CHAPTER 1

Introduction

1. Characteristic factors

A basic problem in dynamics is to understand when two systems are the same. In ergodic theory, this is captured by the notion of isomorphism, and two measure preserving systems are indistinguishable when their measurable and dynamical structures are identified via a measure preserving transformation. More precisely, if $(X, \mathcal{X}, \mu, T)$ and $(Y, \mathcal{Y}, \nu, S)$ are measure preserving systems, they are *isomorphic* if there is a measurable isomorphism $\phi\colon (X, \mathcal{X}, \mu) \to (Y, \mathcal{Y}, \nu)$ between the measure spaces additionally satisfying $\phi \circ T = S \circ \phi$.

A related notion is when a dynamical system is a subsystem of another system, and in ergodic theory, this is made precise by the notion of a measurable factor map: this is a homomorphism preserving the measurable and dynamical structures, but not necessarily in an invertible way. More precisely, a *factor map* is a measurable map $\pi\colon (X, \mathcal{X}, \mu) \to (Y, \mathcal{Y}, \nu)$, and again this map preserves the dynamics, meaning that $\pi \circ T = S \circ \pi$. A factor map $\pi\colon (X, \mathcal{X}, \mu, T) \to (Y, \mathcal{Y}, \nu, S)$ gives rise to a *factor* of the system $(X, \mathcal{X}, \mu)$ by pulling back the σ-algebra $\mathcal{Y}$ of ν-measurable sets to a T-invariant sub-σ-algebra of $\mathcal{X}$; for simplicity, this σ-algebra is often denoted by $\mathcal{Y}$, instead of the formally more correct notation $\pi^{-1}(\mathcal{Y})$, and we refer to $\mathcal{Y}$ as a factor of the system $(X, \mathcal{X}, \mu, T)$.

Given a factor $(Y, \mathcal{Y}, \nu, S)$ of the system $(X, \mathcal{X}, \mu, T)$, there is a naturally associated subspace of $L^2(\mu)$, consisting of functions that are measurable with respect to the σ-algebra $\mathcal{Y}$ defining the factor. Using the natural identification induced by the factor map, this T-invariant subspace of functions is identified with $L^2(\nu)$. The *conditional expectation* $\mathbb{E}_\mu(\cdot \mid \mathcal{Y})\colon L^2(\mu) \to L^2(\nu)$ is the operator defined by taking orthogonal projection onto the subspace induced by the factor. A key property is that the conditional expectation map gives rise to an orthogonal decomposition of $L^2(\mu)$, providing a tool for analyzing the behavior of functions in $L^2(\mu)$.

If $(X, \mathcal{X}, \mu, T)$ is a measure preserving system and $f \in L^2(\mu)$, taking the conditional expectation of f with respect to a factor $\mathcal{Y}$ gives techniques to detect dynamical properties of the system via the decomposition

$$f = \mathbb{E}_\mu(f \mid \mathcal{Y}) + \big(f - \mathbb{E}_\mu(f \mid \mathcal{Y})\big).$$

Considering different factors of the system, we obtain different orthogonal decompositions, giving us tools for viewing the components of a function in various ways. Sometimes in studying a particular problem, the first term of the decomposition carries all of the information and the second term plays no role, and then we say that $\mathcal{Y}$ is a *characteristic factor* for the given problem. Such characteristic factors capture certain information about the system, and the choice of which factor is appropriate depends on the type of question. However, without context, the notion

of a characteristic factor is not mathematically precise, as the meaning depends on the problem being studied. We give some simple examples that illustrate this idea.

