Entropy and Sinai's Theorem

Tim D. Austin

These short notes cover the basic definitions and properties of entropy, with an emphasis on its use in dynamical systems. They are intended to show how the different notions of entropy that appear in probability, mathematical statistical physics can be modified for use in the study of measurable dynamical systems, culminating in one of the most striking such uses of entropy: the proof of Sinai's Theorem that entropic constraints are enough to find all Bernoulli factors of an ergodic invertible probability measure-preserving system.

Our presentation loosely follows that in Smorodinsky's monograph [7].

Contents

1 Orientation

Our concern will be with the definition and use of entropy in the study of measurable dynamical systems (what has traditionally come to be called ergodic theory, even when the underlying system need not be ergodic). Ergodic theory breaks into several sub-areas:

1. The classical ergodic theorems about convergence of averages;

2. Spectral ergodic theory, traditionally concerned with establishing relationships between the properties of a measure-preserving self-map τ of a measure space $(\Omega, \Sigma, \mu, \tau)$ and the associated isometric operator acting on the Hilbert space $L^2(\mu)$ (over $\mathbb{R}$ or $\mathbb{C}$, according to taste);
3. Structural ergodic theory, concerned with finding different properties and invariants of measure-preserving systems that can (in principle) distinguish them, or possibly allow us to relate more abstract examples, about which we know only a little information, to more concrete ones;
4. Ergodic theory in some more specialized category – such as for autohomeomorphisms of topological spaces or, more popularly, for various special kinds of diffeomorphism of differentiable manifolds;
5. Ergodic Ramsey Theory, pioneered by Furstenberg and Katznelson, which provides a clean and powerful language for stating and proving various results in density-Ramsey Theory (the version of Ramsey Theory concerned with substructures that are 'large' in some quantitative sense, rather than with those defined by colourings).

Each of these can be extended in various directions – even the first, and oldest, can be made to fill a book by including results about averages of functions from different function spaces, about different notions of convergence, about the actions of different amenable groups and about ergodic averages taken from special subsequences. However, perhaps only the fourth and fifth are still the scene of much new research, and only the fourth (which, naturally, can be viewed as being subsumed into the theory of whichever other category is being studied) is still a major industry.

In this essay we will be concerned with the centrepiece of structural ergodic theory: the theory of entropy, introduced to ergodic theory by Kolmogorov in 1958 [3], motivated in part by the work of Shannon in information theory during the 1940's and 50's (see, for example, [6]). There have also developed various additional methods in the structure theory of measure-preserving transformations, concerned mainly with finding systematic ways to construct examples of particular phenomena; however, only the entropy theory stands out as having yielded results that apply to very general classes of system.

More particularly, the ultimate goal of this essay is the present one of the culminating results of the entropy theory: Sinai's 1964 result describing all of the factors of a measure-preserving system that are of a particularly simple form – a Bernoulli shift – purely in terms of entropic constraints. This presentation serves a twofold purpose: Sinai's Theorem is worth learning in its own right, and the journey to reach it serves as a very thorough introduction to dynamical system entropy. In fact, we shall see that the whole theory of dynamical system entropy can be conceived quite naturally as a systematic means for comparing an arbitrary system with Bernoulli shifts.

2 Entropy in measure theory and probability

In this section we review various ideas and constructions from probability. Proofs are omitted; they are all routine, or can be found in standard texts that cover entropy, such as Ruelle's monograph [5]. (Probably it will be clearer to find them in a book on probability or statistical mechanics, rather than ergodic theory, owing to slight differences in the conventional definitions; we will have more to say on this later.)

We will assume some basic ideas from measure theory and probability, including the Radon-Nikodým Theorem. We will adopt throughout the convention that all measures are semi-finite (or even σ-finite, if the reader prefers).

2.1 Relative entropy

In the first instance, entropy – in the form of relative entropy – is a tool for studying measures. Specifically, it gives us a way of comparing one measure with another.

Definition 2.1 *Suppose that* (Ω,Σ) *is a measurable space and that* μ,ν *are two measures on it. Then we define the* ***relative entropy of*** μ ***with respect to*** ν *by*

$$s(\mu:\nu) := \begin{cases} \int_\Omega f\log f\,\mathrm{d}\nu & \textit{if } \mu \ll \nu \textit{ and } \mu = \nu\llcorner f\textit{, and the integral makes sense} \\ \infty & \textit{else.} \end{cases}$$

Lemma 2.2 *If* μ *and* ν *are both* probability *measures then* $s(\mu:\nu) \geq 0$*, with equality if and only if* $\mu = \nu$*.*

Proof The function $F:[0,\infty)\to\mathbb{R}:x\mapsto x\log x$ is bounded below and strictly convex (this is an elementary exercise in calculus), and so the result follows from the conditions for equality in Jensen's inequality, since in this case $\int_\Omega f\,\mathrm{d}\nu = 1$ and $F(1) = 0$. □

So, in some sense, in the probabilistic setting entropy provides a measure of how far μ is from being equal to ν. For example, if f is always close to 1, and so in some sense μ is but a small re-weighting of ν, then $f\log f$ will be close to zero and the relative entropy is small. (More formally, the entropy is controlled by the $L^\infty(\nu)$-norm of $(f-1)$; note, however, that this is not true for the $L^1(\nu)$-norm.) On the other hand, if μ is not absolutely continuous with respect to ν at all – or even if f exists but has is too singular to lie in $L\log L(\nu)$ – then $s(\mu:\nu)$ is infinite.

It seems worth remarking right away that our function $s(\,\cdot : \cdot\,)$ is by no means unique in having the above properties. Indeed, letting $F:[0,\infty)\to(-\infty,\infty]$ be any continuous strictly-convex function satisfying $F(0) = F(1) = 1$, precisely the above reasoning leads us to the same conclusions for

$$s_F(\mu:\nu) := \begin{cases} \int_\Omega F\circ f\,\mathrm{d}\nu & \text{if } \mu \ll \nu \text{ and } \mu = \nu\llcorner f\text{, and the integral makes sense} \\ 0 & \text{else.} \end{cases}$$

(note that this is not a standard definition). Moreover, there seems no reason to believe that such s_F exhaust all possible comparators of two measures with these properties.

Yet, entropy defined using the function $F:x\mapsto x\log x$ is afforded a special status, having proved useful in a range of different kinds of argument. After defining absolute entropy in the next subsection, we will begin to see the reasons for the usefulness of these ideas. We end this subsection with one more definition and elementary result.

Definition 2.3 *Given measurable spaces* (Ω,Σ) *and* (X,T)*, a probability measure* μ *on* (Ω,Σ) *and a measurable function* $\psi:(\Omega,\Sigma)\to(X,\mathrm{T})$*, we will write* $\mu\circ\psi^{-1}$ *for the* ***push-forward measure*** *obtained from* μ *under the action of* ψ*:*

$$\mu\circ\psi^{-1}(A) := \mu(\psi^{-1}A) \quad \textit{for } A\in\mathrm{T}.$$

If we want to emphasize the function ψ *more than the measure* μ*, we may refer to the* ***law of*** ψ ***under*** μ*.*

The following property follows at once from the conditional Jensen inequality:

Proposition 2.4 (Non-increase of entropy under push-forward) *Suppose that* (Ω_i,Σ_i)*,* $i=1,2$*, are two measurable spaces and that we have two probability measures,* μ *and* ν*, on the first space* (Ω_1,Σ_1)*, and also a map* $\phi:(\Omega_1,\Sigma_2)\to(\Omega_2,\Sigma_2)$*. Then*

$$s(\mu\circ\phi^{-1}:\nu\circ\phi^{-1}) \leq s(\mu:\nu).$$

2.2 Absolute entropy

In some respects entropy seems to be most easily understood in its relative form; however, there are strong reasons – particularly in the applications to dynamical systems that we will see later – for introducing separately a certain specialization of relative entropy: absolute entropy.

We first observe that measurable maps $\phi : (\Omega, \Sigma) \to ([k], \mathcal{P}[k])$ to the finite set $[k] := \{1, 2, \ldots, k\}$ (with all its subsets regarded as measurable) are in a natural bijective correspondence with ordered measurable partitions of Ω of length k: we simply take our partition to have cells $\phi^{-1}\{0\}$, $\phi^{-1}\{1\}$, ..., $\phi^{-1}\{k-1\}$. More generally, a measurable map into any finite set can be thought of as a labelled partition. Of course, some of the corresponding cells may be empty if our map is not surjective. Henceforth we shall assume these identifications and speak of such maps as being themselves partitions (and will at times find notating a partition as a finite-valued function quite handy).

