in \mathbb{N} \) is a basis for the topology on \( \mathcal{M}(X) \). Moreover, by the density of the rationals, the collection of finite intersections of sets of the form

\[ U_{(n,r,q)} = \left\{ \lambda \in \mathcal{M}(X) : |\int g_n \, d\lambda - r| < q \right\} \]

for some \( n \in \mathbb{N} \) and \( r, q \in \mathbb{Q} \) is a countable basis for the topology on \( \mathcal{M}(X) \). Now,

\[ S^{-1}U_{(n,r,q)} = \left\{ \lambda \in \mathcal{M}(X) : |\int Tg_n \, d\lambda - r| < q \right\} \]

is clearly measurable. Moreover, for any \( U \subseteq \mathcal{M}(X) \) open, \( S^{-1}U \) is a countable union of finite intersections of measurable sets and therefore measurable. The open subsets of \( \mathcal{M}(X) \) form a generating algebra for \( \mathcal{D} \) and hence \( S \) is \( \mathcal{D} \)-measurable.

It remains to prove \( S \) preserves the measure \( \nu \). By Corollary 9.6, for all \( A \in \mathcal{B} \) we have \( \mu_{T\nu}(A) = S\mu_{\nu}(A) \) and so

\[ S \circ \alpha = \alpha \circ T \]

(9.12)
It follows that for any $D \in \mathcal{D}$ we have

$$\alpha^{-1}S^{-1}(D) = T^{-1}\alpha^{-1}(D)$$

Finally

$$\alpha^*\mu(S^{-1}D) = \mu(\alpha^{-1}S^{-1}(D)) = \mu(T^{-1}\alpha^{-1}(D)) = \nu(D)$$

and hence $S$ preserves $\nu$.

Now that the measure preserving system $(Y, \mathcal{D}, \nu, S)$ has been constructed, most of the work in proving Theorem 9.9 has been done. Through a slight abuse of notation consider

$$\alpha : (X, B, \mu, T) \rightarrow (Y, \mathcal{D}, \nu, S)$$

Then, as we have shown, $(Y, \mathcal{D}, \nu, S)$ is a measure preserving system and by (9.12), $\alpha$ is an extension of measure preserving systems.

We are required to prove $\mu = \pi^{-1}\mathcal{D}$. To see this note that as $(X, B, \mu)$ is a Borel probability space, by Proposition 9.8 there exists some countably generated sub-$\sigma$-algebra $\mathcal{A}'$ such that $\mathcal{A}' \subseteq \mathcal{D}$. As the $\subseteq$ relation is transitive, we may assume therefore without loss of generality $A_n = \sigma(A_1, A_2, \ldots)$ is countably generated.

To see $\mu = \pi^{-1}\mathcal{D}$ it is enough to show that for each $n \in \mathbb{N}$ there exists some set $A_n \in \alpha^{-1}D$ such that $\mu(A_n \Delta A_n) = 0$.

Consider the $\mathcal{D}$-measurable set $D = \{\nu \in Y : \nu(A_n) = 1\}$. Note that

$$\alpha^{-1}D = \{x \in X : \mu_x(A_n) = 1\}$$

Recalling that the conditional measures are defined through the trivial extension $\iota$ it is easy to see

$$\mu_x(A_n) = 1_{A_n}(x) \quad \text{almost everywhere } x \in X$$

Define

$$X' = \bigcap_{n \in \mathbb{N}} \{x \in X : \mu_x(A_n) = 1_{A_n}(x)\}$$

Then $\mu(X') = 1$ and so, if we set $A_n = A_n \cap X' \in \mathcal{D}$ then $\mu(A_n \Delta A_n) = 0$ for all $n \in \mathbb{N}$.

Conversely, to prove $\alpha^{-1}D \subseteq \mathcal{A}'$ it suffices to show $\alpha^{-1}(U) \in \mathcal{A}'$ for all open sets $U$ of the form (9.3) as these sets form a basis for the topology on $\mathcal{M}(X)$ which generates $\mathcal{D}$. But this is immediate as for any $f \in C(X)$ and any $r, \epsilon > 0$ we have

$$\alpha^{-1}\left\{\nu \in \mathcal{M}(X) : \left| \int f \, d\nu - r \right| < \epsilon \right\} = \left\{x \in X : \left| \int f \, d\mu_x - r \right| < \epsilon \right\}$$

is $\mathcal{A}'$-measurable by the construction of the conditional measures.

We now have enough measure-theoretic tools to return to the proof of the Multiple Recurrence Theorem.

### 10 Existence of Maximal SZ Factors

As payoff for the work done in the previous section we turn to proving the following theorem, an important step towards the proof of the Multiple Recurrence Theorem.

**Theorem 10.1.** Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and $\mathcal{F}$ the set of factors $\mathcal{A} \subseteq \mathcal{B}$ of the system for which the action of $T$ on $(X, \mathcal{A}, \mu, T)$ is SZ. Then $\mathcal{F}$ contains a maximal element with respect to inclusion.
Proof (of Theorem 10.1). Let \( \mathcal{A}_i \) be a chain, or totally ordered subset, \( \{ \mathcal{A}_i : i \in I \} \subseteq \mathcal{F} \) of \( \mathcal{F} \) and define

\[
\sup_{i \in I} \mathcal{A}_i := \sigma \left( \bigcup_{i \in I} \mathcal{A}_i \right)
\]

To prove Theorem 10.1 it suffices to show \( \sup_{i \in I} \mathcal{A}_i \) is an upper bound for the chain. Clearly

- \( \sup \mathcal{A}_i \) is a \( \sigma \)-algebra
- \( \mathcal{A}_i \subseteq \sup \mathcal{A}_i \) for all \( i \in I \)
- \( \sup \mathcal{A}_i \) is \( T \)-invariant

Note that for the final bullet point, the set \( \{ A \in \sup_{i \in I} \mathcal{A}_i : T^{-1}A \in \sup_{i \in I} \mathcal{A}_i \} \) is a \( d \)-system and thus by Dynkin’s Lemma it suffices to show \( T \)-invariance only for sets in the generating algebra \( \cup_{i \in I} \mathcal{A}_i \). For such sets the result is trivial.

The proof of Theorem 10.1 therefore boils down to showing that the action of \( T \) on \( (X, \sup_{i \in I} \mathcal{A}_i, \mu, T) \) is SZ. Explicitly, for any \( A \in \sup_{i \in I} \mathcal{A}_i \) such that \( \mu(A) > 0 \) we need to show that given any \( k \in \mathbb{N} \),

\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k} T^{-jn}A \right) > 0
\]  \hspace{1cm} (10.1)

**Proof (of Theorem 10.1).** Let \( A \in \sup_{i \in I} \mathcal{A}_i \) be such that \( \mu(A) > 0 \) and \( k \in \mathbb{N} \). By the above discussion it suffices to show (10.1). By Lemma 3.8, if \( \epsilon = \frac{1}{2} \eta \mu(A) \) where \( \eta = \frac{1}{2}(k+1)^{-1} \) then there exists a set \( A_1 \in \mathcal{A}_i \) for some \( i_0 \in I \) such that

\[
A \subseteq A_1 \quad \text{and} \quad \mu(A \triangle A_1) < \epsilon
\]  \hspace{1cm} (10.2)

Note that \( \mu(A_1) > 0 \). In particular,

\[
\mu(A_1) \geq \mu(A) - \mu(A \triangle A_1) > \mu(A) - \frac{1}{4} \eta \mu(A)
\]  \hspace{1cm} (10.3)

Now we apply the results of the previous section. By Theorem 9.9, there exists a Borel probability system \( (Y, \mathcal{G}, \nu, S) \) and extension \( \alpha : (X, \mathcal{F}, \mu, T) \to (Y, \mathcal{G}, \nu, S) \) such that \( \mathcal{A}_0 = \mathcal{A}_1 = \mathcal{A}_{i_0} \). In particular, \( A_1 = \alpha^{-1} D_1 \) for some set \( D_1 \in \mathcal{G} \).

Disintegrate the measure \( \mu \) with respect to \( (Y, \mathcal{G}, \nu) \) to form a collection of probability measures \( \mu_y \) defined for almost every \( y \in Y \). Note that

\[
\mu_y(A_1) = E(1_{D_1} | Y)(y) = 1_{D_1}(y)
\]

Now \( A_1 \) may not give a good approximation to \( A \) in terms of the conditional measures \( \mu_y \). To this end we define

\[
D_0 := \{ y \in Y : \mu_y(A) \geq 1 - \eta \}
\]

and let \( A_0 = \alpha^{-1}(D_0) \). We claim \( \mu(A_0) > \frac{1}{2} \mu(A) > 0 \). Indeed, using the properties of the conditional measure

\[
\mu(A_1 \setminus A) = \int \mu_y(A_1 \setminus A) \, d\nu(y)
\]

\[
\geq \int_{D_1 \setminus D_0} \mu_y(A_1) - \mu_y(A) \, d\nu(y)
\]

\[
= \int_{D_1 \setminus D_0} 1 - \mu_y(A) \, d\nu(y)
\]

\[
\geq \eta \nu(D_1 \setminus D_0)
\]

\[
\geq \eta(\nu(D_1) - \nu(D_0))
\]

50
Now α is measure preserving and so
\[
\mu(A_0) \geq \mu(A_1) - \frac{1}{\eta} \mu(A_1 \setminus A) \\
= \mu(A_1) - \frac{1}{\eta} \mu(A_1 \triangle A) \quad (10.4)
\]
Finally, by applying (10.2) and (10.3) to (10.4) we get
\[
\mu(A_0) > \frac{3 - \eta}{4} \mu(A) > \frac{1}{2} \mu(A)
\]
By hypothesis the action of \(T\) on \((X, \mathcal{A}, \mu, T)\) is \(SZ\). In particular,
\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k} T^{-jn} A_0 \right) > 0 \quad (10.5)
\]
We claim that
\[
\mu_y \left( \bigcap_{j=0}^{k} T^{-jn} A \right) > \frac{1}{2}
\]
Once the claim is proven it is easy to see that
\[
\mu \left( \bigcap_{j=0}^{k} T^{-jn} A \right) \geq \int_{\bigcap_{j=0}^{k} S^{-jn} D_0} \mu_y \left( \bigcap_{j=0}^{k} T^{-jn} A \right) d\nu \\
\geq \frac{1}{2} \nu \left( \bigcap_{j=0}^{k} S^{-jn} D_0 \right) \\
= \frac{1}{2} \mu \left( \bigcap_{j=0}^{k} T^{-jn} A_0 \right) \quad (10.6)
\]
and hence comparing (10.5) and (10.6) concludes the proof.