In a measure preserving system $(X, \mathcal{X}, \mu, T)$, the simplest nontrivial factor is the T-invariant sub-σ-algebra of $\mathcal{X}$, which is the smallest sub-σ-algebra of $\mathcal{X}$ guaranteeing that all T-invariant functions in $L^2(\mu)$ are measurable. This T-invariant sub-σ-algebra of $L^2(\mu)$ is used to give one of the usual proofs of von Neumann's Mean Ergodic Theorem [**206**]. A function in $L^2(\mu)$ can be decomposed into the orthogonal sum of a function measurable with respect to the T-invariant sub-σ-algebra (obtained by conditional expectation onto the T-invariant sub-σ-algebra of $\mathcal{X}$) and a function in the orthogonal complement. The analysis of the averages for von Neumann's Theorem is then simplified: if a function $f \in L^2(\mu)$ is measurable with respect to the T-invariant sub-σ-algebra, then the average of this function converges to a constant, while if a function lies in the orthogonal complement, then it can be approximated in norm as a coboundary $g - g \circ T$ for some $g \in L^2(\mu)$ and the average of such a function converges to zero. The more interesting part of the proof lies in the decomposition itself, rather than in the computation of the averages.

A more interesting factor of $(X, \mathcal{X}, \mu, T)$ is the *Kronecker factor*, which is the smallest σ-algebra $\mathcal{Z}$ of $\mathcal{X}$ such that all of the eigenfunctions of the unitary operator on $L^2(\mu)$ defined by $f \mapsto f \circ T$ are measurable. While it may not be apparent from this description, this factor is closely related to the $(T \times T)$-invariant sub-σ-algebra of the Cartesian square of the system $(X, \mathcal{X}, \mu, T)$: specifically, the invariant sub-σ-algebra of this Cartesian square is measurable with respect to $\mathcal{Z} \times \mathcal{Z}$. Furthermore, in an ergodic system, the Kronecker factor has hidden algebraic structure, and in a sense that can be made precise, it is the largest abelian rotational factor of a system. As for the invariant sub-σ-algebra, this gives rise to a decomposition of any function in $L^2(\mu)$ into a function measurable with respect to the Kronecker factor and a function lying in the orthogonal complement. Such a decomposition was essentially known to Koopman and von Neumann [**137**], who used it to detect if the system is weakly mixing: a measure preserving system is weakly mixing if and only if it has no nontrivial Kronecker factor.

The Kronecker factor is characteristic in another sense, relating to the double average

$$\frac{1}{N} \sum_{n=0}^{N-1} T^n f \cdot T^{2n} f,$$

where $(X, \mathcal{X}, \mu, T)$ is a measure preserving system and $f \in L^\infty(\mu)$. Furstenberg [**86**] showed that these averages converge in $L^2(\mu)$, and his proof relied on the decomposition of $L^2(\mu)$ into the Kronecker factor and its orthogonal complement. In particular, he showed that the Kronecker factor is characteristic for the double average, meaning that instead of computing the average in an arbitrary system, it suffices to compute this average when the system is a compact abelian group endowed with a rotation. This vastly simplifies the analysis, as one now has the tools of Fourier analysis available.

Analogous notions can be defined for a continuous transformation acting on a compact metric space: if (X, T) and (Y, S) are topological dynamical systems, then (Y, S) is a *factor* of (X, T) if there exists a continuous surjection $\pi \colon X \to Y$ such that $\pi \circ T = S \circ \pi$. Again, particular factors can be used to detect dynamical properties of the systems. An early such structural theorem is Furstenberg's

characterization of topological distal flows in [**84**]. However, there is no analog of conditional expectation, and this makes the topological factors less adaptable for our purposes.

2. Towers of factors

The invariant and Kronecker factors are the first two factors in a tower of factors that we construct, and as such are the first steps in understanding hidden behaviors and structures in an arbitrary system. Infinite towers of factors have long been a key tool for analyzing the structure of systems. In the topological setting, Furstenberg [**84**] showed that any distal system can be decomposed as a tower of factors, each of which is a simple extension of the previous one. The algebraic characterization that he gave for these factors gives insight into the general structure of topological dynamical systems, and is useful in numerous applications. In the ergodic setting, analogs of this result were given by Furstenberg [**86**] and Zimmer [**219**], who showed that any measure preserving system can be obtained as tower of extensions. First one builds an infinite tower of certain compact structures starting with the trivial factor, capturing along the way the invariant σ-algebra, the Kronecker factor, and others, uncovering structures that describe the (measurable) distal factor of the system, and then a further extension gives the full system (capturing the weakly mixing behavior).