In working with partitions, we will need to call on the notion of **refinement**: one partition $\theta : \Omega \to [K]$ **refines** another $\phi : \Omega \to [k]$ if each cell of the latter is a union of cells of the former, or, equivalently, if there is some $\gamma : [K] \to [k]$ such that $\phi = \gamma \circ \theta$ (given this interpretation, we may also say that ϕ **factors through** θ). We write $\theta \Rightarrow \phi$.

We will also make use of an approximate version of the above idea: Given $\varepsilon > 0$, we will say that $\theta : \Omega \to [K]$ is a ε-**refinement** of $\phi : \Omega \to [k]$ if there is some small 'exclusion set' $\Omega_0 \subseteq \Omega$, $\mu(\Omega_0) < \varepsilon$, and some function $\gamma : [K] \to [k]$ such that $\phi(\omega) = \gamma \circ \theta(\omega)$ for $\omega \in \Omega \setminus \Omega_0$. Thus this definition is saying that images under ϕ can be recovered from images under θ with high probability. Note that this property depends on our choice of μ, although it does not appear explicitly in the notation. In this case we will write $\theta \Rightarrow^{\varepsilon}_{\gamma} \phi$ or just $\theta \Rightarrow^{\varepsilon} \phi$.

Given two partitions ϕ, ψ, they have a common refinement as *unordered* partitions obtained by considering all possible intersections of one cell from each; however, for our purposes it will be more natural to think of their refinement as the map $(\phi, \psi) : \Omega \to [K] \times [k] : \omega \mapsto (\phi(\omega), \psi(\omega))$, which amounts to endowing the unordered common refinement with labels and possibly introducing some extra empty cells. We adopt here the non-standard name of the **composite partition of** ϕ **and** ψ for this construction, and extend it to larger families of partitions in the obvious way.

Note also that on the finite measurable space $([k], \mathcal{P}[k])$, a probability measure can be identified with a stochastic vector $\mathbf{p} = (p_0, p_1, \ldots, p_{k-1})$ in the usual way; we will sometimes refer to such measures as probability **distributions**, to mark them out as special.

Definition 2.5 *Given a probability distribution* $\mathbf{p} = (p_0, p_1, \ldots, p_{k-1})$ *on the set* $[k]$*, its* ***absolute entropy****, or just* ***entropy****,* $H(\mathbf{p})$*, will refer to* minus *its relative entropy with respect to to counting measure* $\#$*: this can be written out explicitly as*

$$H(\mathbf{p}) := -s(\mathbf{p} : \#) = -\sum_{0 \leq i \leq k-1} p_i \log p_i = \log k - \frac{1}{k} s\Big(\mathbf{p} : \frac{1}{k}\#\Big)$$

(where the last expression gives the relation with the counting measure $\#$ *normalized to the probability measure* $\frac{1}{k}\#$*).*

Similarly, given probability space (Ω, Σ, μ) *and a partition* $\phi : \Omega \to [k]$*, the* ***absolute entropy of*** ϕ ***with respect to*** μ*,* $H_\mu(\phi)$*, will be understood to be the absolute entropy of the push-forward measure* $\mu \circ \phi^{-1}$*:*

$$H_\mu(\phi) := H(\mu \circ \phi^{-1}) = -\sum_{0 \leq i \leq k-1} \mu\{\omega \in \Omega : \ \phi(\omega) = i\} \log(\mu\{\omega \in \Omega : \ \phi(\omega) = i\}).$$

We note first that for absolute entropy there is a valuable inequality which goes some way towards reversing our earlier result Proposition 2.4 for relative entropy.

Proposition 2.6 *Suppose that* θ *and* ϕ *are partitions of* Ω *with* θ *a refinement of* ϕ*. Then* $H_\mu(\theta) \geq H_\mu(\phi)$*.*

Proof This is an elementary exercise in manipulation and convexity. □

2.3 Relative versus absolute entropy

Why have we introduced these two separate notions of entropy, and even graced each with its own notation?

The difference here is really one of our underlying emphasis: relative entropy is suitable for the study of *measures*, while absolute entropy is the better choice for the study of *maps*.

In several settings in probability theory and mathematical statistical physics, it happens that we have a naturally-occurring underlying measure space, and then a number of other measures also defined on it as a result of some random process (for example, empirical measures drawn from a Markov chain or a spatially-distributed statistical physical model). In this setting we may wish to compare these other measures with the first. In many cases, so long as the underlying probabilistic model has a enough independence built into it, relative entropy will provide a useful comparator; it may even appear as a natural parameter in results describing some other asymptotic behaviour of the system. Perhaps the most striking such applications occur, on the one hand, in the theory of large deviations, where entropy is often found to characterize (up to exponential order of magnitude) the small probability of some very rare event in the evolution of a stochastic process; and on the other, in the thermodynamic formalism of mathematical statistical physics, where a suitably normalized version of relative entropy has its own modest calculus, putting on a rigorous footing the heuristic associated with entropy in physics since Boltzmann's work on thermodynamics in the $19^{\rm th}$ century. For nice treatments of these subjects see, for example, Deuschel and Stroock [1] and Ruelle [5] respectively.

However, starting in the next section, our concern will be with the structure of measurable dynamical systems. In this field the questions are a little different. We wish to study morphisms (that is, measure-preserving maps that respect the group action) that map one such system onto another; in particular, in building a structure theory, we will typically wish to study maps either from some simple, special class of system into our general system, or vice-versa. In this study, empirical measures – for example, defined by the ergodic averages up to a particular time – do appear naturally, but as *auxiliary* objects. Our choice to work with absolute entropy now follows from our more specific choice to single out finite sets with counting measure as the building blocks of our simple, special systems, and to make a comparison of our general system with these in particular by studying possible maps between them. Note that in this case, measurable maps out of our general space into a finite set are just measurable partitions.

How we use these building blocks will become clear shortly (when we define Bernoulli shifts, our preferred non-trivial-but-simple measure-preserving systems); what should be stressed now is that, given our intention to work with partitions and finite sets, the monotonicity and continuity properties of absolute entropy, regarded as a property that characterizes a partition, will be more helpful than the related (but not immediately equivalent) properties of relative entropy, regarded as characterizing of a measure (or, typically, both an empirical measure and an underlying 'reference measure').

Before moving on to dynamical systems, we need just a little more background.

2.4 Independence and ε-independence

The definition of entropy (in either its relative or absolute form) has one additional property that both singles it out uniquely and ensures its worth:

Proposition 2.7 (Additivity of entropy: relative version) *Suppose that* (Ω_i, Σ_i)*,* $i = 1, 2$*, are two measurable spaces and that on each we have two probability measures, say* μ_i*,* ν_i *for* $i = 1, 2$*. Then, working with the usual product measures on the product space* $(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2)$*,*

$$s(\mu_1 \otimes \mu_2 : \nu_1 \otimes \nu_2) = s(\mu_1 : \nu_1) + s(\mu_2 : \nu_2).$$

□

Proposition 2.8 (Additivity of entropy: absolute version) *Suppose that $\phi_i : \Omega_i \to [k]$, $i = 1, 2$, are two independent partitions of Ω. Then*

$$H_\mu((\phi_1, \phi_2)) = H_\mu(\phi_1) + H_\mu(\phi_2).$$

□

In fact the above are an immediate consequence of a rather stronger result, described below. To state this we need the following definition:

Definition 2.9 *Suppose $(\Omega_i, \Sigma_i, \mu_i)$ are probability spaces for $i = 1, 2$. By a **joining** of μ_1 and μ_2 we understand a probability measure μ on $(\Omega_1 \times \Omega_2, \Sigma_1 \otimes \Sigma_2)$ with marginals μ_1, μ_2: that is, such that $\mu \circ \pi_i^{-1} = \mu_i$ for $i = 1, 2$.*

Example The simple product measure $\mu_1 \otimes \mu_2$ is always a joining of μ_1, μ_2. ◁

Proposition 2.10 (Superadditivity of relative entropy) *Suppose that (Ω_1, Σ_1), (Ω_2, Σ_2) are two measurable spaces and that on each we have two probability measures, say μ_i and ν_i for $i = 1, 2$. Suppose that in addition we have a joining μ of μ_1 and μ_2. Then*

$$s(\mu : \nu_1 \otimes \nu_2) \geq s(\mu_1 : \nu_1) + s(\mu_2 : \nu_2),$$

with equality if and only if $\mu = \mu_1 \otimes \mu_2$. □

Proposition 2.11 (Subadditivity of absolute entropy) *Suppose that $\phi_i : \Omega \to [k]$ are two partitions of Ω. Then*

$$H_\mu((\phi_1, \phi_2)) \leq H_\mu(\phi_1) + H_\mu(\phi_2),$$

with equality if and only if ϕ_1 and ϕ_2 are independent. □

This behaviour under tensor products is certainly clean and satisfying, but it is perhaps still obscure under what circumstances it could be useful. The point is that it allows us to make *calculations* or estimates of entropy values in a variety of situations and thus obtain good estimates on it as a comparator of two probability measures. However, it will be made more useful still once we also have ways to deduce other properties of a system from the knowledge that it satisfies certain entropy constraints.