To prove the claim suppose \(y \in \bigcap_{j=0}^{k} S^{-jn} D_0\). Then for \(j = 0, \ldots, k\), we have \(S^{jn}y \in D_0\) and thus, applying Corollary 9.6,
\[
\mu_y(T^{-jn} A) = \mu_{S^{jn}y}(A) > 1 - \eta \quad \text{for all } j \in \{0, \ldots, k\}
\]
In particular the complement of each \(T^{-jn} A\) has measure \(\mu_y((T^{-jn} A)^c) < \eta\). The measure of the intersection \(\bigcap_{j=0}^{k} T^{-jn} A\) is smallest when the complements \((T^{-jn} A)^c\) are all disjoint. Hence
\[
\mu_y \left( \bigcap_{j=0}^{k} T^{-jn} A \right) > 1 - (k + 1)\eta = \frac{1}{2}
\]
and we are done. \(\square\)

11 Weak Mixing Extensions

We will briefly recap what we have thus achieved and remind the reader of the proof strategy laid out earlier in the essay. So far we have shown that any system \((X, \mathcal{B}, \mu, T)\) contains some maximal (non-trivial) factor with respect to inclusion, \((X, \mathcal{A}, \mu, T)\) say. The proof of the Multiple Recurrence Theorem is concluded by demonstrating no proper factor can be maximal. The method used to achieve this rather ingeniously mirrors our previous work.

51
In this section we describe what it is meant for a system \((X, \mathcal{B}, \mu, T)\) to be weak mixing relative to a factor \((X, \mathcal{A}, \mu, T)\). As factors correspond precisely to extensions, this is equivalent to defining a particular kind of extension \(\alpha : (X, \mathcal{B}, \mu, T) \rightarrow (Y, \mathcal{D}, \nu, S)\) known as a weak mixing extension. We will then show that weak mixing extensions preserve the SZ property: if the action of \(S\) on \((Y, \mathcal{D}, \nu, S)\) is SZ and \(\alpha : (X, \mathcal{B}, \mu, T) \rightarrow (Y, \mathcal{D}, \nu, S)\) is a weak mixing extension then the action of \(T\) on \((X, \mathcal{B}, \mu, T)\) is also SZ. Thus we will have shown:

If \((X, \mathcal{A}, \mu, T)\) is a proper maximal SZ factor of \((X, \mathcal{B}, \mu, T)\) then \((X, \mathcal{B}, \mu, T)\) cannot be weak mixing relative to \((X, \mathcal{A}, \mu, T)\).

Similarly, in Section 12 we define what is meant for a system to be compact relative to a factor via the notion of compact extensions. We will show that compact extensions also preserve the SZ property. Hence if \((X, \mathcal{B}, \mu, T)\) is compact relative to \((X, \mathcal{A}, \mu, T)\) then \((X, \mathcal{B}, \mu, T)\) cannot be compact relative to \((X, \mathcal{A}, \mu, T)\). Furthermore, we deduce a more general statement:

Suppose \((X, \mathcal{A}, \mu, T)\) is a proper maximal SZ factor of \((X, \mathcal{B}, \mu, T)\) and \(\mathcal{B}^*\) is some intermediate factor, that is \((X, \mathcal{B}^*, \mu, T)\) is a factor of \((X, \mathcal{B}, \mu, T)\) and \(\mathcal{A} \subseteq \mathcal{B}^* \subseteq \mathcal{B}\). If \((X, \mathcal{A}, \mu, T)\) is a proper factor of \((X, \mathcal{B}^*, \mu, T)\), then \((X, \mathcal{B}^*, \mu, T)\) cannot be compact relative to \((X, \mathcal{A}, \mu, T)\).

In the final section we conclude the proof of the Multiple Recurrence Theorem by demonstrating a dichotomy between weak mixing and compact extensions. We demonstrate that for any proper factor \((X, \mathcal{A}, \mu, T)\) of \((X, \mathcal{B}, \mu, T)\) either:

- \((X, \mathcal{B}, \mu, T)\) is weak mixing relative to \((X, \mathcal{A}, \mu, T)\) or
- There exists some intermediate factor \(\mathcal{A} \subseteq \mathcal{B}^* \subseteq \mathcal{B}\) such that \(\mathcal{A}\) is a proper factor of \(\mathcal{B}^*\) and \((X, \mathcal{B}^*, \mu, T)\) is compact relative to \((X, \mathcal{A}, \mu, T)\).

Bringing all this together we conclude no proper factor can be maximal.

Presently we will use the notions of conditional expectation and measure to define the concept of a ‘relative weak mixing’ system. Recall that a system \((X, \mathcal{B}, \mu, T)\) is weak mixing if and only the product system \(X \times X\) is ergodic. This inspires us to consider defining a ‘conditional product’.

### 11.1 The Relative Square

This subsection is based on ([8] Chapter 5).

**Definition 22.** Suppose \((X_1, \mathcal{B}_1, \mu_1, T_1)\) and \((X_2, \mathcal{B}_2, \mu_2, T_2)\) are measure preserving systems defined on Borel probability spaces which both extend the system \((Y, \mathcal{D}, \nu, S)\). Explicitly, suppose there exists extensions

\[
\alpha_1 : (X_1, \mathcal{B}_1, \mu_1, T_1) \rightarrow (Y, \mathcal{D}, \nu, S) \\
\alpha_2 : (X_2, \mathcal{B}_2, \mu_2, T_2) \rightarrow (Y, \mathcal{D}, \nu, S)
\]

Thus we may disintegrate the measures \(\mu_1\) and \(\mu_2\) with respect to \((Y, \mathcal{D}, \nu)\) to form two systems of conditional measures \(\mu_{1,y}\) and \(\mu_{2,y}\), defined for almost all \(y \in Y\). It is not difficult to see \(\mu_{1,y} \otimes \mu_{2,y}\) is \(\nu\)-measurable (for instance, see [8]). We define the relative independent joining \(X_1 \times_Y X_2\) of the systems \(X_1 = (X_1, \mathcal{B}_1, \mu_1, T_1)\) and \(X_2 = (X_2, \mathcal{B}_2, \mu_2, T_2)\) as the system

\[(X_1 \times X_2, \mathcal{B}_1 \otimes \mathcal{B}_2, \mu_1 \times_Y \mu_2, T_1 \times T_2)\]

where the measure \(\mu_1 \times_Y \mu_2 = \int \mu_{1,y} \otimes \mu_{2,y} \, d\nu(y)\).

**Lemma 11.1.** Let \(f_1 \in L^2(X_1, \mathcal{B}_1, \mu_1)\) and \(f_2 \in L^2(X_2, \mathcal{B}_2, \mu_2)\). Then \(f_1 \otimes f_2 \in L^2(\mu_1 \times_Y \mu_2)\) and

\[
\int f_1 \otimes f_2 \, d(\mu_1 \times_Y \mu_2) = \int E(f_1 | Y) E(f_2 | Y) \, d\nu
\]

\[52\]
Proposition 11.3. For any systems where

\[ \text{Definition 23.} \text{Suppose } \alpha : (X, \mathcal{B}, \mu, T) \rightarrow (Y, \mathcal{D}, \nu, S) \text{ is an extension of measure preserving systems where } (X, \mathcal{B}, \mu, T) \text{ is defined on a Borel probability space. Then we can form the relative independent joining } X \times_Y X. \text{ We say } X \times_Y X \text{ is the Relative Square of the system } (X, \mathcal{B}, \mu, T) \text{ with respect to the factor } (Y, \mathcal{D}, \nu, S). \text{ Note by Lemma 11.1, for any } f \in L^2(X, \mathcal{B}, \mu) \text{ we have}
\]

\[ \int f \otimes f \, d(\mu \times_Y \mu) = \int \int f_1 \otimes f_2 \, d(\mu_1 \otimes \mu_2) \, d\nu(y) \]
\[ = \int \int f_1 \, d\mu_1 \int f_2 \, d\nu(y) \]
\[ = \int E(f_1 \mid Y)(y)E(f_2 \mid Y)(y) \, d\nu(y) \]

\[\square\]

Lemma 11.2. For any \( f \in L^2(\mu_1 \times_Y \mu_2) \) we have

\[ E(f \mid Y) = E(E(f \mid X) \mid Y) \]

Proof. Considering orthogonal projections

\[ P_1 : L^2(X \times_Y X) \rightarrow L(X, \mathcal{B}, \mu) \]
\[ P_2 : L(X, \mathcal{B}, \mu) \rightarrow L(Y, \mathcal{D}, \nu) \]
\[ Q : L^2(X \times_Y X) \rightarrow L(Y, \mathcal{D}, \nu)^{\alpha \circ \pi} \]

it is easy to see \( Q = P_2^\pi \circ P_1 \). The result then follows immediately by the definition of conditional expectation. \[\square\]

Proposition 11.3. For any \( f_1, f_2 \in L^2(X, \mathcal{B}, \mu) \) we have

\[ E(f_1 \otimes f_2 \mid Y) = E(f_1 \mid Y)E(f_2 \mid Y) \]

Proof. First we show that \( E(f_1 \otimes f_2 \mid X) = f_1E(f_2 \mid Y)^\alpha \). Indeed, for any \( h \in L^2(X, \mathcal{B}, \mu) \), by applying Lemma 11.1 and the standard properties of conditional expectation we have

\[ \int f_1E(f_2 \mid Y)^\alpha h \, d\mu = \int E(f_1E(f_2 \mid Y)^\alpha h \mid Y) \, d\nu \]
\[ = \int E(f_1h \mid Y)E(f_2 \mid Y) \, d\nu \]
\[ = \int f_1 \otimes f_2 h^\pi \, d(\mu \times_Y \mu) \]

Further, now considering the extension \( \pi_1 : X \times_Y X \rightarrow X \) observe

\[ \int f_1 \otimes f_2 h^\pi \, d(\mu \times_Y \mu) = \int E(f_1 \otimes f_2 \mid X)h \, d\mu \] (11.2)
Hence, by comparing the left hand side of (11.1) and the right hand side of (11.2), and noting that the equalities hold for all \( h \in L^2(X, \mathcal{B}, \mu) \) we conclude \( E(f_1 \otimes f_2 | X) = f_1 E(f_2 | Y) \alpha \) by the elementary properties of the inner product.

The proposition now follows by applying the preceding lemma. Explicitly,
\[
E(f_1 \otimes f_2 | Y) = E(E(f_1 \otimes f_2 | X) | Y) = E(f_1 E(f_2 | Y) \alpha | Y) = E(f_1 | Y) E(f_2 | Y)
\]

\[\square\]

11.2 Weak Mixing Extensions

The following discussion of weak mixing extensions is adapted from ([10], §8).

**Definition 24.** Let \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) be an extension of measure preserving systems. We say the extension is weak mixing if the relative-square system \( X \times_Y X \) is ergodic.

Alternatively, in this case we may also say that \( (X, \mathcal{B}, \mu, T) \) is weak mixing relative to the factor \( (X, \mathcal{A}, \mu, T) \) where \( \mathcal{A} = \alpha^{-1}(\mathcal{D}) \) or weak mixing relative to \( (Y, \mathcal{D}, \nu, S) \).

Following the argument given in ([10], §8) we first note the following lemma.