Furstenberg [**86**] used this description of an arbitrary measure preserving system in an essential way in his proof of Szemerédi's [**191**] famous theorem stating that a subset of integers with positive upper density contains arbitrarily long arithmetic progressions. Szemerédi's original proof was combinatorial in nature, relying on an intricate counting argument. Furstenberg's dynamical approach was fundamentally different. First he translated the combinatorial problem of finding patterns in large sets of integers (meaning those with positive upper density) into a purely measure theoretic statement about multiple averages, using what is now referred to as the Furstenberg Correspondence Principle. The second step was to show that the limit superior of a resulting multiple ergodic average is positive (in fact he proved a stronger result, showing that the limit inferior of this average is positive). For example, to prove Szemerédi's Theorem for an arithmetic progression of length k, the associated average that is studied is

$$\frac{1}{N}\sum_{n=0}^{N-1} f(T^n x)\cdot f(T^{2n}x)\cdot \ldots \cdot f(T^{(k-1)n}x),$$

where T is a measure preserving transformation on a probability measure space $(X,\mathcal{X},\mu)$ and f is a bounded function. The analysis does not require existence of the limit, but rather only needs that the limit inferior of the integral of this average is positive when f is the indicator function of a set with positive measure; this result is known as Furstenberg's Multiple Recurrence Theorem.

The use of ergodic theory to prove a combinatorial statement, in particular such a Ramsey theoretic statement, turned out to be the first of many deep advances using these techniques. The method is sufficiently robust to show the existence of numerous other patterns in sets of integers with positive upper density. This led to vast generalizations of Szemerédi's Theorem, including the existence of polynomial patterns in sets with positive upper density, along with multidimensional versions of all of these results (see for example [**19**, **89**]). However, these methods shed

no insight on what features of the measure space control the associated multiple ergodic averages and in what sense, if any, these averages converge.

Although Furstenberg's paper [**86**] included a proof of the mean convergence for the double average, for the multiple average with $k \geq 3$ terms, Furstenberg's Theorem is a recurrence statement: it suffices to deduce positivity of the averages, but it does not give sufficient information on the limiting behavior to prove convergence of the averages or to describe the resulting limit. While the Kronecker factor is characteristic for the double average, it is not characteristic for these more general averages. This general convergence problem was one of the motivating questions that led to the theory we develop.

3. Cubes, norms, nilfactors, and structure theorems

Our main focus is the description of structures in an arbitrary measure preserving system, with the goal of decomposing an arbitrary function into a sum of a function measurable with respect to such a structure, a function that is orthogonal in some sense, and possibly a small error function. The structures are described by algebraically defined systems that are a class that includes abelian rotations, along with their natural generalizations. Such systems are *nilsystems*: these are inductively constructing starting with a trivial one point system, and ultimately capture all of the nilpotent behavior in the system. Each subsequent nilsystem is obtained by a compact abelian group extension of the previous one, and the associated measure space can be described as a quotient of a nilpotent group by a discrete cocompact subgroup, with the transformation becoming translation by a fixed element of the nilpotent group. A function measurable with respect to such a nilpotent factor, or *nilfactor*, has a certain type of rigidity, and this can be exploited to prove various recurrence and convergence results.

The study of nilsystems and their properties has a long history in ergodic theory (see for example [**6**, **83**, **171**, **172**]). Due to the essential role of nilsystems in the theory we develop, we review many of their well known properties, formulating the results in ways that are convenient for our viewpoint.