The first steps in this direction are to prove related ε, δ-versions of our above results. Here we describe a relationship with a kind of approximate independence, and in the next section we will also prove such 'continuity' results relating entropy to approximate refinements.

Definition 2.12 *Suppose that $\phi_i : (\Omega, \Sigma) \to (X_i, \mathrm{T}_i)$ for $i = 1, 2$ are two measurable maps and that μ is a probability measure on (Ω, Σ), and consider their joint distribution $\mu \circ (\phi_1, \phi_2)^{-1}$ on $(X_1 \times X_2, \mathrm{T}_1 \otimes \mathrm{T}_2)$. Given an $\varepsilon > 0$ we will say that ϕ_1 and ϕ_2 are **ε-uncorrelated** if*

$$\|\mu \circ (\phi_1, \phi_2)^{-1} - (\mu \circ \phi_1^{-1}) \otimes (\mu \circ \phi_2^{-1})\| < \varepsilon.$$

Remark In fact we will need a slightly different formulation of approximate independence when we come to use it later, relating the law of ϕ_1 with its law conditioned on ϕ_2: if $\varepsilon > 0$, then we can choose some other $\varepsilon_1 > 0$ so that if ϕ_1 is ε_1-uncorrelated with ϕ_2, then it follows that the law of ϕ_1 conditioned on ϕ_2 taking a particular value x_2 is within ε of the unconditional law of ϕ_1, for x_2 outside some exclusion set in X_2 of measure at most ε. In this case we will say ϕ_1 and ϕ_2 are **ε-independent**, and will write $\phi_1 \perp^\varepsilon \phi_2$. A similar equivalence runs in the other direction; both follow quickly from a careful unpacking of the relevant definitions. Of course, this latter notion depends on technicalities about the existence of conditional laws in quite general spaces, but in our special case of finite partitions we will find it to be the more directly useful. ◁

Proposition 2.13 (Robust superadditivity for relative entropy) *Given $\varepsilon > 0$ and ν_1, ν_2 measures on our target spaces we can choose $\eta > 0$ depending on ε, ν_1 and ν_2 such that for any ϕ_1, ϕ_2 as above, if*

$$s(\mu \circ (\phi_1, \phi_2)^{-1} : \nu_1 \otimes \nu_2) < s(\mu \circ \phi_1^{-1} : \nu_1) + s(\mu \circ \phi_2^{-1} : \nu_2) + \eta$$

then $\phi_1 \perp^{\varepsilon} \phi_2$.

Proposition 2.14 (Robust subadditivity for absolute entropy) *Given $\varepsilon > 0$ and $\phi_1 : \Omega \to [k]$, $\phi_2 : \Omega \to [\ell]$ partitions of Ω we can choose $\eta > 0$, depending now on ε, k and ℓ, such that if*

$$H_\mu((\phi_1, \phi_2)) > H_\mu(\phi_1) + H_\mu(\phi_2) - \eta$$

then $\phi_1 \perp^{\varepsilon} \phi_2$.

Proof For either version this is another, slightly more elaborate, exercise in manipulation and the use of standard inequalities, going via the notion of approximate-uncorrelation. □

3 Entropy in dynamical systems

3.1 Motivation

Our interest in this section and the rest of this essay is in **probability measure-preserving systems**: quadruples $(\Omega, \Sigma, \mu, \tau)$ such that (Ω, Σ, μ) is a probability space and $\tau : \Omega \to \Omega$ is a measurable, measure-preserving self-map. In fact, all our results can be made to apply to the case of μ any finite measure simply by renormalizing; some of them will also extend to the case of infinite μ, although here the work would be more involved. We will not comment on these points further.

Popular examples of such systems are abundant: in particular, they arise naturally as discrete-time samples from flows in geometry and in the study of Hamiltonian dynamics, and also from other kinds of stochastic process in probability. Here we will describe some of the centrepieces of a structure theory for such systems in quite an abstract setting, without the assumption of any other underlying structure.

We will make the assumption throughout that our system is invertible; this is not strictly necessary for most of the results that follow, but will make the presentation easier to follow without suppressing any of the main ideas.

Our basic program will be one of comparing our abstract system $(\Omega, \Sigma, \mu, \tau)$ with some collection of more concrete systems that we feel we understand. In doing this we must choose whether to try to find them as factors of $(\Omega, \Sigma, \mu, \tau)$, or vice versa – it turns out that a more interesting theory can be built for the latter.

We will first specify a suitable collection of reference systems:

Definition 3.1 *Given a finite set $[k]$ and a probability distribution $\mathbf{p}$ on it, form its infinite tensor power indexed by $\mathbb{Z}$ as usual: $X = [k]^{\mathbb{Z}}$, $\mathrm{T} = (\mathcal{P}[k])^{\otimes\mathbb{Z}}$, $\nu = \mathbf{p}^{\otimes\mathbb{Z}}$. Then the* ***bilateral Bernoulli shift*** *σ on (X, T, ν) is defined by $\sigma((x_i)_{i\in\mathbb{Z}}) \mapsto (x_{i+1})_{i\in\mathbb{Z}}$.*

Remarks 1. We could also construct the analogous unilateral Bernoulli shift on the space $[k]^{\mathbb{N}}$; we will not need this here, owing to our earlier assumption of invertibility.

2. We could easily perform this construction in rather more generality, building a bilateral Bernoulli shift σ on the product space $(X_0^{\mathbb{Z}}, \mathrm{T}_0^{\otimes\mathbb{Z}}, \nu_0^{\otimes\mathbb{Z}})$ for any underlying probability space $(X_0, \mathrm{T}_0, \nu_0)$; however, although this more abstract picture might be intuitively useful, later on working with a finite underlying set will give us some very important extra traction.

In fact, the theory we will describe below *can* be extended to more general spaces, but not completely arbitrary ones. *Really*, this whole theory is about measure algebras, not measure spaces; as is, arguably, much of the rest of ergodic theory. The problem with adopting a picture painted in terms of measure spaces is one of being able to represent homomorphisms of the underlying measure algebras by actual maps between those spaces. This not possible in full generality, and the theory rests crucially on our ability to construct such homomorphisms or maps (depending on the adopted picture) with certain properties.

It is possible to develop this theory by making some weaker assumptions on our target space $(X_0, \mathrm{T}_0, \nu_0)$, to ensure that we can construct measurable functions into it: for example, one of many equivalent approaches would be to assume that it is some Polish space with its Borel σ-algebra. However, this leads to considerably trickier technical difficulties than surface while working only with finite partitions, since building suitable functions (and various other things we will need) or such more general structured spaces is still not straightforward. Here we will content ourselves with the traditional form of our structure theory, in terms of partitions, while doffing our hats to the observation that a more high-minded understanding of our work would involve thinking only about measure algebras and not whole measure spaces anyway. For an introduction to measure algebras (including a rather detailed discussion of Sinai's Theorem), see Fremlin [2]. $\triangleleft$

The observation underlying the whole of the program to follow is this: Suppose that we are given any partition $\phi : \Omega \to [k]$. Then using it and τ we can define for each $I \subseteq \mathbb{Z}$ the map

$$(\phi \circ \tau^i)_{i \in I} : \Omega \to [k]^I : \omega \mapsto (\phi(\tau^i(\omega)))_{i \in I}.$$

We use the shorthand name $\phi_I^{(\tau)}$ for this, and abbreviate further $\phi_{\mathbb{Z}}^{(\tau)}$ as $\phi^{(\tau)}$. Now the push-forward $\mu \circ (\phi^{(\tau)})^{-1}$ is a shift-invariant probability measure on $(X, \mathrm{T}) = ([k]^{\mathbb{Z}}, (\mathcal{P}[k])^{\otimes \mathbb{Z}})$.