**Lemma 11.4.** Suppose \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) is a weak mixing extension and \( f, g \in L^\infty(X, \mathcal{B}, \mu) \) then the following holds:
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int |E(fT^n g | Y) - E(f | Y) S^n E(g | Y)|^2 \, d\nu = 0
\]

**Proof.** Let \( f, g \in L^\infty(X, \mathcal{B}, \mu) \) and first consider the case \( E(f | Y) \equiv 0 \). By hypothesis the system \( X \times_Y X \) is ergodic. Apply Lemma 11.1 and the Mean Ergodic Theorem to the functions \( f \otimes g \in L^\infty(X \times_Y X) \) to obtain
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int |E(fT^n g | Y)|^2 \, d\nu = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int f \otimes f(T \times T)^n (g \otimes g) \, d(\mu \times_Y \mu)
\]
\[
= \int f \otimes f \, d(\mu \times_Y \mu) \int g \otimes g \, d(\mu \times_Y \mu)
\]
\[
= \int E(f | Y)^2 \, d\nu \int E(g | Y)^2 \, d\nu = 0
\]

The general case where \( f \in L^\infty(X, \mathcal{B}, \mu) \) has arbitrary conditional expectation now follows by considering the function \( F = f - E(f | Y) \). Clearly \( E(F | Y) = E(f | Y) - E(E(f | Y) | Y) = E(f | Y) - E(f | Y) = 0 \) and by the previous case
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int |E(FT^n g | Y)|^2 \, d\nu = 0
\]
But \( E(FT^n g | Y) = E(fT^n g | Y) - E(f | Y) E(g | Y) \), as required. \[\square\]

**Lemma 11.5.** Suppose \( (X, \mathcal{B}, \mu, T) \) is weak mixing relative to \( (Y, \mathcal{D}, \nu, S) \) and \( T \) is an ergodic transformation. Then the relative square system \( X \times_Y X \) is also weak mixing relative to \( (Y, \mathcal{D}, \nu, S) \).

**Proof.** Write \( \bar{X} = X \times_Y X \). In particular, let \( \bar{\mu} = \mu \times_Y \mu \) and \( \bar{T} = T \times T \). We are required to prove that the system \( \bar{X} \times_Y \bar{X} \) is ergodic.

Now write \( \bar{X} = \bar{X} \times_Y \bar{X} \) so that \( \bar{\mu} = \bar{\mu} \times_Y \bar{\mu} \) and \( \bar{T} = \bar{T} \times \bar{T} \). By the Mean Ergodic Theorem, specifically Corollary 3.13, it suffices to show that there exists a dense subset \( H \subseteq L^2(\bar{\mu}) \) such that for all \( F, G \in H \)
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \int \bar{F} T^n \bar{G} \, d\bar{\mu} - \int F \, d\bar{\mu} \int G \, d\bar{\mu} = 0
\]  

(11.3)
In particular, it is suffices to show (11.3) holds for functions $F$ and $G$ of the form

$$F(x_1, x_2, x_3, x_4) = \bigoplus_{i=1}^{4} f_i(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} f_i(x_i)$$

$$G(x_1, x_2, x_3, x_4) = \bigoplus_{i=1}^{4} g_i(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} g_i(x_i)$$

where each $f_i, g_i \in L^\infty(X, \mathcal{B}, \mu)$. By applying Lemma 11.1 and Proposition 11.3 we obtain

$$\int FT^n G \, d\hat{\mu} = \int E(f_1 T^n g_1 \oplus f_2 T^n g_2 \mid Y) E(f_3 T^n g_3 \oplus f_4 T^n g_4 \mid Y) \, d\nu$$

$$= \int \prod_{i=1}^{4} E(f_i T^n g_i \mid Y) \, d\nu$$

Similarly,

$$\int F \, d\hat{\mu} = \int \prod_{i=1}^{4} E(f_i \mid Y) \, d\nu$$

and

$$\int G \, d\hat{\mu} = \int \prod_{i=1}^{4} E(g_i \mid Y) \, d\nu$$

By using these equalities to rewrite (11.3) and then applying the triangle inequality it suffices to show the following two statements:

$$\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{4} E(f_i T^n g_i \mid Y) - \prod_{i=1}^{4} E(f_i \mid Y) S^n E(g_i \mid Y) \, d\nu \right| = 0 \quad (11.4)$$

and

$$\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{4} E(f_i \mid Y) S^n E(g_i \mid Y) \, d\nu - \int \prod_{i=1}^{4} E(f_i \mid Y) \, d\nu \int \prod_{i=1}^{4} E(g_i \mid Y) \, d\nu \right| = 0 \quad (11.5)$$

Statement (11.4) is easily deduced by rewriting the integrand as an appropriate sum and applying Lemma 11.1. To verify (11.5) first apply the triangle and Cauchy-Schwarz inequalities to see that for each $N \in \mathbb{N}$ the $N^{th}$ term of the sequence is bounded above by

$$\left\| \prod_{i=1}^{4} E(f_i \mid Y) \right\|_{L^2(\nu)} \left\| \frac{1}{N} \sum_{n=1}^{N} S^n \left( \prod_{i=1}^{4} E(g_i \mid Y) \right) - \int \prod_{i=1}^{4} E(g_i \mid Y) \, d\nu \right\|_{L^2(\nu)}$$

Recall the transformation $T$ is ergodic. This implies $S$ is also ergodic and we are done by the Mean Ergodic Theorem.  

Analogous to the result proved during our study of weak mixing systems, we now show relative weak mixing in some sense implies relative weak mixing of all orders.

**Theorem 11.6.** Suppose $(X, \mathcal{B}, \mu, T)$ is weak mixing relative to $(Y, \mathcal{D}, \nu, S)$. Then for any $k \in \mathbb{N}_{\geq 2}$ and all $f_0, \ldots, f_k \in L^\infty(X, \mathcal{B}, \mu)$ the following hold:

$$\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \left( \prod_{l=0}^{k-1} T^{ln} f_l \mid Y \right) - \prod_{l=0}^{k-1} S^{ln} E(f_l \mid Y) \right|^2 \, d\nu = 0 \quad (11.6)$$

and

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \left( \prod_{l=1}^{k} T^{ln} f_l - \prod_{l=1}^{k} T^{ln} E(f_j \mid Y) \right) \right\|_{L^2(\mu)} = 0 \quad (11.7)$$
Proof. The work for the most part is analogous to that of proof of Theorem 6.4. Once again we use an inductive method; Lemma 11.4 is precisely (11.6) case $k = 2$ whilst (11.7) case $k = 2$ holds for the system $\mathbf{X} \times \mathbf{Y} \times \mathbf{X}$ by the Mean Ergodic Theorem. We shall show for $k \geq 3$

- **Claim (1):** (11.6) case $k - 1 \Rightarrow$ (11.7) case $k$
- **Claim (2):** (11.7) case $k$ for the system $\mathbf{X} \times \mathbf{Y} \times \mathbf{X}$

By Proposition 11.5, if the system $(\mathbf{X}, \mathcal{B}, \mu, T)$ is a relatively weak mixing extension of $(\mathbf{Y}, \mathcal{D}, \nu, S)$ then so is $\mathbf{X} \times \mathbf{Y} \times \mathbf{X}$. Therefore the Theorem follows directly from these two claims.

Claim (1). Suppose for all $g_0, \ldots, g_{k-1} \in L^\infty(\mathbf{X}, \mathcal{B}, \mu)$ the following holds:

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int \left| E \left( \prod_{l=0}^{k-2} T^{in} g_l \mid \mathbf{Y} \right) - \prod_{l=0}^{k-2} S^{in} E(g_l \mid \mathbf{Y}) \right|^2 d\nu = 0 \tag{11.8}
\]

Then given $f_1, \ldots, f_k \in L^\infty(\mathbf{X}, \mathcal{B}, \mu)$ we have

\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} T^{in} f_l - \prod_{l=1}^{k} E(f_l \mid \mathbf{Y})^a \right\|_{L^2(\mu)} = 0
\]

As in the proof of the analogous claim in Theorem 6.4, this is an application of van der Corput’s trick\(^1\). Given $f_1, \ldots, f_k \in L^\infty(\mathbf{X}, \mathcal{B}, \mu)$, first consider the case where for some $j_0 \in \{1, \ldots, k\}$ the function $f_{j_0}$ has zero conditional expectation: $E(f_{j_0} \mid \mathbf{Y}) = 0$. Then by taking

\[
u_n = \prod_{l=1}^{k} T^{in} f_l
\]

it suffices to show

\[
\lim_{H \to \infty} \frac{1}{H} \sum_{h=0}^{H-1} s_h = 0 \tag{11.9}
\]

where

\[
s_h = \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle \nu_{n+h}, \nu_n \rangle \right|
\]

As in the proof of Theorem 6.4 we have

\[
\langle \nu_{n+h}, \nu_n \rangle = \int \prod_{l=0}^{k-1} T^{in} (f_{l+1} T^{(l+1)h} f_{l+1}) d\mu \tag{11.10}
\]

Define $g_{h,l} = f_{l+1} T^{(l+1)h} f_{l+1} \in L^\infty(\mathbf{X}, \mathcal{B}, \mu)$ for each $h \in \mathbb{N}$ and each $l \in \{0, \ldots, k - 1\}$. Then (11.10) becomes

\[
\langle \nu_{n+h}, \nu_n \rangle = \int \prod_{l=0}^{k-1} T^{in} g_{h,l} d\mu = \int E \left( \prod_{l=0}^{k-1} T^{in} g_{h,l} \mid \mathbf{Y} \right) d\nu
\]

\(^1\)The proof is very similar so we work rapidly.
By applying the hypothesis of the claim applied to the collection of functions $g_{h,0}, \ldots, g_{h,k-1}$ for each $h \in \mathbb{N}$ observe

$$s_h = \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \langle u_{n+h}, u_n \rangle \right|$$

$$= \limsup_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \int \left( \prod_{l=0}^{k-1} T_l^{n} g_{h,l} \mid Y \right) \, d\nu \right|$$

$$= \int \prod_{l=0}^{k-1} S_l^{n} E(g_{h,l} \mid Y) \, d\nu \quad (11.11)$$

We claim that for some set $Z \subset \mathbb{N}$ of zero density

$$\lim_{h \to \infty, h \notin Z} s_h = 0 \quad (11.12)$$

and thus, by Theorem 2.1, establish (11.9). Indeed, as $(X, \mathcal{A}, \mu, T)$ is weak mixing relative to $(Y, \mathcal{B}, \nu, S)$, we may apply Lemma 11.4 to the function $f_{j_0}$ to obtain

$$\lim_{H \to \infty} \frac{1}{H} \sum_{h=1}^{H} \left| E(f_{j_0} T^h f_{j_0} \mid Y) - E(f_{j_0} \mid Y) S^h E(f_{j_0} \mid Y) \right|^2 \, d\nu = 0$$

By hypothesis $E(f_{j_0} \mid Y) = 0$ and so

$$\lim_{h \to \infty, h \notin Z} \int \left| E(g_{h,j_0-1} \mid Y) \right|^2 \, d\nu = 0 \quad (11.13)$$

for some set $Z \subset \mathbb{N}$ of zero density. Finally, applying the Cauchy Schwarz inequality to (11.11) gives

$$|s_h| \leq \left\| E(g_{h,j_0-1} \mid Y) \right\|_{L^2(\nu)} \left\| \prod_{l=0}^{k-1} S_l^{n} E(g_{h,l} \mid Y) \right\|_{L^2(\nu)}$$

$$\leq \left\| E(g_{h,j_0-1} \mid Y) \right\|_{L^2(\nu)} \prod_{l=0}^{k-1} \left\| f_l \right\|_{\infty}^2 \quad (11.14)$$

Thus, by comparing (11.13) and (11.14) we obtain (11.12) as desired.