One of the central problems we face is to show how nilsystems arise naturally. They are unearthed via parallelepiped structures, which we call *cubic structures*, or *cubes* for short. In the Euclidean setting, parallelepipeds are familiar objects that are easy to define and study, but we need a more sophisticated version to capture dynamical behavior. The simplest such example is described by considering the two dimensional topological dynamical cubes. More precisely, in the four fold product X^4 of a compact metric space X endowed with a homeomorphism $T\colon X \to X$, we consider quadruples of the form $(x, T^n x, T^m x, T^{n+m} x)$ for some $n, m \in \mathbb{N}$ and point $x \in X$. Taking the closure with respect to all $x \in X$ and $n, m \in \mathbb{N}$ gives rise to the two dimensional topological cubic structure. These higher dimensional cubes and the measure theoretic versions of these notions give rise to k-dimensional cubic structures. However, even the definition of a corresponding parallelepiped structure takes significant work in the measurable setting.

The interest is that the cubes correspond to patterns that capture rotational behavior, including not just the standard rotation on a circle or higher dimensional torus, but a more general rotation on a homogeneous space associated to a nilpotent group. While the Kronecker factor can only capture linear behavior in the system, the nilpotent factors allow us to capture nonlinear objects.

Using the measurable cubic structures, we then define seminorms on measurable functions, and in turn, we use the seminorms to define some special systems on which the seminorms are actually norms. Inductively, we build these systems: for the base case of $k = 1$, the factor is the Kronecker factor, and for each $k \geq 1$, we build a *system of order k* and this is the maximal factor with the property that the associated $(k+1)^{st}$-seminorm is a norm. This leads to a tower of factors, and these factors are characteristic for many dynamical problems. In particular, they are characteristic for the behavior of numerous multiple averages (including those arising in Furstenberg's proof of Szemerédi's Theorem), for the study of multiple correlation sequences, and for some recurrence problems. This would be of little interest if these factors had no description useful for applications, but it turns out that they have a simple algebraic and geometric characterization, and this is given by the nilsystems. This is the content of the main result in the theory, and we refer to it as the Structure Theorem, with various formulations of this result given in Chapter 16. Precise definitions of the terms are developed in Parts 2 and 3 of the book, but for the moment we give a roughly stated version to convey the type of results we prove:

THEOREM. *Assume that $(X, \mathcal{X}, \mu, T)$ is an ergodic measure preserving system. For each $k \in \mathbb{N}$, there is a seminorm $\|\cdot\|_{k+1}$ on $L^\infty(\mu)$ defining an associated factor of the system and this factor is an inverse limit of k-step nilsystems.*

This theorem is the main ingredient for proving numerous multiple ergodic theorems, with the simplest being the mean convergence of the averages

$$\frac{1}{N}\sum_{n=0}^{N-1} f_1(T^n x)\cdot f_2(T^{2n}x)\cdot\ldots\cdot f_k(T^{(k-1)n}x)$$

that arise in Furstenberg's proof of Szemerédi's Theorem (again T is a measure preserving transformation on a probability measure space $(X, \mathcal{X}, \mu)$ and $f_1, f_2, \ldots, f_k \in L^\infty(\mu)$). The first progress on this general convergence problem came with additional dynamical assumptions, such as weakly mixing, on the measure preserving system. New techniques were required to show mean convergence along arbitrarily long arithmetic progressions, convergence along more general exponents such as polynomials, convergence in the more general setting of a multiple average for commuting transformations with polynomial exponents, and convergence in the most general setting of a group of nilpotent transformations with polynomial exponents. For a measure preserving system with a single transformation T, there were numerous partial results, including the work of Conze and Lesigne [**48**–**50**] for the average of an arithmetic progression with three terms, work of Furstenberg and Weiss [**91**] for a polynomial average with two terms, as well as other partial convergence results [**117**, **118**]. Mean convergence for arithmetic progressions was proven in Host and Kra [**120**], and this is the proof that motivated the new structural description of an arbitrary measure preserving system.