This gives us a way of building shift-invariant measures on (X, T), and so of finding many factors of our original system $(\Omega, \Sigma, \mu, \tau)$. However, so far this is not telling us very much about abstract measure preserving systems (rather, it is telling us that shift-invariant measures on the product space (X, T) can exhibit almost any kind of behaviour). A more interesting question, however, is whether we can find some special partition $\phi : \Omega \to [k]$ for which $\mu \circ (\phi^{(\tau)})^{-1}$ is a product measure, so that the corresponding factor of $(\Omega, \Sigma, \mu, \tau)$ is a Bernoulli shift – a member of our chosen class of more concrete reference systems.

This is the point at which entropy becomes relevant: Given its properties relating to independence and approximate independence, it can be used tell us how close is the shift system associated to a partition of Ω to being independent.

It turns out that, suitably bent to our purpose, entropy provides a measure of this 'closeness' of a factor to a Bernoulli shift that is both very clean and also enables a proof that certain factors can arise – this latter existence result is contained in Sinai's Theorem.

3.2 More definitions

We first need to define dynamical system entropy and derive some of its more elementary properties.

For any finite $I \subseteq \mathbb{Z}$ write π_I for the canonical projection map $X = [k]^{\mathbb{Z}} \to [k]^I$, and endow $[k]^I$ with its product measurable space structure also. Now, having pushed μ forward to $\mu \circ (\phi^{(\tau)})$ on X, we can further consider the finite-dimensional image measures of μ under these projections: of course, these are just $\mu \circ (\phi^{(\tau)})^{-1} \circ \pi_I^{-1} = \mu \circ \phi_I^{(\tau)}$. Since μ is τ-invariant, we find that the projected measures $\mu \circ (\phi^{(\tau)})^{-1} \circ \pi_I^{-1}$ are preserved under translations $I \to I + n$. In particular, the one-dimensional distributions $\mu \circ (\phi^{(\tau)})^{-1} \circ \pi_n^{-1} = \mu \circ (\phi \circ \tau^n)^{-1}$, regarded as measures on $[k]$, all agree.

It follows that if our factor measure $\mu \circ (\phi^{(\tau)})^{-1}$ is just a Bernoulli shift, we can recover it precisely as $(\mu \circ \phi^{-1})^{\otimes \mathbb{Z}}$, since in this case the coordinate maps π_n are independent. In general, this will not be the case, but we can now introduce how, for each finite $I \subseteq \mathbb{Z}$, entropy can be used as a measure of 'shortfall' from this independence. As foreseen earlier, here we will be concerned only with absolute entropy. Given a partition $\phi : \Omega \longrightarrow [k]$, we will consider the values

$$\frac{1}{|I|} H\big(\mu \circ (\phi_I^{(\tau)})^{-1}\big).$$

The normalizing constant $\frac{1}{|I|}$ is worth remarking here: as we will see shortly, it gives us the right control over the size of these quantities as $|I|$ grows. For $|I|$ large, the un-normalized relative entropy is very large: as $|I|$ increases our two measures $\mu \circ (\phi_I^{(\tau)})^{-1}$ and $(\mu \circ \phi^{-1})^{\otimes I}$ tend towards being mutually singular. Considering the situation 'at infinity', the measures $\mu \circ (\phi^{(\tau)})^{-1}$ and $(\mu \circ \phi^{-1})^{\otimes \mathbb{Z}}$ may actually be mutually singular; but looking at finite-dimensional distributions allows us to judge the 'rate' at which they tend towards mutual singularity as $|I|$ increases.

In our definition of dynamical system entropy we use this idea restricted to the special case $I = [n]$:

Definition 3.2 (Absolute dynamical entropy) *Suppose $\phi : \Omega \longrightarrow [k]$ is a partition. The **entropy of** τ **with respect to** μ **measured by** ϕ is*

$$h_\mu(\tau, \phi) := \limsup_{n\to\infty} \frac{1}{n} H\big(\mu \circ (\phi_{[n]}^{(\tau)})^{-1}\big),$$

*and the **entropy of** τ **with respect to** μ is*

$$h_\mu(\tau) := \sup_{k\geq 1,\ \phi:\Omega\longrightarrow[k]} h_\mu(\tau, \phi).$$

Remark The second of the above quantities can take any value in $[0, \infty]$. ◁

Before going on to prove more useful properties of dynamical systems entropy, we should witness the following:

Lemma 3.3 *The limit supremum in the definition of $h_\mu(\tau, \phi)$ is actually a limit, and its value is less than or equal to any given term of the sequence.*

Proof First, observe that for $m, n \geq 1$ the partition $\phi_{[m+n]}^{(\tau)}$ is a composite of the two partitions $\phi_{[n]}^{(\tau)} : \Omega \longrightarrow [k]^n$ and $(\phi \circ \tau^n)_{[m]}^{(\tau)} : \Omega \longrightarrow [k]^m$, and this latter has the same distribution (and hence the same entropy) as $\phi_{[m]}^{(\tau)}$. Therefore, by Proposition 2.11,

$$H\big(\mu \circ (\phi_{[m+n]}^{(\tau)})^{-1}\big) \leq H\big(\mu \circ (\phi_{[n]}^{(\tau)})^{-1}\big) + H\big(\mu \circ (\phi_{[m]}^{(\tau)})^{-1}\big);$$

that is, the sequence of values

$$H\big(\mu \circ (\phi_{[n]}^{(\tau)})^{-1}\big) \quad \text{for } n = 1, 2, \ldots$$

is subadditive. It is now a standard exercise that the limit

$$\lim_{n\to\infty} \frac{1}{n} H\big(\mu \circ (\phi_{[n]}^{(\tau)})^{-1}\big)$$

exists and is less than or equal to each term of the sequence. □

3.3 Dynamical system entropy and refinement

We have now covered enough ground to prove two fundamental clusters of results, firstly on the relation of entropy to approximate refinement, and secondly on its relation to approximate independence. Our first result, proved in this subsection, is a 'direct' result, showing how approximate refinement controls dynamical system entropy. This should be contrasted with the 'inverse' result of Proposition 2.14 and its dynamical system version, Proposition 3.8 in the next section, which show how dynamical system entropy controls approximate independence.

Proposition 3.4 *Suppose that θ and ϕ are partitions of Ω with θ a refinement of ϕ. Then $h_\mu(\tau,\theta) \geq h_\mu(\tau,\phi)$*

Proof This follows at once from Proposition 2.6. □

Proposition 3.5 *For any $\varepsilon > 0$ and fixed $K \geq 1$ there is some δ such that the following holds: If $\theta : \Omega \to [K]$ is a partition and $\phi : \Omega \to [k]$ is a δ-refinement of it then*

$$h_\mu(\tau,\phi) > h_\mu(\tau,\theta) - \varepsilon.$$

Remark 'Approximate refinements catch almost all entropy'. ◁

Proof Suppose that $\phi \Rightarrow^\delta_\gamma \theta$ with exclusion set $\Omega_0 \subseteq \Omega$. Let $\psi : \Omega \to [k+1]$ be any other partition which agrees with θ on the exclusion set and takes the constant value $k+1$ outside that set. The idea here is that ψ looks to see whether we are in the exclusion set, and emulates θ if we are; otherwise, it always takes the value $k+1$. Since the exclusion set is small, ψ tells us very little everywhere except on a small part of Ω, and so we expect it to be 'entropically small'. This means that we can recover θ exactly by introducing, in addition to ϕ, only a new partition which tells us very little on most of Ω, and so should have little effect on the entropies.