The general case reduces to the above case by writing

$$\prod_{l=1}^{k} T_l^{n} f_l - \prod_{l=1}^{k} E(f_l \mid Y)^\alpha = \sum_{j=1}^{k} \left( \prod_{l=1}^{j-1} T_l^{n} f_l \right) T_j^{n} \left( f_j - E(f_j \mid Y)^\alpha \right) \left( \prod_{l=j+1}^{k} E(f_l \mid Y)^\alpha \right) \quad (11.15)$$

and noting $E(f_j - E(f_j \mid Y)^\alpha \mid Y) = E(f_j \mid Y) - E(f_j \mid Y) = 0$.

For the second claim we turn our attention to the system $X \times Y$. We adopt the notation $\hat{\mu} = \mu \times \mathcal{Y} \mu$ and $\hat{T} = T \times T$ and let $\hat{\alpha} : X \times Y \to Y$ denote the weak mixing extension given by composing $\alpha$ with the projection onto the first co-ordinate.

**Claim (2).** Suppose for all $f_1, \ldots, f_k \in L^{\infty}(X \times \mathcal{Y} X)$ the following holds:

$$\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} \hat{T}_l^{n} f_l - \prod_{l=1}^{k} \hat{T}_l^{n} E(f_l \mid Y)^\hat{\alpha} \right\|_{L^2(\hat{\mu})} = 0 \quad (11.16)$$

Then given $g_0, \ldots, g_k \in L^{\infty}(X, \mathcal{A}, \mu)$ we have

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \int \left| E\left( \prod_{l=0}^{k-1} T_l^{n} g_{l} \mid Y \right) - \prod_{l=0}^{k-1} S_l^{n} E(g_{l} \mid Y) \right|^2 \, d\nu = 0 \quad (11.17)$$

57
Here we use a slightly different approach than that utilised in dealing with the analogous claim in the proof of Theorem 6.4. The following argument is adapted from ([8] Chapter 7).

Let \( g_0, \ldots, g_k \in L^\infty(X, \mathscr{A}, \mu) \). First we consider the case where \( E(g_{j_0} \mid Y) = 0 \) for some \( j_0 \in \{0, \ldots, k\} \). By Proposition 11.5 the relative square system \( X \otimes \otimes X \) is weak mixing relative to \((Y, \mathscr{D}, \nu, S)\). Apply the hypothesis (11.16) of the claim to the functions \( g_1 \otimes g_1, \ldots, g_k \otimes g_k \) so that

\[
\lim_{N \to \infty} \left\| \frac{1}{N} \sum_{n=1}^{N} \prod_{l=1}^{k} \hat{T}^{ln} g_l \otimes g_l - \prod_{l=1}^{k} \hat{T}^{ln} E(g_l \otimes g_l \mid Y) \hat{\alpha} \right\|_{L^2(\hat{\mu})} = 0
\]

Norm convergence implies weak convergence so that

\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} \int g_0 \otimes g_0 \left( \prod_{l=1}^{k} \hat{T}^{ln} g_l \otimes g_l \right) \, d\hat{\mu} - \int g_0 \otimes g_0 \left( \prod_{l=1}^{k} \hat{T}^{ln} E(g_l \otimes g_l \mid Y)^{\alpha} \right) \, d\hat{\mu} \right| = 0
\]

We may rewrite the left hand integral as

\[
\int \prod_{l=0}^{k} \hat{T}^{ln} g_l \otimes \prod_{l=0}^{k} \hat{T}^{ln} g_l \, d\hat{\mu} = \int E \left( \prod_{l=0}^{k} T^{ln} g_l \mid Y \right)^{2} \, d\nu
\]

Do the same for the right hand integral and apply the assumption \( E(g_{j_0} \mid Y) = 0 \) to give

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left| \int E \left( \prod_{l=0}^{k} T^{ln} g_l \mid Y \right)^{2} \, d\nu \right| = 0
\]

Thus we have established the result for this case. Note that by Theorem 2.1 we also have

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left\| E \left( \prod_{l=0}^{k} T^{ln} g_l \mid Y \right) \right\|_{L^2(\nu)} = 0 \tag{11.18}
\]

For the general case observe

\[
E \left( \prod_{l=0}^{k-1} T^{ln} g_l \mid Y \right) \prod_{l=0}^{k-1} S^{ln} E(g_l \mid Y) = E \left( \prod_{l=0}^{k-1} T^{ln} g_l - \prod_{l=0}^{k-1} T^{ln} E(g_l \mid Y)^{\alpha} \mid Y \right)
\]

Using the identity stated in (11.15) (replacing the \( f_t \) with \( g_t \)), the above expression may be written as

\[
\sum_{j=1}^{k} E \left( \prod_{l=1}^{k} T^{ln} G_{j,l} \mid Y \right) \tag{11.19}
\]

where

\[
G_{j,l} = \begin{cases} f_l & \text{if } l = 1, \ldots, j-1 \\ f_j - E(f_j \mid Y)^{\alpha} & \text{if } l = j \\ E(f_l \mid Y)^{\alpha} & \text{if } l = j+1, \ldots, k 
\end{cases}
\]

For each \( j = 1, \ldots, k \) we have \( E(G_{j,l} \mid Y) = 0 \) and so, by taking the \( L^2 \)-norm of (11.19) and applying the triangle inequality we can reduce the general case to the previous case. Specifically, by applying (11.18) to each collection of functions \( G_{j,1}, \ldots, G_{j,k} \) it follow

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \left\| E \left( \prod_{l=0}^{k-1} T^{ln} g_l \mid Y \right) - \prod_{l=0}^{k-1} S^{ln} E(g_l \mid Y) \right\|_{L^2(\nu)} = 0
\]

and we obtain (11.17) by once again applying Theorem 2.1. \( \square \)
We now use the previous theorem to deduce the main result of this section: that the SZ property is preserved under weak mixing extensions.

**Theorem 11.7.** Suppose \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) is a weak mixing extension and that the action of \( S \) on \((Y, \mathcal{D}, \nu, S)\) is SZ. Then so too is the action of \( T \) on \((X, \mathcal{B}, \mu, T)\).

**Proof.** Let \( A \in \mathcal{B} \) be a set with positive measure \( \mu(A) > 0 \) and \( k \in \mathbb{N} \). Define

\[
D_n := \left\{ y \in Y : \mu_y(A) \geq \frac{1}{n} \right\}
\]

Then each \( D_n \) is \( \nu \)-measurable and

\[
\lim_{n \to \infty} \left| \int_{D_n} \mu_y(A) d\nu(y) - \mu(A) \right| = 0
\]

and hence there must exist \( n_0 \in \mathbb{N} \) such that \( \nu(D_{n_0}) > 0 \). In other words, there exists a constant \( a > 0 \) such that if we define \( D := \{ y \in Y : \mu_y(A) \geq a \} \) then \( \nu(D) > 0 \).

Let \( \epsilon > 0 \). By Theorem 11.6 there exists an \( N_0 \in \mathbb{N} \) such that

\[
\frac{1}{N} \sum_{n=1}^{N} \left\| E\left( \prod_{l=0}^{k} T^{ln} 1_A \big| Y \right) - \prod_{l=0}^{k} S^{ln} E(1_A \big| Y) \right\|_{L^2(\nu)} < \epsilon
\]

for all \( N > N_0 \)

Thus, fixing \( N > N_0 \), by applying the triangle and Cauchy-Schwarz inequalities we have

\[
\left| \frac{1}{N} \sum_{n=1}^{N} \int E\left( \prod_{l=0}^{k} T^{ln} 1_A \big| Y \right) d\nu - \int \prod_{l=0}^{k} S^{ln} E(1_A \big| Y) d\nu \right| < \epsilon
\]

Note that

\[
\mu\left( \bigcap_{l=0}^{k} T^{-ln} A \right) = \int E\left( \prod_{l=0}^{k} T^{ln} 1_A \big| Y \right) d\nu
\]

and as \( E(1_A \big| Y) \geq 0 \) it follows \( E(1_A \big| Y) \geq a1_D \). Thus

\[
\frac{1}{N} \sum_{n=1}^{N} \mu\left( \bigcap_{l=0}^{k} T^{-ln} A \right) > \frac{1}{N} \sum_{n=1}^{N} \int \prod_{l=0}^{k} S^{ln} E(1_A \big| Y) d\nu - \epsilon
\]

\[
> a^{k+1} \frac{1}{N} \sum_{n=1}^{N} \int \prod_{l=0}^{k} S^{ln} 1_D d\nu - \epsilon
\]

\[
> a^{k+1} \frac{1}{N} \sum_{n=1}^{N} \nu\left( \bigcap_{l=0}^{k} S^{-ln} D \right) - \epsilon
\]

As \( \epsilon \) was arbitrary and the SZ property holds for the system \((Y, \mathcal{D}, \nu, S)\) it follows

\[
\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu\left( \bigcap_{l=0}^{k} T^{-ln} A \right) \geq \liminf_{N \to \infty} a^{k+1} \frac{1}{N} \sum_{n=1}^{N} \nu\left( \bigcap_{l=0}^{k} S^{-ln} D \right) > 0
\]

as required.  

\[ \square \]

12 Compact Extensions

We now turn to defining compact extensions and showing that they too preserve the SZ property.
\textbf{Definition 25.} Let \(\alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S)\) be an extension of a measure preserving system and \(f \in L^2(X, \mathcal{B}, \mu)\). We say \(f\) is almost periodic relative to \(Y\) \((\text{AP rel. } Y)\) if for every \(\epsilon > 0\) there exists functions \(g_1, \ldots, g_m \in L^2(X, \mathcal{B}, \mu)\) such that
\[
\min_{1 \leq j \leq m} \| T^n f - g_j \|_{L^2(\mu)} < \epsilon \quad \text{almost everywhere } y \in Y
\]
Moreover, the extension is compact if every the set of AP rel. \(Y\) functions is dense in \(L^2(X, \mathcal{B}, \mu)\).

Complementing the result of the previous section, the purpose of this section will be to prove the following theorem:

\textbf{Theorem 12.1.} Let \(\alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S)\) be a compact extension and suppose the action of \(S\) on \((Y, \mathcal{D}, \nu)\) is SZ. Then so is the action of \(T\) on \((X, \mathcal{B}, \mu)\).

The proof of this result builds upon our work on compact systems in a far from trivial manner. The proof is divided into a number of results: the first reduces the problem, and the remainder to some extent provide an analog of the method used to show the SZ property for compact systems in Theorem 7.2.

Recall that in order to prove Theorem 12.1 we must show that for all \(A \in \mathcal{B}\) with \(\mu(A) > 0\) and for all \(k \in \mathbb{N}\)
\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mu \left( \bigcap_{j=0}^{k-1} T^{-jn} A \right) > 0 \quad (12.1)
\]
We begin by demonstrating that without loss of generality we may only consider \(A \in \mathcal{B}\) with certain “nice” properties. The following lemma is adapted from ([5], Chapter 7).