The study of mean convergence for more general averages culminated in the breakthrough theorem of Walsh [**207**], proving convergence for the multiple ergodic averages where the transformations are taken from a nilpotent group of transformations and the exponents are arbitrary integer valued polynomials. This vast generalization includes the well studied cases of the polynomial averages, and in particular the averages along arithmetic progressions. Walsh's innovative proof is

not ergodic in nature, but rather following previous work of Tao [**194**], he reformulated the problem as a quantitative convergence property for bounded sequences. He then proved this statement by showing that such sequences can be decomposed into a structured piece, a piece that averages to zero, and a small error term. While this new method suffices to prove numerous new convergence results, it lacks the structural description of the averages, the associated description of the limit, and the combinatorial theorems that can be derived from this structure. We make use of Walsh's Theorem when convenient, but omit the proof as it lies outside the central themes of this book.

Instead, we focus on the structural results, and such results are not yet available in the general setting of nilpotent groups of transformations, nor even just for commuting transformations. The structural results allow us to transfer a problem in a general system to a problem in a nilsystem. This allows us to prove numerous convergence results, including special cases of Walsh's Theorem, especially multiple convergence results involving powers of a single transformation. Moreover, the explicit nature of the structures that control these averages allow us to deduce information on the form of the limit, which in turn give rise to further applications in other settings.

The ergodic structural results also prove to be useful in other dynamical settings, such as that of topological dynamics. Similar to the ergodic setting, we define a chain of topological nilfactors that describe features of a topological dynamical system, and this is carried out in [**127**], followed by other topological structure theorems [**52**, **130**, **187**]. In a surprising turn, these results can be used to prove pointwise results for multiple ergodic averages, at least for certain classes of systems, such as distal ones. Starting with the work of Huang, Shao, and Ye [**131**,**132**], several other pointwise convergence results [**53**–**55**] have now been proven. However, the general question of pointwise convergence for multiple ergodic averages remains a deep open problem, other than for the special case of cubic averages. At present, the strongest result is still that of Bourgain [**35**] for double averages, and even pointwise convergence for the multiple averages with three terms remains open.

4. Nilsequences in ergodic theory and in combinatorics

While the mean convergence of the averages was the original goal, in retrospect the structural description used to prove the convergence is both more interesting and more useful. This opened up the question of understanding which other averages are controlled by the same nilpotent structures, and ultimately which combinatorial and number theoretic objects can be described using analogous structures.

There are some obvious parallels between the ergodic and combinatorial settings. In giving a description of the structures in [**120**], we start by inductively building factors of an arbitrary system, and these factors are defined by functions such that an associated seminorm is zero. These seminorms turn out to be the infinite analog of the Gowers uniformity norms, introduced by Gowers [**97**] in his deeply influential and new proof of Szemerédi's Theorem. There are numerous differences between the Gowers uniformity norms and the ergodic seminorms; the former are defined on the finite space $\mathbb{Z}/N\mathbb{Z}$ and are norms, while the latter are defined on the infinite space $L^\infty(\mu)$ and are only seminorms. However, the objects

that describe the spaces defined by these seminorms turn out to give key insight in both the finite and infinite settings.

We give an example of such a result. If $(X, \mathcal{X}, \mu, T)$ is a measure preserving system, $k \geq 1$, and $f \in L^\infty(\mu)$, consider the *multiple correlation sequence*

$$\int_X f(x) \cdot f(T^n) \cdot \ldots \cdot f(T^{kn}x)\, d\mu(x).$$

For $k = 1$, Herglotz's Theorem gives the classical result that such a sequence is the Fourier transform of a positive finite measure on the torus. The decomposition of this measure into its continuous and discrete part decomposes the simple correlation $\int_X f(x) \cdot f(T^n x)\, d\mu(x)$ into a sum of two sequences, one of which tends to 0 in uniform density and the other of which is almost periodic. An almost periodic sequence can be defined as the evaluation of a continuous real valued function on an orbit in a compact abelian group and so this result fits within the framework of characteristic factors: the Kronecker factor is characteristic for the simple correlation. More generally, in [**18**] it is shown that a multicorrelation sequence can be decomposed into the sum of two sequences, one of which is small in uniform density and the other is a *nilsequence*, meaning it arises from evaluating a continuous real valued function on an orbit in a nilsystem. In other words, nilsystems are characteristic for multiple correlations.