To be precise, we observe that (ϕ,ψ) strictly refines θ, since

$$\theta(\omega) = \begin{cases} \gamma(\phi(\omega)) & \text{if } \psi(\omega) = k+1 \\ \psi(\omega) & \text{if } \psi(\omega) \in [k]. \end{cases}$$

Therefore, by Proposition 3.4,

$$h_\mu\big(\tau,(\phi,\psi)\big) \geq h_\mu(\tau,\theta).$$

On the other hand, we can compute explicity from the formula for absolute entropy that

$$h_\mu\big(\tau,(\phi,\psi)\big) \leq h_\mu(\tau,\phi) + h_\mu(\tau,\psi) \leq h_\mu(\tau,\phi) + (1-\delta)\log(1-\delta) - \delta\log\delta + \delta\log k.$$

Now for a given k, by choosing δ sufficiently small we can make the error term on the right hand side above as small as we please; by forcing it to be less than ε we prove the result. □

Another result in a slightly different direction, but which we will need to combine with the above, is the following:

Proposition 3.6 *Suppose $\phi : \Omega \to [k]$ is a partition. Then for any integers $k > m$ we have*

$$h_\mu\big(\tau,(\phi\circ\tau^m,\phi\circ\tau^{m+1},\dots,\phi\circ\tau^{k-1})\big) = h_\mu(\tau,\phi).$$

Proof This follows quickly from the definition of entropy, by observing first that by translating m and k together we may reduce to the case $m = 0$, and next that for any $n \geq 1$ the doubly-refined partition obtained by combining the single-refinements

$$(\phi,\phi\circ\tau,\dots,\phi\circ\tau^{k-1}),(\phi\circ\tau,\phi\circ\tau^2,\dots,\phi\circ\tau^k),\dots,(\phi\circ\tau^{(n-1)k},\phi\circ\tau^{(n-1)k+1},\dots,\phi\circ\tau^{nk-1})$$

(that is, $(n-1)k$ copies shifted by composition with τ at each step) is actually no greater a refinement than that obtained by taking only every k^{th} of the above sequence,

$$(\phi,\phi\circ\tau,\dots,\phi\circ\tau^{k-1}),(\phi\circ\tau^k,\phi\circ\tau^{k+1},\dots,\phi\circ\tau^{2k-1}),\dots,(\phi\circ\tau^{(n-1)k},\phi\circ\tau^{(n-1)k+1},\dots,\phi\circ\tau^{nk-1}).$$

However, the values

$$\frac{1}{nk}H\big((\phi,\phi\circ\tau,\dots,\phi\circ\tau^{k-1}),(\phi\circ\tau^{k},\phi\circ\tau^{k+1},\dots,\phi\circ\tau^{2k-1}),\dots,(\phi\circ\tau^{(n-1)k},\phi\circ\tau^{(n-1)k+1},\dots,\phi\circ\tau^{nk-1})\big)$$

are drawn from a subsequence of

$$\frac{1}{n}H(\phi,\phi\circ\tau,\dots,\phi\circ\tau^{n-1}),$$

and so converges to the same limit, $h_\mu(\tau,\phi)$. □

Combining the last two results as promised yields:

Corollary 3.7 *For any $\varepsilon > 0$ and fixed $K \geq 1$ there is some δ such that the following holds: If $\theta : \Omega \to [K]$ is a partition and $\phi : \Omega \to [k]$ is such that for some sufficiently large n the composite partition $(\phi,\phi\circ\tau,\dots,\phi\circ\tau^{n-1})$ is a δ-refinement of θ then*

$$h_\mu(\tau,\phi) > h_\mu(\tau,\theta) - \varepsilon.$$

□

This result will be crucial in proving Sinai's Theorem. Colloquially, it tells us that, in order to catch almost all of the entropy of some large partition θ, we need only ensure that the smaller partition ϕ approximately refines θ eventually, via $(\phi,\phi\circ\tau,\dots,\phi\circ\tau^{n-1})$ for some n, and not necessarily all at once (which is generally impossible if k is strictly smaller than K).

3.4 Dynamical system entropy and independence

Here our main result is the following:

Proposition 3.8 *For any $\varepsilon > 0$ there is some $\eta(\varepsilon) > 0$ such that whenever $\phi : \Omega \to [k]$ is a partition which satisfies $h_\mu(\tau,\phi) > H_\mu(\phi) - \eta(\varepsilon)$, then for any $n \geq 1$ the τ-shifted partition $\phi\circ\tau^n$ is ε-independent of the composite of its predecessors $(\phi,\phi\circ\tau,\dots,\phi\circ\tau^{n-1})$.*

Remark Compare this with the universal inequality in the opposite direction, $h_\mu(\tau,\phi) \leq H_\mu(\phi)$. ◁

Proof This follows from manipulating Proposition 2.14 and observing that for any n our hypotheses give

$$H_\mu(\phi) \geq \frac{1}{n+1}H\big(\mu\circ\big((\phi,\phi\circ\tau,\dots,\phi\circ\tau^{n-1}),\phi\circ\tau^n\big)^{-1}\big) \geq h_\mu(\tau,\phi) > H_\mu(\phi) - \eta(\varepsilon).$$

□

The following immediate corollary will be needed separately from the above:

Corollary 3.9 *If τ is invertible and $\phi : \Omega \to [k]$ is any partition for which $h_\mu(\tau,\phi) = H_\mu(\phi)$, then it is a Bernoulli partition; that is, the compositions $\phi\circ\tau^i$ for $i \in \mathbb{Z}$ are independent.* □

We will sometimes refer colloquially to a partition $\phi : \Omega \to [k]$ for which the shortfall $H_\mu(\phi) - h_\mu(\tau,\phi)$ is small as being 'entropically efficient'.

4 The Shannon-McMillan-Breiman Theorem

This is perhaps the first result about dynamical system entropy whose proof requires some serious effort. We will state two separate versions below, the first for relative and the second for absolute entropy. In fact it is a little tricky to recover the first from the second, since the second does not tell us anything about convergence almost everywhere, but it is all we will need. This is because, although it used almost exclusively in its absolute form, this result at least does make sense in both versions, and illustrates the connection with large deviation theory more clearly when stated relatively.

Theorem 4.1 (Shannon-McMillan-Breiman Theorem: relative version) *Let $(\Omega, \Sigma, \mu, \tau)$ be an* ergodic *probability measure-preserving system, $(X_0, \mathrm{T}_0, \nu_0)$ a target space and $\phi : \Omega \to X_0$ a measurable map such that $s := s_{\tau,\phi}(\mu : \nu_0) < \infty$. Then for μ-almost every $\omega \in \Omega$ and any $\varepsilon > 0$ there is some $n_1 \geq 1$ such that*

$$\mathrm{e}^{n(s-\varepsilon)} < \frac{\mathrm{d}\big(\mu \circ (\phi_{[n]}^{(\tau)})^{-1}\big)}{\mathrm{d}\nu_0^{\otimes [n]}}(\phi_{[n]}^{(\tau)}(\omega)) < \mathrm{e}^{n(s+\varepsilon)}$$

whenever $n \geq n_1$.

Theorem 4.2 (Shannon-McMillan-Breiman Theorem: weak absolute version) *Let $(\Omega, \Sigma, \mu, \tau)$ be an ergodic probability measure-preserving system and $\phi : \Omega \to [k]$ a partition such that $h := h_\mu(\tau, \phi) < \infty$. Then for any $\varepsilon > 0$ and $\delta > 0$ there is some $n_1 \geq 1$ such that for any $n \geq n_1$ the refined partition $(\phi, \phi \circ \tau, \dots, \phi \circ \tau^{n-1})$ covers all but at most δ of Ω by cells A whose measures satisfy*

$$\mathrm{e}^{-n(h+\varepsilon)} < \mu(A) < \mathrm{e}^{-n(h-\varepsilon)}.$$

Remarks 1. The assumption of ergodicity here is essential.

2. One consequence of the above absolute version of the theorem is that, for large n, the work of covering Ω is not shared equally among all the cells of $(\phi, \phi \circ \tau, \dots, \phi \circ \tau^{n-1})$; in fact, asymptotically there is some sub-collection of roughly e^{nh} cells (up to exponential order), each of which is doing roughly e^{-nh} of the work (up to exponential order), and most of the other cells – which might be the numerical majority – are very much smaller. ◁

We omit the proof[1].

5 Rokhlin's Theorem

In order to prove Sinai's Theorem, we will make essential use of the following result of Rokhlin, which is also of independent interest.

Theorem 5.1 *Let $(\Omega, \Sigma, \mu, \tau)$ be an atomless ergodic invertible probability-preserving system, and let n be any natural number and $\varepsilon > 0$. Then we can find a set $F \in \Sigma$ such that F, $\tau^{-1}F$, $\tau^{-2}F$, . . ., $\tau^{-(n-1)}F$ are disjoint and*

$$\mu\Big(\bigcup_{0 \leq i \leq n-1} \tau^{-i}F\Big) > 1 - \varepsilon.$$

Remark This theorem is easily extended to the case of non-invertible τ, but we omit these details for brevity. ◁

[1] I may put a full proof here later if I have the time.