\textbf{Lemma 12.2.} Let \(\alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S)\) be a compact extension and \(A \in \mathcal{B}\) a set of positive measure. Then there exists a measurable subset \(\tilde{A} \subseteq A\) of positive measure such that
1. For almost every \(y \in Y\) either \(\mu_y(\tilde{A}) = 0\) or \(\mu_y(\tilde{A}) > \frac{1}{2} \mu(\tilde{A})\).
2. The characteristic function \(1_{\tilde{A}} \in L^2(X, \mathcal{B}, \mu)\) is A.P. rel. \(Y\).

\textbf{Remark 12.} Clearly (12.1) holds for \(A\) automatically if it holds for any measurable subset \(\tilde{A} \subseteq A\). Thus, in proving Theorem 12.1 we may restrict our attention without loss of generality to showing (12.1) holds for sets \(A \in \mathcal{B}\) with \(\mu(A) > 0\) which satisfy properties 1 and 2.

\textbf{Proof of Lemma 12.2.} To begin we simply remove the parts of the set which contradict property 1. To do this let
\[
B := \{ y \in Y : \mu_y(A) \leq \frac{1}{2} \mu(A) \}
\]
Then \(B\) is measurable and hence so is \(A' = A \setminus \alpha^{-1}(B)\). For almost every \(y \in Y \setminus B\) the set \(\alpha^{-1}(B)\) is \(\mu_y\)-null and so
\[
\mu_y(A') = \mu_y(A) > \frac{1}{2} \mu(A) \geq \frac{1}{2} \mu(A')
\]
On the other hand, for almost every \(y \in B\) then the support of the measure \(\mu_y\) is contained in \(\alpha^{-1}(B)\) and so
\[
\mu_y(A') = 0
\]
Hence for almost every \(y \in Y\) either \(\mu_y(A') > \frac{1}{2} \mu(A')\) or \(\mu_y(A') = 0\). It remains to show \(A'\) is a set of positive measure. If \(y \in B\) then \(\mu_y(A \setminus A') = \mu_y(A) \leq \frac{1}{2} \mu(A)\) and if \(y \in Y \setminus B\) then \(\mu_y(A \setminus A') = 0\). Integrating gives
\[
\mu(A \setminus A') = \int \mu_y(A \setminus A') \, d\nu(y) \leq \frac{1}{2} \mu(A)
\]
and so \(\mu(A') \geq \frac{1}{2} \mu(A) > 0\).

We claim there exists a set \(B' \in \mathcal{D}\) such that if we define \(\tilde{A} = A' \setminus \alpha^{-1}(B')\) then \(\mu(\tilde{A}) > 0\) and the function \(1_{\tilde{A}}\) is AP rel \(Y\). This set then also has property 1: for almost every \(y \in B'\)
we have \( \mu_y(\tilde{A}) = 0 \). On the other hand, for almost every \( y \notin B' \) it follows \( \mu_y(\tilde{A}) = \mu_y(A') > \frac{1}{2} \mu(A') \geq \frac{1}{2} \mu(\tilde{A}) \).

Let \((\epsilon_l)_{l \in \mathbb{N}}\) be the sequence \( \epsilon_l = \frac{1}{2^n} \mu(A') \) so that

\[
\sum_{l=1}^{\infty} \epsilon_l = \frac{1}{2} \mu(A') < \frac{1}{2} \mu(A') \quad (12.2)
\]

By hypothesis the set of AP rel \( Y \) functions is dense in \( L^2(X, \mathcal{B}, \mu) \). For every \( l \in \mathbb{N} \) there exists an \( f_l \in L^2(X, \mathcal{B}, \mu) \) such that \( f_l \) is AP rel \( Y \) and

\[
\|1_{A'} - f_l\|_{L^2(\mu)} < \epsilon_l
\]

Now consider the sets

\[
B_l = \{ y \in Y : \|1_{A'} - f_l\|_{L^2(\mu_y)}^2 \geq \epsilon_l \}
\]

defined for every \( l \in \mathbb{N} \). Then each \( B_l \) is measurable and

\[
\nu(B_l) = \int_{B_l} d\nu(y) \leq \frac{1}{\epsilon_l} \int_{B_l} \|1_{A'} - f_l\|_{L^2(\mu_y)}^2 d\nu(y) \\
\leq \frac{1}{\epsilon_l} \int \|1_{A'} - f_l\|_{L^2(\mu_y)}^2 d\nu(y) = \frac{1}{\epsilon_l} \|1_{A'} - f_l\|_{L^2(\mu)}^2 < \epsilon_l 
\]  

(12.3)

Define \( \tilde{A} = A' \setminus \alpha^{-1}(\cup_{l \in \mathbb{N}} B_l) \). By (12.2) and (12.3) one may easily deduce \( \mu(\tilde{A}) > \frac{1}{2} \mu(A) \). We claim \( 1_{\tilde{A}} \) is AP rel \( Y \). Indeed, given \( \epsilon > 0 \) there exists \( l_0 \in \mathbb{N} \) such that \( \epsilon_l < \frac{1}{2} \epsilon \) and functions \( g_1, \ldots, g_m \in L^2(X, \mathcal{B}, \mu) \) such that

\[
\min_{1 \leq j \leq m} \|T^n f_{l_0} - g_j\|_{L^2(\mu_y)} < \frac{\epsilon}{2} \quad \text{for all } n \in \mathbb{N} \text{ and almost everywhere } y \in Y
\]

Note, for almost every \( y \in Y \) if \( S^ny \notin \bigcup_{l \in \mathbb{N}} B_l \) then

\[
\|T^n 1_{\tilde{A}} - T^n f_l\|_{L^2(\mu_y)} = \|1_{\tilde{A}} - f_l\|_{L^2(\mu_y)} = \|1_{A'} - f_l\|_{L^2(\mu_y)} < \frac{\epsilon}{2}
\]

On the other hand, if \( S^ny \in \bigcup_{l \in \mathbb{N}} B_l \) then

\[
\|T^n 1_{\tilde{A}}\|_{L^2(\mu_y)} = \|1_{\tilde{A}}\|_{L^2(\mu_y)} = 0
\]

If we take \( g_0 \equiv 0 \) then by the triangle inequality

\[
\min_{0 \leq j \leq m} \|T^n 1_{\tilde{A}} - g_j\|_{L^2(\mu_y)} \leq \min_{0 \leq j \leq m} \|T^n 1_{\tilde{A}} - T^n f_l\|_{L^2(\mu_y)} + \|T^n f_l - g_j\|_{L^2(\mu_y)} < \epsilon
\]

as required. \( \Box \)

At the heart of the multiple recurrence result for compact systems was the idea expounded in Proposition 7.1, which we presently recall. If a system \((X, \mathcal{B}, \mu, T)\) is compact, then for all \( f \in L^2(X, \mathcal{B}, \mu) \) A.P. there exists a set \( K \subset \mathbb{N} \) of positive lower density (in particular a syndetic set) such that

\[
\|T^nf - f\|_{L^2(\mu)} < \epsilon \quad \text{for all } n \in K \quad (12.4)
\]

Following the argument given in [5, 10], our recourse in proving Theorem 12.1 will be to formulate and prove a similar statement but concerning compact extensions \( \alpha : (X, \mathcal{B}, \mu, T) \rightarrow (Y, \mathcal{B}, \nu, S) \) and A.P. rel. \( Y \) functions and in particular we consider \( f = 1_{A'} \).
Proposition 12.3. Let $f = 1_A$ where we assume $A \in \mathcal{B}$ is of positive measure and satisfies properties 1 and 2 of Lemma 12.2 and let $\epsilon > 0$ be given. There exists a set $D \in \mathcal{D}$ of positive measure and finite set $F \subset \mathbb{Z}$ such that given any $n \in \mathbb{N}$, for any $y \in \bigcap_{l=0}^{k} S^{-lm} D$ there exists an $m \in F$ such that

$$\|T^{(n+m)} f - f\|_{L^2(\mu_y)} < \epsilon \quad \text{for all } l \in \{1, \ldots, k\}$$

(12.5)

As in the proof of Proposition 7.1 we consider finite maximum $\epsilon$–separated sets. The main problem we encounter here is that we now need to find finite sets

$$\{T^{t_{1}}, f, \ldots, T^{t_{r}} f\} \subset L^2(X, \mathcal{B}, \mu_y) \quad \text{for all } y \in \bigcap_{l=0}^{k} S^{-lm} D$$

which are maximal $\epsilon$–separated in many different vector spaces. To this end we make the following definition:

Definition 26. For each $y \in Y$ let $\bigoplus_{l=0}^{k} L^2(\mu_y)$ denote the vector space of $(k+1)$–tuples $(f_0, \ldots, f_k)$, where each $f_i \in L^2(\mu_y)$ with the addition and scalar multiplication defined in the obvious way. Define a norm $\| \cdot \|_{\oplus}$ on $\bigoplus_{l=0}^{k} L^2(\mu_y)$ by

$$\|(f_0, \ldots, f_k)\|_{\oplus} = \max_{0 \leq i \leq k} \|f_i\|_{L^2(\mu_y)}$$

Proof of Proposition 12.3. We devide the proof into three claims, first focusing our attention on subsets of the following form

$$\mathcal{L}(y) := \{ (f, T^{m} f, \ldots, T^{km} f) : f \in \bigoplus_{l=0}^{k} L^2(\mu_y) : m \in \mathbb{Z}, \mu \}$$

By the proof of Lemma 12.2 there exists a set $D' \in \mathcal{D}$ of positive measure such that $\mu_y(A) > \frac{1}{2} \mu(A)$ for almost every $y \in D'$ and $\mu_y(A) = 0$ for almost every $y \in Y \setminus D'$. Consider the subset $\mathcal{L}^*(y) \subseteq \mathcal{L}(y)$ given by

$$\mathcal{L}^*(y) := \{ (f, T^{m} f, \ldots, T^{km} f) : m \in \mathbb{Z}, y \in \bigcap_{l=0}^{k} S^{-lm} D' \}$$

Explicitly, for any $m \in \mathbb{Z}$, $(f, T^{m} f, \ldots, T^{km} f) \in \mathcal{L}^*(y)$ provided $y \in \bigcap_{l=0}^{k} S^{-lm} D'$. This is precisely the subset consisting of the $(f, T^{m} f, \ldots, T^{km} f)$ for which each entry has non-zero $L^2(\mu_y)$-norm. Indeed, suppose $(f, T^{m} f, \ldots, T^{km} f) \in \mathcal{L}^*(y)$ then for any $l \in \{1, \ldots, k\}$ we have $S^{lm} y \in D'$ and so

$$\|T^{lm} f\|_{L^2(\mu_y)} = \|f\|_{L^2(\mu_{S^lm y})} > \sqrt{\frac{1}{2} \mu(A)} > 0$$

Claim (1). Let $\epsilon > 0$ be given. Then there exists a (non-empty) finite set $F_0 \subset \mathbb{Z}$ such that

$$\{ (f, T^{m} f, \ldots, T^{km} f) : m \in F_0 \}$$

is maximal $\epsilon$–separated in $\mathcal{L}^*(y)$ for all $y \in \bigcap_{l=0}^{k} S^{-lm} D'$.

Proof. As $f$ is AP rel $Y$, by definition the set $\{T^n f : n \in \mathbb{N}\}$ is totally bounded in $L^2(X, \mathcal{B}, \mu_y)$. It follows that $\mathcal{L}(y)$ and, moreover, $\mathcal{L}^*(y)$ are totally bounded in $\bigoplus_{l=0}^{k} L^2(\mu_y)$.