An alternate version of the Structure Theorem gives a decomposition for functions in a measurable system. If $(X, \mathcal{X}, \mu, T)$ is an ergodic system, then any $f \in L^2(\mu)$ can be decomposed as a nilsequence, a bounded function that lies in the orthogonal complement of an associated nilfactor, and a function that has arbitrarily small norm in $L^p(\mu)$ for any $p \in [1, \infty)$. This functional form of the structural results is given in Theorem 5 of Chapter 16, and other structural results are described in the same chapter.

In a further twist, nilsequences turn out to provide an appropriate interface between ergodic theory and combinatorics, giving insight into what sorts of objects describe the Gowers uniformity norms. In particular, finite analogs of the nilsystems and structural results play a key role in understanding certain patterns in subsets of the integers that do not have positive upper density, such as the primes. A major ingredient in this program is the Inverse Theorem for Gowers norms of Green, Tao, and Ziegler [**108**], and this is the analog, in the setting of finite systems, of the functional version of the Structure Theorem.

These number theoretic results, in turn, can be used in the ergodic setting, for example showing that certain other multiple ergodic averages converge and in giving descriptions of associated ergodic structures. This has continued the two way flow between combinatorics and ergodic theory that started with Furstenberg's seminal paper [**86**], with combinatorial methods and results being adapted for use in ergodic theory, and ergodic advances and techniques being used in combinatorics, number theory, and probability (see, for example, [**87**, **115**, **140**, **141**, **143**, **193**]).

Organization of the book

This book is intended for advanced graduate students and researchers in ergodic theory, along with those working in the related areas of combinatorics and number theory. As such we only briefly cover some of the standard background material in ergodic theory and dynamical systems, focusing mainly on the point of view of the cubic structures and nilsystems in ergodic theory.

Chapters 2 to 5 contain basic material; readers with a background in ergodic theory and topological dynamics, and the connections to combinatorial results, can skip these chapters and refer back to them as needed.

The next section of the book, Chapters 6 through 8 introduces one of the main topics of the book: these are the cubes, which are described in algebraic, topological, and ergodic settings. We use the cubes in Chapter 9 to describe the systems of order k, which are the systems used to describe the basic structures in the decomposition results.

Chapters 10 through 15 cover nilmanifolds and nilsystems. The main definitions and results are in the first two of these chapters, and the more detailed analysis of Chapters 12 to 15 can be omitted at a first reading.

The heart of the subject is described in Chapter 16, including the statement of the Ergodic Structure Theorem that connects the ergodic cubic structures and nilsystems. This chapter includes some consequences and applications of the structure theory, and starting in Chapter 17, we turn to other forms of such structural results, including in topological dynamics and in combinatorics.

Chapters 18 through 20 are devoted to the proof of the Structure Theorem, and the material in these sections is not used elsewhere in the book. Applications of the theory are described in Chapters 21 through 25.

Throughout most of this book, we restrict ourselves to $\mathbb{Z}$-actions. We do this to simplify the notation, but much of the theory extends easily to $\mathbb{Z}^d$-actions (this should not be confused with the extension of structural results to d independent actions, meaning d commuting transformations, and understanding such actions remains an open problem). Additionally, much of the theory carries over to the continuous setting, such as an $\mathbb{R}$-action, and there are some simplifications of the theory in this setting.

We defer the references to the end of each chapter, where we include bibliographic references and further notes on results presented. Numbering of theorems, propositions, and other results is done chapter by chapter, and so within each chapter these results are only referred to by number, but when we refer to a result in another chapter we also include this in the reference.

Acknowledgments

We are grateful to the many friends and colleagues who encouraged us to complete this project, and who sent us numerous suggestions and corrections. In particular, we thank Nikos Frantzikinakis, Song Shao, Terence Tao, and Jean-Paul Thouvenot. The first author thanks the Center for Mathematical Modeling (CMM), University of Chile for its hospitality and support.