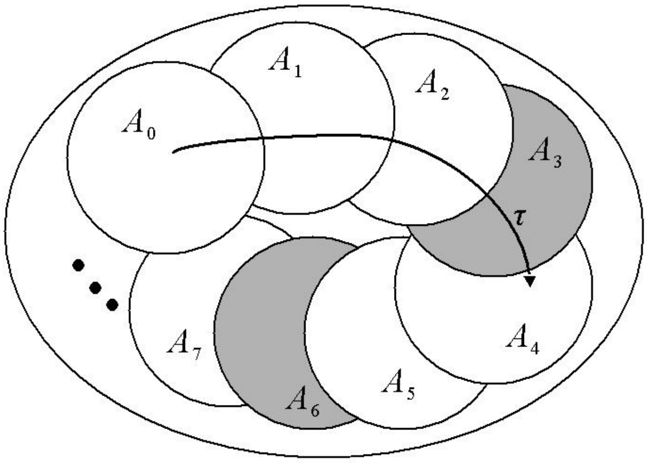

Figure 1: The construction of $A_0, A_1, A_2, \ldots$ and F in Rokhlin's Theorem: the components of F areshaded grey for the case $n = 3$.

Proof This is proved by just building a suitable set. First choose some $A_0 \in \Sigma$ with $0 < \mu(A_0) < \frac{\varepsilon}{n}$. Now, since the set $\bigcup_{i \geq 0} \tau^i A_0$ is τ-invariant and of positive measure, and τ is ergodic, it must actually fill almost all of X. We break it up into pieces: in addition to A_0, set

$$\begin{array}{rcl} A_1 & := & \tau A_0 \setminus A_0 \\ A_2 & := & \tau^2 A_0 \setminus (A_0 \cup A_1) \\ & \vdots & \end{array}$$

This decomposition is shown in Figure 1.

Finally, set $F := \bigcup_{k \geq 1} A_{kn}$; that is, the union of every n^{th} term in our disjoint sequence A_0, $A_1, \ldots$, starting from A_n. The desired properties for F now follow at once. □

Remark In some sense, this is our first 'structure-theoretic' result. ◁

6 Sinai's Theorem

Sinai's Theorem is the surprising converse to the following easy observation. Suppose we have a probability measure-preserving system $(\Omega, \Sigma, \mu, \tau)$ that we wish to compare with Bernoulli shifts built on the power space $([k]^{\mathbb{Z}}, (\mathcal{P}[k])^{\otimes \mathbb{Z}})$. Let us assume henceforth that $(\Omega, \Sigma, \mu, \tau)$ is invertible and ergodic (these are necessary for the results that follow). Suppose that we have also a 'target' probability distribution $\mathbf{p}$ on $[k]$, so that our specific question is whether $([k]^{\mathbb{Z}}, (\mathcal{P}[k])^{\otimes \mathbb{Z}}, \mathbf{p}^{\otimes \mathbb{Z}}, \sigma)$ is a factor of $(\Omega, \Sigma, \mu, \tau)$. One test for this is to ask whether $h_\mu(\tau) \geq H(\mathbf{p})$: certainly this is necessary, for if $\phi : \Omega \to \mathbb{N}$ is a suitable Bernoulli partition then $h_\mu(\tau) \geq h_\mu(\tau, \phi) = H(\mathbf{p})$.

Sinai's Theorem tells us that this condition is also sufficient.

Theorem 6.1 (Sinai, 1964) *Suppose that* $\mathbf{p}$ *is some target probability distribution on* $[k]$ *and that* $(\Omega, \Sigma, \mu, \tau)$ *is an invertible ergodic probability measure-preserving system for which* $\infty > h_\mu(\tau) \geq H(\mathbf{p})$. *Then this is* *enough to guarantee that we can find some measurable partition $\phi : \Omega \to [k]$ such that $([k]^{\mathbb{Z}}, (\mathcal{P}[k])^{\otimes\mathbb{Z}}, \mu \circ (\phi^{(\tau)})^{-1}, \sigma)$ is the Bernoulli shift $([k]^{\mathbb{Z}}, (\mathcal{P}[k])^{\otimes\mathbb{Z}}, \mathbf{p}^{\otimes\mathbb{Z}}, \sigma)$.*

Thus the main content of Sinai's Theorem must be a procedure for finding a suitable Bernoulli partition $\phi : \Omega \to \mathbb{N}$. To do this, we will seek a ϕ that has the following two properties:

- (ϕ has the correct law)
$$\mu \circ \phi^{-1} = \mathbf{p};$$
- (ϕ is entropically perfectly efficient)
$$h_\mu(\tau, \phi) = H(\mathbf{p}) = H_\mu(\phi)$$
(the second of these equalities follows from ϕ having the correct law).

By Corollary 3.9, the second of these forces ϕ to be a Bernoulli partition.

Let us pause to think informally about how we might go about this. We need to obtain a partition ϕ, given only knowledge of various entropy constraints. We will seek a method which obtains ϕ as a limit of some sequence of progressively-improved partition, where each improvement is made essentially 'by hand': given that each term of the sequence is a finite partition, we will refine it further, chopping up Ω into many smaller pieces, and then find some way to re-assign these smaller pieces, giving most of them back to their original cell of ϕ, but moving some of them to give our new partition better properties. This sequence $(\phi_i)_{i\geq 1}$ will be Cauchy in the sense that the 'error probabilities'

$$\mu\{\omega \in \Omega : \ \phi_i(\omega) \neq \phi_j(\omega)\} \to 0$$

for $i < j$ as $i \to \infty$; it is easy to see that we may then extract a limit partition.

The point here is that we have very few ways of constructing maps from one arbitrary measure space into another; even into a finite set. On the one hand, the lack of structure in the domain space puts very few ingredients at our disposal, and so it is hard to specify any single map at all. However, on the other hand, being measurable is very easy – in most more concrete situations it is one of the weakest assumptions one can make about a map – and so if there are any measurable maps having the desired properties at all, then typically there are very many of them, with no obviously canonical examples.

Thus the most basic device in our procedure for upgrading a given partition to a better one will be a way of carefully introducing certain more concrete ways of specifying that better partition as a modification of the first partition. These must give us enough freedom to reach a partition with the desired improved properties, but not so much freedom that we then cannot find that improved partition.

Our upgrading procedure is contained in the following proposition, whose proof forms the bulk of the work. Note that, for now, we are actually going to assume equality in our entropy constraint: $h_\mu(\tau) = H(\mathbf{p})$.

Proposition 6.2 (Partition upgrade procedure) *Suppose that $\mathbf{p}$ is some target probability distribution on $[k]$ and that $(\Omega, \Sigma, \mu, \tau)$ is an invertible ergodic probability measure-preserving system for which $\infty > h_\mu(\tau) = H(\mathbf{p})$. Given any $\varepsilon_1 > 0$ there is some $\delta_1 = \delta_1(\varepsilon_1) > 0$ such that the following holds. Whenever $\phi_1 : \Omega \to [k]$ is a partition such that:*

1. *(law of ϕ_1 close enough to p)*
$$\|\mu \circ \phi_1^{-1} - \mathbf{p}\| < \delta_1,$$
2. *(ϕ_1 is entropically efficient enough)*
$$h_\mu(\tau, \phi_1) > H_\mu(\phi_1) - \delta_1,$$

and whenever $0 < \delta_2 < \delta_1$, we can find a partition $\phi_2 : \Omega \to [k]$ such that:

1. *(law of ϕ_2* very *close to p)*
$$\|\mu \circ \phi_2^{-1} - \mathbf{p}\| < \delta_2;$$
2. *(ϕ_2 is* very *entropically efficient)*
$$h_\mu(\tau, \phi_2) > H_\mu(\phi_2) - \delta_2;$$
3. *(ϕ_2 equals ϕ_1 outside a small set)*
$$\mu\{\omega \in \Omega : \ \phi_2(\omega) \neq \phi_1(\omega)\} < \varepsilon_1.$$

Remark It is key to our use of this proposition that we are free to select any value of $\delta_2 > 0$, however small, even after ϕ_1 has been specified; only $\varepsilon_1 > 0$ affects the assumptions we need to make about δ_1 and ϕ_1. This is a reflection of the fact that, among our three desired conclusions about ϕ_2, only the last, involving ε_1, relates ϕ_2 to our original function ϕ_1; and so this proposition is telling us that, *however* much better we want to make ϕ_2 than ϕ_1 in the sense of conclusions one and two, we can always do it with some ϕ_2 within ε_1 of ϕ_1.