For any finite set $F \subset \mathbb{Z}$ let Sep$_F$ be the set of all $y \in D'$ such that the set

$$\{ (f, T^{m} f, \ldots, T^{km} f) : m \in F \}$$

(12.6)
is maximal \( \epsilon \)-separated in \( \mathcal{L}^*(y) \). Alternatively, define the function \( g_F \) on \( Y \) by

\[
g_F(y) = \min_{m \neq m' \in F} \max_{j=0, \ldots, k} \|T^{im}f - T^{im'}f\|_{L^2(\mu_y)} \quad \text{if } y \in \bigcap_{m \in F} \bigcap_{j=0}^{k} S^{-jm}D \tag{12.7}
\]

\[
g_F(y) = 0 \quad \text{otherwise}
\]

If the value of \( g_F(y) > 0 \), then \( g_F(y) \) gives the separation of (12.6) in \( \mathcal{L}(y) \) and it follows \( \text{Sep}_\epsilon(F) \subset Y \) can be expressed in the following form

\[
\text{Sep}_\epsilon(F) = \{ y \in Y : g_F(y) > \epsilon \text{ but } g_F(y) \leq \epsilon \text{ whenever } |\tilde{F}| > |F| \}
\]

We claim that there exists some finite set \( F_0 \subset \mathbb{Z} \) such that \( \nu(\text{Sep}_\epsilon(F)) > 0 \). To prove the claim it suffices to show the set system \( \{\text{Sep}_\epsilon(F) : F \subset \mathbb{Z} \text{ finite} \} \) is (almost) a covering for \( \mathcal{D}' \). As there are only countably many finite subsets of integers, all the \( \text{Sep}_\epsilon(F) \) cannot have zero measure for otherwise

\[
0 < \nu(D') \leq \sum_{F \subset \mathbb{Z} \text{ finite}} \nu(\text{Sep}_\epsilon(F)) = 0
\]

For any \( y \in \mathcal{D}' \), as \( (f, \ldots, f) \in \mathcal{L}'(y) \) the set \( \mathcal{L}'(y) \) is non-empty and totally bounded and thus there exists a maximal \( \epsilon \)-separated subset of \( \mathcal{L}'(y) \). It follows directly that there exists \( F_y \subset \mathbb{Z} \) finite such that \( y \in \text{Sep}_\epsilon(F_y) \) and we are done.

Henceforth we fix \( F = F_0 \subset \mathbb{Z} \) as a set of integers such that \( \nu(\text{Sep}_\epsilon(F)) > 0 \). Note that for each \( l \in \mathbb{N} \)

\[
\text{Sep}_\epsilon^l(F) := \{ y \in Y : g_F(y) > \epsilon + \frac{1}{l} \text{ but } g_F(y) \leq \epsilon \text{ whenever } |\tilde{F}| > |F| \}
\]

is a measurable subset of \( \text{Sep}_\epsilon(F) \) and \( \text{Sep}_\epsilon(F) = \bigcup_{l \in \mathbb{N}} \text{Sep}_\epsilon^l(F) \). There must be some \( l_0 \in \mathbb{N} \) such that \( \nu(\text{Sep}_\epsilon^{l_0}(F)) > 0 \). Thus, defining \( \eta = \frac{1}{l_0} \) the set

\[
\tilde{D} := \{ y \in Y : g_F(y) > \epsilon + \eta \text{ but } g_F(y) \leq \epsilon \text{ whenever } |\tilde{F}| > |F| \}
\]

is of positive measure.

**Claim** (2). **There exists a subset \( D \subseteq \tilde{D} \) of positive measure such that for each pair \( m, m' \in F \), each \( l = 0, \ldots, k \) and each pair \( y, y' \in D \)

\[
\|T^{im}f - T^{im'}f\|_{L^2(\mu_y)} - \|T^{im}f - T^{im'}f\|_{L^2(\mu_y')} < \eta
\]

Proof. For each choice of \( m, m' \in F \) and \( l \in \{0, \ldots, k\} \) define the function \( \theta(m, m', l) : \tilde{D} \to [0, 1] \) by

\[
\theta(m, m', l) : y \mapsto \|T^{ml}f - T^{ml'}f\|_{L^2(\mu_y)}
\]

Then these functions are measurable. Let \( I_1, \ldots, I_l \) be a partition of \([0, 1]\) into disjoint closed intervals, each of length < \( \eta \). Note that for each choice of \( m, m' \) and \( l \) the sets

\[
\theta(m, m', l)^{-1}(I_1), \ldots, \theta(m, m', l)^{-1}(I_k)
\]

partition \( \tilde{D} \) into a finite number of measurable sets. Let \( D_1, \ldots, D_M \subseteq \tilde{D} \) be the partition which refines the collection of partitions corresponding to every choice of \( m, m' \) and \( l \). Then each \( D_i \) is an intersection of measurable sets and \( \nu(\tilde{D}) = \sum_{i=1}^{M} \nu(D_i) \). There must be some set \( D = D_{i_0} \) with positive measure. It is easy to check \( D \) satisfies the required property. \( \square \)

Finally, Proposition 12.3 proceeds directly from the following claim.
Claim (3). For any \( n \in \mathbb{N} \) the set

\[ B_n := \{(f, T^{n+m} f, \ldots, T^{k(n+m)} f) : m \in F \} \]

is an \( \epsilon \)-separated subset of \( L^*(y) \) for all \( y \in \bigcap_{j=0}^k S^{-jn} D \).

Once we have proven Claim 3, fixing \( n \in \mathbb{N} \) and \( y \in \bigcap_{j=0}^k S^{-jn} D \subseteq \tilde{D} \), it follows from the definition of \( \tilde{D} \) given in (12.8) that for the set \( B_n \) is maximal \( \epsilon \)-separated in \( L^*(y) \). In particular, if we define \( \tilde{F} = (n \cdot F) \cup \{0\} \) then as \( |\tilde{F}| > |F| \), the set \( \{ (f, \ldots, f) \} \cup B_n \) cannot be \( \epsilon \)-separated in \( L^*(y) \). Moreover, we must have some \( m \in F \) such that for each \( l = 0, \ldots, k \)

\[ \|T^{l(m+n)} f - f\|_{L^2(\mu_\nu)} \leq \| (f, T^{m+n} f, \ldots, T^{k(m+n)} f) - (f, \ldots, f)\|_{\tilde{D}} < \epsilon \]

Proof (of Claim 3). By hypothesis the action of \( S \) on \( (Y, \mathcal{B}, \nu) \) is SZ, and as \( \nu(D) > 0 \) this ensures that the set \( \bigcap_{j=0}^k S^{-jn} D \) in non-empty for infinitely many values of \( n \) and therefore the claim is not vacuous. Explicitly,

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^N \nu \left( \bigcap_{j=0}^k S^{-jn} D \right) > 0 \]

and so there exists a set of positive lower density such that for all \( n \) belonging to this set

\[ \nu \left( \bigcap_{j=0}^k S^{-jn} D \right) > 0 \]  \hspace{1cm} (12.9)

Fix \( n \in \mathbb{N} \) such that \( \bigcap_{j=0}^k S^{-jn} D \) is non-empty. We will demonstrate that the set \( B_n \) has the desired properties.

First we show that \( B_n \subseteq L^*(y) \). Recalling the definition of \( L^*(y) \), this is true provided that for each \( y \in \bigcap_{l=0}^k S^{-ln} D \) we have \( y \in \bigcap_{l=0}^k S^{-l(m+n)} D' \) for every \( m \in F \).

Let \( y \in \bigcap_{l=0}^k S^{-ln} D \), then \( S^{ln} y \in D \) for every \( l \in \{0, \ldots, k\} \) and so it follows \( g_F(S^{ln} y) > \epsilon + \eta > 0 \). We must have \( S^{ln} y \in \bigcap_{m \in F, l} S^{-jm} D' \) by the definition of \( g_F \). In particular \( S^{ln} y \in \bigcap_{m \in F} S^{-lm} D' \) and so \( S^{ln} y \in \bigcap_{m \in F} S^{-lm} D' \) as required.

Now we show the \( \epsilon \)-separation. As \( y \in D \subseteq \tilde{D} \), if \( m \neq m' \in F \) then

\[ \|T^{lm} f - T^{lm'} f\|_{L^2(\mu_\nu)} > \epsilon + \eta \]  \hspace{1cm} (12.10)

On the other hand, by Claim 2,

\[ \| T^{lm} f - T^{lm'} f\|_{L^2(\mu_\nu)} = \| (T^{lm} f - T^{lm'} f)\|_{L^2(S^{ln} y)} < \eta \]  \hspace{1cm} (12.11)

and so combining (12.10) and (12.11) gives

\[ \|T^{ln} f - T^{lm} f\|_{L^2(\mu_\nu)} > \epsilon \]

But

\[ \|T^{l(n+m)} f - T^{l(n+m')} f\|_{L^2(y)} = \|T^{ln}(T^{lm} f - T^{lm'} f)\|_{L^2(y)} = \|T^{lm} f - T^{lm'} f\|_{L^2(S^{ln} y)} < \epsilon \]

as required.

This concludes the proof of Proposition 12.3. 

We now turn to the proof of Theorem 12.1. Recall we are trying to show if \( \alpha : (X, \mathcal{B}, \mu, T) \rightarrow (Y, \mathcal{D}, \nu, S) \) is a compact extension and the action of \( S \) on \( (Y, \mathcal{D}, \nu, S) \) is SZ then it follows that the action of \( T \) on \( (X, \mathcal{B}, \mu, T) \) is SZ.
Proof (of Theorem 12.1). Take $\epsilon = \frac{1}{4(k+1)}\mu(A)$ in Proposition 12.3 to obtain a set $D \in \mathcal{F}$ of positive measure and finite set $F \subset \mathbb{Z}$ with the stated properties. Temporarily fixing $n \in \mathbb{N}$ and $y \in \bigcap_{l=0}^{k} S^{-ln}D$ there exists $m \in F$ such that

$$\|T^{l(n+m)}f - f\|_{L^2(\mu_y)} < \epsilon \quad \text{for all } l \in \{0, \ldots, k\}$$

Thus, for any $l \in \{0, \ldots, k\}$ we have

$$\left| \int \prod_{l=0}^{k} T^{l(n+m)} f \, d\mu_y - \int f^{k+1} \, d\mu_y \right| = \left| \int \sum_{j=0}^{k} \prod_{l=0}^{j-1} T^{l(n+m)} f |T^{j(n+m)} f - f| f^{k-j} \, d\mu_y \right|
\leq \sum_{j=0}^{k} \int \prod_{l=0}^{j-1} T^{l(n+m)} f |T^{j(n+m)} f - f| f^{k-j} \, d\mu_y
\leq \sum_{j=0}^{k} \int |T^{j(n+m)} f - f| f^{k-j} \, d\mu_y$$

By the Cauchy Schwarz inequality $\int |T^{j(n+m)} f - f| \, d\mu_y \leq \|T^{j(n+m)} f - f\|_{L^2(\mu_y)} < \epsilon$. Hence

$$\left| \int \prod_{l=0}^{k} T^{l(n+m)} f \, d\mu_y - \int f^{k+1} \, d\mu_y \right| < (k+1)\epsilon$$