One consequence of this is that if we take ε_1 large, say at least 1, then our third conclusion relating ϕ_1 to ϕ_1 trivializes, but the other two conclusions do not. In this case, we can excise ϕ_1 altogether, to be left concluding the existence, for any δ_2, of a partition $\phi_2 : \Omega \to [k]$ satisfying

1. (law of ϕ_2 *very* close to p)
$$\|\mu \circ \phi_2^{-1} - \mathbf{p}\| < \delta_2;$$
2. (ϕ_2 is *very* entropically efficient)
$$h_\mu(\tau, \phi_2) > H_\mu(\phi_2) - \delta_2.$$

Thus we have concluded that, in the two particular senses at work above, we can find partitions ϕ_2 that approximate the desired Bernoulli behaviour arbitrarily well. Already this is a non-trivial result, and certainly it gives us hope that we can cover the extra distance to actually obtaining such a Bernoulli partition. This is what the full form of the proposition enables us to do. While the underlying construction cannot cover this last gap in a single leap, our extra conclusions involving ε_1 will tell us that, simply by being ambitious enough in our choice of $\delta_2 > 0$ at each step in an iterative procedure, we can make the distance ε_1 that we will need to move our partition next time decrease to zero as fast as we like. Thus we can form a sequence of partitions which, in addition to being better and better approximations to Bernoulli behaviour, actually converges to a perfectly Bernoulli limit partition in the sense of the third conclusion. Note that there is no useful notion of compactness in the topological space corresponding to this measure of the difference between two partitions, hence our need for a careful, deliberate construction of a sequence that converges. We will see the details of this argument in our final proof of Sinai's Theorem below. $\lhd$

Proof This is quite delicate and breaks into several steps. If only for didactic reasons, we will actually start by proving first the weakened form of the conclusion mentioned in the remarks above, in which we take $\varepsilon_1 \geq 1$. Having elaborated the ideas needed for this, we will run through the construction again, showing how we can guarantee closeness to ϕ_1 in addition. Note that we have gone to no effort to optimize the various constants that will appear in the process (preferring instead to use consistently $\frac{1}{100}$ as a convenient small fraction).

Steps 0.1 through 0.5 handle the case $\varepsilon_1 \geq 1$, forgetting about ϕ_1.

Step 0.1

We begin by observing that we may assume that $\mathbf{p}$ is not the uniform distribution on $[k]$, by embedding $[k]$ into $[k+1]$ and re-labelling $k+1$ as k if necessary; we will need this when we come to choose a 'more uniform' auxiliary distribution $\mathbf{q}$ on $[k]$ later. When, in the final proof of Sinai's Theorem, we eventually recover $\mathbf{p}$ as the law of a partition ϕ, we can just descend back to the original value k with adjustment by some cell of zero measure if necessary.

Let us introduce a new, auxiliary partition $\theta : \Omega \to [K]$, whose rôle is to catch most of the possible entropy of τ: say $h_\mu(\tau, \theta) > h_\mu(\tau) - \frac{1}{100}\delta_2$.

We will now try to construct a suitable good partition ϕ_2 by considering first some much finer partition and then carefully assigning its cells into small families, whose unions will then be the cells of ϕ_2. Our strategy is to do this in such a way that for some sufficiently large N and some sufficiently small ε_2 we have

$$(\phi_2, \phi_2 \circ \tau, \ldots, \phi_2 \circ \tau^N) \Rightarrow^{\varepsilon_2} \theta.$$

By the continuity result of Corollary 3.7 this would then imply that $h_\mu(\tau, \phi_2)$ is not much smaller than $h_\mu(\tau, \theta)$ (which we have chosen to be almost as large as possible); say we have chosen N and ε_2 so that this gap is also at most $\frac{1}{100}\delta_2$. Next, we will find that we can arrange the modification ϕ_2 so that the entropy $H_\mu(\phi_2)$ of its law is not much larger than $H(\mathbf{p})$, say again not by more than $\frac{1}{100}\delta_2$. Finally, since $H_\mu(\phi_2) \geq h_\mu(\tau, \phi_2)$, we will find that these two quantities will have been sandwiched together:

$$\begin{aligned} h_\mu(\tau) = H(\mathbf{p}) \quad &\geq \quad H_\mu(\phi_2) - \frac{1}{100}\delta_2 \\ &\geq h_\mu(\tau, \phi_2) - \frac{1}{100}\delta_2 \\ &\geq h_\mu(\tau, \theta) - \frac{2}{100}\delta_2 \\ &> h_\mu(\tau) - \delta_2, \end{aligned}$$

and so in particular the gap $H_\mu(\phi_2) > h_\mu(\tau, \phi_2)$ is less than δ_2, as we wanted. $\surd$

Step 0.2

Having decided to form the cells of ϕ_2 from cells of some finer partition, our first thought might be to use θ as this finer partition; this would amount to selecting a suitable map $\gamma_2 : [K] \to [k]$ and setting $\phi_2 = \gamma_2 \circ \theta$. However, it turns out that this is not yet enough freedom: we will have to do a little more, relying (in our current special case, and also in the full case to follow) on a trick using Rokhlin's Theorem 5.1. In this step we explain how this trick will work.

We will translate the problem into an (approximately) equivalent form. Suppose that for some $n \geq 1$ and some $\varepsilon_3 > 0$ (we will choose these later) we have used Rokhlin's Theorem to find some $F \subseteq \Omega$ with F, $\tau^{-1}F, \ldots, \tau^{-(n-1)}F$ disjoint and filling up all but $1 - \varepsilon_3$ of Ω.

(In fact, the construction underlying Rokhlin's Theorem gives us a way to find many such sets F, and it will turn out that we need it to have certain additional properties, and so must select it with a little more care from among these. We will first try to make these properties clearer by looking at the other parts of the proof, and will observe later that it is easy to make sure that our F is suitable.)

The partition θ induces a partition on each $\tau^{-i}F$, $0 \leq i \leq n-1$ by restriction, and now τ defines a natural bijection between functions ϕ defined on $\bigcup_{0 \leq i \leq n-1} \tau^{-i}F$ and n-tuples $(\phi^{(0)}, \phi^{(1)}, \ldots, \phi^{(n-1)})$ of functions defined on F by setting $\phi^{(i)} := (\phi|_{\tau^{-i}F}) \circ \tau^{-i} = (\phi \circ \tau^{-i})|_F$. Clearly this can be reversed to reconstruct ϕ on $\bigcup_{0 \leq i \leq n-1} \tau^{-i}F$.

Furthermore, so long as we have chosen ε_3 sufficiently small, it will suffice for us to find a suitable ϕ_2 restricted to the set $\bigcup_{0 \leq i \leq n-1} \tau^{-i}F$, and then define ϕ_2 arbitrarily on the remainder of Ω; this follows from our Proposition 3.5, which told us that 'approximate refinements catch almost all entropy'. Therefore, by the above identification, our problem can be studied instead by seeking a suitable n-tuple $(\phi_2^{(0)}, \phi_2^{(0)}, \ldots, \phi_2^{(n-1)})$ of functions on F.

This has the important advantage that, working with n-tuples defined only on F, we can make an n-tuple on F any way we like, and still be able to reconstruct a single function on $\bigcup_{0\leq i\leq n-1}\tau^{-i}F$ afterwards.

Clearly, we may also apply the above procedure to θ to obtain $\theta^{(0)}$, $\theta^{(1)}$, …, $\theta^{(n-1)}$.

Now, instead of seeking merely to write $\phi_2^{(i)} = \gamma_2\circ\theta^{(i)}$ — that is, $\phi_2 = \gamma_2\circ\theta$ — for some $\gamma_2 : [K]\to[k]$, we can try to choose a map $\tilde{\gamma_2} : [K]^n \to [k]^n$ to define ϕ_2 via the relationships

$$(\phi_2^{(0)},\phi_2^{(1)},\ldots,\phi_2^{(n-1)}) = \tilde{\gamma_2}\circ(\theta^{(0)},\theta^{(1)},\ldots,\theta^{(n-1)}).$$

We can then use our ability to reconstruct from this data a partition ϕ_2 on $\bigcup_{0\leq i\leq n-1}\tau^{-i}F$, and thence extend it to all of Ω, as observed above.

In this formulation our initial, naïve approach, choosing just $\gamma_2 : [K]\to[k]$, would amount to choosing a map $\tilde{\gamma_2}$ under which each coordinate of the image depends only on the corresponding coordinate in the domain, and each according to the same map γ_2. However, now we can choose some $\tilde{\gamma_2}$ for which the i^{th} coordinate of the image occasionally depends also on other coordinates of the domain than the i^{th}: since $\tilde{\gamma_2}$ is defined between much higher dimensional spaces, we have gained for ourselves more freedom in our choice, while still guaranteeing a single function ϕ_2 on Ω at the end of the process.