By our choice of $\epsilon$ and property 1 of the set $A$ as given in Lemma 12.2 we have

$$\int \prod_{l=0}^{k} T^{l(n+m)} f \, d\mu_y \geq \int f^{k+1} \, d\mu_y - (k+1)\epsilon \quad (12.12)$$

Now the choice of $m$ does depend on $y \in \bigcap_{l=0}^{k} S^{-ln}D$, but now we exploit the fact there are only finitely many possible choices of $m$. Consider the sum

$$\sum_{m \in F} \int \prod_{l=0}^{k} T^{l(n+m)} f \, d\mu_y$$

For any $y \in \bigcap_{l=0}^{k} S^{-ln}D$, the inequality (12.12) must hold for some term in this summation and all terms are non-negative. Hence

$$\sum_{m \in F} \int \prod_{l=0}^{k} T^{l(n+m)} f \, d\mu_y \geq \frac{1}{4}\mu(A) \quad (12.13)$$

Integrating (12.13) over $\bigcap_{l=0}^{k} S^{-ln}D$ therefore gives

$$\sum_{m \in F} \int \prod_{l=0}^{k} T^{l(n+m)} f \, d\mu_y \geq \frac{1}{4}\mu(A) \nu\left(\bigcap_{l=0}^{k} S^{-ln}D\right)$$

Averaging both sides over $1 \leq n \leq N$ and taking the limit as $N \to \infty$ gives

$$|F| \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \nu\left(\bigcap_{l=0}^{k} T^{-ln}A\right) \geq \frac{1}{4}\mu(A) \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \nu\left(\bigcap_{l=0}^{k} S^{-ln}D\right) > 0$$

$\square$

65
13 The Dichotomy Between Weak Mixing and Compact Extensions

We are now able to conclude the proof of Furstenberg’s Multiple Recurrence Theorem. The final piece of the argument is the following Theorem:

**Theorem 13.1.** Let \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) be an extension of measure preserving systems. Then one of the following holds:

- Either \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) is a weak mixing extension,
- Or there exists a non-trivial intermediate compact extension. Explicitly, there exists a system \( (Z, \mathcal{E}, \theta, R) \) defined on a Borel Probability space and an extension \( \beta : (X, \mathcal{B}, \mu, T) \to (Z, \mathcal{E}, \theta, R) \)

and a non-trivial compact extension

\[
\gamma : (Z, \mathcal{E}, \theta, R) \to (Y, \mathcal{D}, \nu, S)
\]

with the property \( \alpha = \gamma \beta \).

Once Theorem 13.1 has been established the Multiple Recurrence Theorem follows directly as we shall now illustrate:

**Proof (of Furstenberg’s Multiple Recurrence Theorem).** Suppose \( (X, \mathcal{B}, \mu, T) \) is an invertible, ergodic measure preserving system defined on a Borel probability space. By Theorem 10.1 there exists some maximal SZ factor \( (X, \mathcal{A}, \mu, T) \), say. Suppose this is a proper factor. Recall that there exists some system \( (Y, \mathcal{D}, \nu, S) \) defined on a Borel probability space corresponding to this maximal factor, namely there is an extension \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) such that \( \mathcal{A} \subseteq \alpha^{-1}(\mathcal{D}) \). Now this extension cannot be weak mixing for otherwise by Theorem 11.7 the action of \( T \) on \( (X, \mathcal{B}, \mu, T) \) is SZ, contradicting the maximality of \( (X, \mathcal{A}, \mu, T) \). Therefore, by Theorem 13.1 there exists some system \( (Z, \mathcal{E}, \theta, R) \) defined on a Borel Probability space; an extension \( \beta : (X, \mathcal{B}, \mu, T) \to (Z, \mathcal{E}, \theta, R) \) and a non-trivial compact extension \( \gamma : (Z, \mathcal{E}, \theta, R) \to (Y, \mathcal{D}, \nu, S) \). By Theorem 12.1 the action of \( R \) on \( (Z, \mathcal{E}, \theta, R) \) is SZ. But this system corresponds to a factor \( (X, \mathcal{B}^*, \mu, T) \) where \( \mathcal{B}^* = \beta^{-1}(\mathcal{E}) \) and \( \mathcal{A} \subseteq \mathcal{B}^* \) is a proper factor. As the action of \( T \) on \( (X, \mathcal{B}^*, \mu, T) \) is SZ, this too contradicts maximality. Hence the factor \( (X, \mathcal{A}, \mu, T) \) cannot be proper and thus the action of \( T \) on \( (X, \mathcal{B}, \mu, T) \) is SZ as required. \( \square \)

The proof of Theorem 13.1 runs similarly to that of Theorem 8.1, albeit with substantial modifications. We will assume that the extension is not weak mixing and show there exists a non-trivial intermediate compact extension. To do this, using the correspondence between extensions and factors, it is enough to demonstrate the existence of a \( T \)-invariant \( \sigma \)-algebra \( \mathcal{B}^* \) with the following properties:

- It is an intermediate factor: \( \alpha^{-1}(\mathcal{D}) \subset \mathcal{B}^* \subset \mathcal{B} \)
- The AP functions in \( L^2(X, \mathcal{B}^*, \mu) \) are dense in \( L^2(X, \mathcal{B}, \mu) \)

**Proof (of Theorem 13.1).** The argument here is adapted from ([5] Chapter 7; [8] §10). We divide the proof into three claims. The first of these claims mirrors the approach used in proving Theorem 8.1 where we considered a convolution operator.

Assume that the extension \( \alpha : (X, \mathcal{B}, \mu, T) \to (Y, \mathcal{D}, \nu, S) \) is not weak mixing. Then \( X \times_Y X \) is not ergodic and so there exists a bounded, non-constant \( T \)-invariant function \( H \in L^\infty(X \times_Y X) \). Define the convolution operator \( H * : L^2(X, \mathcal{B}, \mu) \to L^2(X, \mathcal{B}, \mu) \) by

\[
H * \phi (x) = \int H(x, x') \phi(x') \, d\mu_{\alpha(x)}(x') \quad \text{for all } \phi \in L^2(X, \mathcal{B}, \mu)
\]  

(13.1)
Claim (1). Let $\phi \in L^2(X, \mathcal{B}, \mu)$ and $\epsilon > 0$ be given. Then for almost every $y \in Y$ there exist functions $g_1^{(y)}, \ldots, g_m^{(y)} \in L^2(X, \mathcal{B}, \mu_y)$ such that

$$\min_{1 \leq j \leq m} \left\| T^n H \ast \phi - g_j \right\|_{L^2(\mu_y)} < \epsilon \quad \text{for all } n \in \mathbb{N}_0$$

Remark 13. Note that Claim 1 does not show each $H \ast \phi$ is AP rel $Y$. For this we would need uniform boundedness: that given $\epsilon > 0$ there exists a single set of functions $g_1, \ldots, g_m \in L^2(X, \mathcal{B}, \mu)$ such that

$$\min_{1 \leq j \leq m} \left\| T^n H \ast \phi - g_j \right\|_{L^2(\mu_y)} < \epsilon \quad \text{almost everywhere } y \in Y \text{ for all } n \in \mathbb{N}_0 \quad (13.2)$$

Proof (of Claim 1). For almost every $y \in Y$ the operator $H^* : L^2(X, \mathcal{B}, \mu_y) \to L^2(X, \mathcal{B}, \mu_y)$ is compact. Also, we have the identity

$$T^n(H \ast \phi) = H \ast T^n \phi \quad \text{for all } \phi \in L^2(X, \mathcal{B}, \mu) \text{ and all } n \in \mathbb{N} \quad (13.3)$$

This is easily deduced using the fact $T$ is measure preserving and $H$ is $T \times T$-invariant:

$$T^n(H \ast \phi)(x) = \int H(T^n x, \phi(x')) d\mu_0(T^n x)(x') = \int H(T^n x, \phi(x')) d\mu_0(x)(x') = \int H(T^n x, \phi(x')) d\mu_0(x)(x') = H \ast (T^n \phi)(x)$$

Hence $\text{clos}\{T^n H \ast \phi : n \in \mathbb{N}\}$ is a compact subset in $L^2(X, \mathcal{B}, \mu_y)$ for almost every $y \in Y$ and Claim 1 follows directly.

Although we have not demonstrated that the $H \ast \phi$ are AP rel $Y$ functions\(^{18}\), the next claim shows us each $H \ast \phi$ is arbitrarily close to an AP rel $Y$ function in terms of $L^2(\mathcal{B}, \mu)$-norm.

Claim (2). For any $\phi \in L^\infty(X, \mathcal{B}, \mu)$, given $\epsilon > 0$ there exists a function $f \in L^2(X, \mathcal{B}, \mu)$ such that $f$ is AP rel $Y$ and $\left\| f - H \ast \phi \right\|_{L^2(\mu)} < \epsilon$.

Proof. Let $\phi \in L^\infty(X, \mathcal{B}, \mu)$ be some fixed function. The case $\phi \equiv 0$ is trivial, so we assume $\|\phi\|_\infty \neq 0$.

We will construct a function $f$ equal to $H \ast \phi$ on all of $X$ except for a set where $H \ast \phi$ is ‘badly behaved’ in a sense presently described.

Fix some $y \in Y$ so that the conditional measure $\mu_y$ is defined and some $\delta > 0$. The family of $\delta$-balls $\{B_\delta(n, y)\}_{n \in \mathbb{Z}}$ where

$$B_\delta(n, y) = \{h \in \text{clos}\{T^m f : m \in \mathbb{Z}\} : \|h - T^n H \ast \phi\|_{L^2(\mu_y)} < \delta\}$$

cover the compact subset $\text{clos}\{T^m f : n \in \mathbb{Z}\} \subseteq L^2(X, \mathcal{B}, \mu)$. By taking a finite refinement that there exists some $M \in \mathbb{N}$ such that $\{T^n H \ast \phi : |n| \leq M\}$ is $\delta$-dense in $\{T^m f : m \in \mathbb{Z}\}$

Now let $m_\delta : Y \to \mathbb{N}$ be the almost everywhere defined function which takes $y \in Y$ to the least positive integer such that $\{T^n H \ast \phi : |n| \leq m_\delta(y)\}$ is $\delta$-dense in $\{T^n f : n \in \mathbb{Z}\}$. This is a measurable function. To see this, note that it suffices to show for each $M \in \mathbb{N}$ the set

$$m^{-1}(\{0, \ldots, M\}) = \{y \in Y : m(y) \leq M\}$$

\(^{18}\)It is possible to show that the $H \ast \phi$ are indeed AP rel $Y$ (see [5] Theorem 7.21). However, here we follow a different method and show the functions $H \ast \phi$ can be slightly modified to produce function $f$ easily seen to be AP rel $Y$. 

67
Let $\eta > 0$. Therefore, given any $F \in L^2(X,\mathcal{B},\mu)$ the function $y \mapsto \int F \, d\mu_y$ is measurable by Theorem 9.4. It follows that the function $\Phi : Y \to [0, \infty)$ given by

$$\Phi : y \mapsto \|T^m(H * \phi) - T^n(H * \phi)\|_{L^2(\mu_y)}$$

is measurable. Clearly the set defined by the condition in (13.4) is precisely the set $\Phi^{-1}([0, \delta))$ and so $m^{-1}([\delta, \infty))$ is measurable as required.

Moreover, for each $k \in \mathbb{N}$ the set $A_k = m^{-1}([k, \infty)) \subseteq Y$ is measurable. It is clear that the sequence of sets $(A_k)_{k \in \mathbb{N}}$ is decreasing in measure and that

$$\nu(A_k) \to 0 \quad \text{as} \quad k \to \infty$$

Therefore, given $\eta > 0$ there exists $M(\delta, \eta) \in \mathbb{N}$ such that $\nu(A_n) < \eta$ for all $n \geq M(\delta, \epsilon)$. Let $E(\delta, \eta) = A_{M(\delta, \epsilon)}$. Then $\nu(E(\delta, \eta)) < \eta$ and for almost every $y \in Y \setminus E(\delta, \eta)$ one has $m_y(y) < M(\delta, \eta)$.

Let $\epsilon > 0$ be given. Define

$$\epsilon_n = (2^n\|H\|_\infty\|\phi\|_\infty)^{-1} \epsilon$$

so that $\epsilon = \|H\|_\infty\|\phi\|_\infty \sum_{n=1}^\infty \epsilon_n$. Also let $\delta_n = \frac{1}{n}$ for all $n \in \mathbb{N}$.

Repeat the above argument for each $n \in \mathbb{N}$, taking $\delta = \delta_n$ and $\eta = \epsilon_n$ to obtain a sequence of sets $E(n) \subseteq Y$ of measure $\nu(E(n)) < \epsilon_n$ and integers $M(n) \in \mathbb{N}$ such that for $y \in Y \setminus E(n)$ the set $\{T^m H * \phi : |m| \leq M(n)\}$ is $\delta_n$-dense in $L^2(\mu_y)$.

Let

$$f(x) = \begin{cases} 0 & x \in \alpha^{-1} \left( \bigcup_{n \in \mathbb{N}} E(n) \right) \\ H * \phi(x) & \text{otherwise} \end{cases}$$

Then $\|f - H * \phi\|_{L^2(\mu)} < \|H\|_\infty\|\phi\|_\infty \sum_{n=1}^\infty \epsilon_n = \epsilon$.

It remains to show $f$ is AP rel $Y$. But this is obvious from the construction: for given $\delta > 0$ there exists $n \in \mathbb{N}$ such that $\delta_n < \delta$. Then for almost every $y \in Y \setminus E(n)$ we have $\{T^m H * \phi : |m| \leq M(\delta_n, \epsilon_n)\}$ is $\delta$-dense in $\{T^m H * \phi : m \in \mathbb{Z}\}$ in the $L^2(\mu_y)$-norm. Thus

$$\{0\} \cup \{T^m H * \phi : |m| \leq M(\delta_n, \epsilon_n)\}$$

is $\delta$-dense in $\{T^m f : m \in \mathbb{Z}\}$ in the $L^2(\mu_y)$-norm for almost every $y \in Y$.

\[\text{Claim (3). Let } \mathcal{F} \text{ be the algebra generated by the set}\]

$$\mathcal{H} = \{H * \phi : H \in L^\infty(\mu \times Y \mu), H \text{ is } T \times T \text{-invariant}; \phi \in L^\infty(X)\}$$

Denote $\mathcal{B}^* = \sigma(\mathcal{F})$. Then the following holds:

1. $\mathcal{B}^*$ is $T$-invariant and hence defines a factor $(X, \mathcal{B}^*, \mu, T)$.
2. $\mathcal{B}^*$ is an intermediate factor: $\alpha^{-1}(\mathcal{F}) \subseteq \mathcal{B}^* \subseteq \mathcal{B}$
3. The AP functions in $L^2(X, \mathcal{B}^*, \mu)$ are dense in $L^2(X, \mathcal{B}^*, \mu)$

Proof of Claim 3. 1. The fact $\mathcal{B}^*$ is $T$-invariant follows immediately by the identity

$$T(H * \phi) = H * T \phi$$

from (13.3), which holds for all $H * \phi \in \mathcal{H}$.
Lemma 13.2. Let \( 1 : = 1_{X \times X} \in L^\infty(\mu \times \mu) \) is \( T \times T \)-invariant. For any \( D \in \mathcal{D} \) the function \( 1_{\alpha^{-1}(D)} = 1_{\alpha^{-1}(D)} \in L^\infty(X, \mathcal{B}, \mu) \). Thus the convolution of \( 1 \) and \( 1_{\alpha^{-1}(D)} \) is a \( \mathcal{B}^* \)-measurable and is given by

\[
1 * 1_{\alpha^{-1}(D)}(x) = \int 1_{\alpha^{-1}(D)}(x') d\mu_{\alpha(x)}(x')
\]

\[
= E(1_{\alpha^{-1}(D)} | Y)(x)\alpha
\]

\[
= 1_{\alpha^{-1}(D)}(x) = 1_{\alpha^{-1}(D)}(x)
\]

Therefore for every set \( D \in \mathcal{D} \) the function \( 1_{\alpha^{-1}(D)} \) is \( \mathcal{B}^* \)-measurable and subsequently \( \alpha^{-1}(\mathcal{D}) \subseteq \mathcal{B}^* \).

3. By Claim 2, given any \( H * \phi \in \mathcal{H} \) and \( \epsilon > 0 \) there exists an AP rel \( Y \) function \( f \) such that \( \|f - H * \phi\|_{L^2(\mu)} < \epsilon \).

The sets \( E(\delta, \eta) \) constructed in the proof of Claim 2 are \( \mathcal{D} \)-measurable. Considering the form of the function \( f \) as given in (13.5), it follows that \( f \in L^2(X, \mathcal{B}^*, \mu) \). Thus

\[
\mathcal{H} \subseteq \overline{\{ f \in L^2(X, \mathcal{B}^*, \mu) : f \text{ is AP rel } Y \}} \quad (13.6)
\]

The set on the right hand size of (13.6) is an algebra and it follows

\[
\mathcal{F} \subseteq \overline{\{ f \in L^2(X, \mathcal{B}^*, \mu) : f \text{ is AP rel } Y \}} \quad (13.7)
\]

It suffices to show that the set \( \mathcal{F} \) is dense in \( L^2(X, \mathcal{B}^*, \mu) \). This follows from the following lemma, the proof of which is adapted from [5].

\[\square\]

\[\square\]

**Lemma 13.2.** Let \( (X, \mathcal{B}, \mu) \) be a probability space. Suppose \( \mathcal{F} \subseteq L^\infty(X, \mathcal{B}, \mu) \) is an algebra and \( 1_X \in \mathcal{F} \). Then \( \mathcal{F} \) is dense in \( L^2(X, \mathcal{B}^*, \mu) \) where \( \mathcal{B}^* = \sigma(\mathcal{F}) \) is the \( \sigma \)-algebra generated by \( \mathcal{F} \).

**Proof.** It is enough to prove that for all \( B \in \mathcal{B}^* \) and any \( \epsilon > 0 \) there exists some \( g \in \mathcal{F} \) such that

\[
\|1_B - g\|_{L^2(\mu)} < \epsilon
\]

In other words, if we denote

\[
\mathcal{C} = \{ B \in \mathcal{B}^* : 1_B \in \overline{\mathcal{F}} \}
\]

then it suffices to show \( \mathcal{B}^* = \mathcal{C} \). Once this is established, by considering linear combinations and using the fact \( \mathcal{F} \) is an algebra containing the constants, the space of all simple functions in \( L^2(X, \mathcal{B}^*, \mu) \) must also lie within the \( L^2 \)-closure of \( \mathcal{F} \). By the density of the simple functions we conclude \( \mathcal{F} \) is dense in \( L^2(X, \mathcal{B}^*, \mu) \).

Sets of the form \( f^{-1}([a, b]) \) for some \( f \in \mathcal{F} \) and \( [a, b] \subset \mathbb{R} \) generate \( \mathcal{B} \). Using the Stone-Weierstrass Theorem (see [2]) we show these generators are contained in \( \mathcal{C} \). Indeed, for any function \( f \in \mathcal{F} \) given any compact interval \([a, b] \subset \mathbb{R}\) by the Stone-Weierstrass Theorem the characteristic function \( 1_{[a, b]} \) can be approximated by polynomials on the interval \([-\|f\|_\infty, \|f\|_\infty]\). In particular, for every \( \epsilon > 0 \) there exists some \( p \in \mathbb{R}[x] \) such that

\[
\|1_{[a, b]} - p\|_{L^2(f^* \mu)} < \epsilon \quad (13.8)
\]

where \( f^* \) is the pull-back Borel measure on \( \mathbb{R} \) induced by \( f \), that is

\[
f^* \mu(B) = \mu(f^{-1}(B)) \quad \text{for all Borel sets } B \subseteq \mathbb{R}
\]

69
It is clear the expression in (13.8) is equivalent to
\[ \|1_{f^{-1}([a,b])} - p(f)\|_{L^2(\mu)} < \epsilon \]
As \( \mathcal{F} \) is an algebra, \( p(f) \in \mathcal{F} \) and hence \( f^{-1}([a,b]) \in \mathcal{C} \). It remains to show \( \mathcal{C} \) is a \( \sigma \)-algebra.

Clearly \( X \in \mathcal{C} \). For any \( f \in \mathcal{F} \) we have \( 1_X - f \in \mathcal{F} \) and it follows \( \mathcal{C} \) is closed under complementation. Suppose \( A, B \in \mathcal{F} \) and let \( \epsilon > 0 \) be given. Then there exist \( f_1, f_2 \in \mathcal{F} \) such that
\[ \|1_A - f_1\|_{L^2(\mu)} < \frac{\epsilon}{2} \quad \text{and} \quad \|1_B - f_2\|_{L^2(\mu)} < \frac{\epsilon}{2\|f_1\|_{\infty}} \]
Hence
\[ \|1_A 1_B - f_1 f_2\|_{L^2(\mu)} \leq \|1_B\|_{\infty} \|1_A - f_1\|_{L^2(\mu)} + \|f_1\|_{\infty} \|1_B - f_2\|_{L^2(\mu)} < \epsilon \]
so \( A \cap B \in \mathcal{C} \). Moreover, by considering \( 1_{A \cup B} = 1_A + 1_B - 1_A 1_B \) we have \( A \cup B \in \mathcal{C} \). Thus, given any sequence \( (A_n)_{n \in \mathbb{N}} \subset \mathcal{C} \) and \( \epsilon > 0 \) it is easy to deduce \( \bigcup_{n=1}^{\infty} A_n \in \mathcal{C} \) by noting that the function \( 1_{\bigcup_{n=1}^{\infty} A_n} \) can be approximated in \( L^2(X, \mathcal{B}, \mu) \) by functions of the form \( 1_{\bigcup_{n=1}^{N} A_n} \).

This concludes the proof of the Multiple Recurrence and therefore Szemerédi’s Theorem.

**References**

Ergodic Ramsey Theory: where Combinatorics and Number Theory meet Dynamics; Viewed 17-01-2011