In fact, we will find it easier to choose our $\tilde{\gamma_2}$ — and so construct our new functions $\phi_2^{(0)}$, $\phi_2^{(1)}$, …, $\phi_2^{(n-1)}$ — by referring to the behaviour of θ, and also some other partitions that we have yet to construct, on the *whole* of Ω; then, having done this, we will return to the Rokhlin's Theorem-trick by using *that* construction of $\tilde{\gamma_2}$ to choose some F with certain additional properties, and then build ϕ_2 from the restrictions of $\phi_2^{(0)}$, $\phi_2^{(1)}$, …, $\phi_2^{(n-1)}$ to that particular F. Thus, ultimately we will need $\phi_2^{(0)}$, $\phi_2^{(1)}$, …, $\phi_2^{(n-1)}$ only through their restrictions to some F, and will discard the form they take elsewhere, but to find where is best to put our F we will want to know $\phi_2^{(0)}$, $\phi_2^{(1)}$, …, $\phi_2^{(n-1)}$ on the whole space.

The Rokhlin's Theorem-trick allows us introduce just enough 'parameters' (the coordinates of $\tilde{\gamma_2}$) to put us within reach of a suitable ϕ_2, but, as suggested earlier, not so many that it is impossible to find it. In the first instance, the high-entropy partition θ is the source of this extra data, but by itself this is not enough freedom; in addition, we need the ability to re-arrange smaller cells into a partition ϕ_2 in a slightly more complicated way. This will be just the same in the proof of the full result later. $\surd$

Step 0.3

Let us next ask how the desired properties of ϕ_2 translate into properties of $\phi_2^{(0)}$, $\phi_2^{(1)}$, …, $\phi_2^{(n-1)}$, and thence of $\tilde{\gamma_2}$.

1. For our first outcome, we want $\|\mu\circ\phi_2^{-1} - \mathbf{p}\| < \delta_2$. However,

$$\|\mu\circ\phi_2^{-1} - \mathbf{p}\| = \sum_{j\leq k}|\mu\{\omega\in\Omega :\ \phi_2(\omega) = j\} - p_j|,$$

and now considering separately our sets F, $\tau^{-1}F$, …, $\tau^{-(n-1)}F$ the right-hand side above is bounded by

$$\begin{aligned}&\sum_{j\leq k}\Big|\sum_{0\leq i\leq n-1}\mu\{\omega\in\tau^{-i}F :\ \phi_2(\omega) = j\} - p_j\Big| + \varepsilon_3\\ &= \sum_{j\leq k}\Big|\sum_{0\leq i\leq n-1}\mu\{\omega\in F :\ \phi_2^{(i)}(\omega) = j\} - p_j\Big| + \varepsilon_3.\end{aligned}$$

From this we see that our desired bound will follow if, firstly, ε_3 was chosen small enough (let us say, at most $\frac{1}{100}\delta_2$), and secondly the empirical distribution of the numbers $1,2,\ldots,k$ in the string $(\phi_2^{(0)}(\omega)$, $\phi_2^{(1)}(\omega)$, …, $\phi_2^{(n-1)}(\omega))$ is close enough to $\mathbf{p}$ for a sufficiently large proportion of

$\omega \in F$. This latter condition is now suitable for reformulation in terms of γ_2: we need the images $\gamma_2(\boldsymbol{\ell}) \in [k]^n$ of a point $\boldsymbol{\ell} = (\ell_0, \ell_1, \ldots, \ell_{n-1}) \in [K]^n$ to have empirical distribution close enough to $\mathbf{p}$ (say, within $\frac{1}{200}\delta_2$ in norm), not necessarily for *every* point of $[K]^n$, but for some set of points for which the corresponding cells under θ cover a large enough portion of F; say, all but at most $\frac{1}{n}\varepsilon_3 = \frac{1}{100n}\delta_2$ (it is this condition which will require us to think a little bit when choosing F later).

2. Now let us consider the second condition. As stated earlier, we hope to guarantee this by ensuring that $(\phi_2, \phi_2 \circ \tau, \ldots, \phi_2 \circ \tau^N) \Rightarrow^{\varepsilon_2} \theta$ for some large N and small ε_2. This, in turn, is saying that for all points $\omega \in \Omega$ outside some set of measure at most ε_2, we can tell which cell of θ they fall into by knowing their cells in ϕ_2, $\phi_2 \circ \tau$, ..., and $\phi_2 \circ \tau^N$. This will follow if we know that all but at most ε_2 of Ω is covered by cells of the refinement $(\phi_2, \phi_2 \circ \tau, \ldots, \phi_2 \circ \tau^N)$ each of which lies entirely within one cell of θ. In particular, we can guarantee this for $N = n$ sufficiently large if we know that all but at most ε_2 of Ω is covered by cells which have names under θ which are all mapped to *distinct* members of $[k]^n$ by $\tilde{\gamma_2}$

   Now, numerically the set $[K]^n$ is larger than $[k]^n$, since $K \geq k$; indeed, it may be much larger. But the cells of $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)})$ corresponding to different points of $[K]^n$ may have wildly differing sizes (indeed, it will follow from our argument below that they necessarily do, unless we are in the degenerate case $K = k$ in which case it will not matter), and so we might hope that we can cover almost all of Ω with cells corresponding to some much smaller subset of points in $[K]^n$, and then will need to worry only about the images of these points under $\tilde{\gamma_2}$. Indeed, this is so: it will follow from the Shannon-McMillan-Brieman Theorem 4.2. This gives us hope that we might be able to choose a $\tilde{\gamma_2}$ that also has this "almost-injectivity" property.

3. Of course, in our current special case the third outcome does not apply; we will worry about it more later. $\surd$

Step 0.4

Having decided what we are looking for, we are ready to select a suitable $\tilde{\gamma_2}$. The choice is highly arbitrary, and the way we will make it is not strictly constructive. Even after introducing θ and F we still do not have enough structure to specify a single suitable map easily; rather, we have now imposed a framework within which to prove a relevant non-constructive existence result.

We seek a map $[K]^n \to [k]^n$ which is injective on a certain subset of $[K]^n$ (to be specified shortly), and so we will first identify that subset and then arrange matters so that we can associate to each point in it a fairly large collection of possible choices for its image under $\tilde{\gamma_2}$. This will enable an appeal to Hall's Marriage Lemma to claim that a single suitable map $\tilde{\gamma_2}$ exists as a matching between the two resulting collections of points (this is the non-constructive part).

To do this we need to introduce one more auxiliary construction: another n-tuple of partitions $\Omega \to [k]$, say $(\psi^{(0)}, \psi^{(1)}, \ldots, \psi^{(n-1)})$. The idea is that, although their composite partition defines only k^n cells, these are of more equal sizes than the cells of the composite $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)})$, to such an extent that these cells are actually *smaller* than the (proportionately very few) cells of $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)})$ which are doing most of the work of covering Ω.

To find such $\psi^{(i)}$, we will first let $\mathbf{q}$ be some probability distribution on $[k]$ which is a little closer to uniform than $\mathbf{p}$, in that $H(\mathbf{q}) > H(\mathbf{p}) + \alpha$ for some $\alpha > 0$ (recall that we assumed early on that $\mathbf{p}$ is not the uniform distribution), but still with α very small, say $\alpha < \frac{1}{100}\delta_2$. Now we will choose $\psi^{(0)}, \psi^{(1)}, \ldots, \psi^{(n-1)}$ to be a strictly *independent* sequence of partitions on Ω which have law $\mathbf{q}$. (This time around we do not put any further restrictions on the $\psi^{(i)}$. For the full-strength result we will also need to construct them so as to be 'close to' $\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi_1^{(n-1)}$, making our task rather more subtle at this point, but we will worry about that later.)

We will use these $\psi^{(i)}$ to construct the setting for our application of the Marriage Lemma. Like the $\theta^{(i)}$, together they chop up Ω into many small cells; however, we have deliberately chosen $\mathbf{q} = \mu \circ (\psi^{(i)})^{-1}$

so that $H(\mathbf{q}) > H(\mathbf{p}) + \alpha = h_\mu(\tau) + \alpha \geq h_\mu(\tau,\theta) + \alpha$ (here is why we needed to assume equality in $H(\mathbf{p}) = h_\mu(\tau)$). Given this, an application of the Shannon-McMillan-Breiman Theorem 4.2 (for θ) and another of the Law of Large numbers (for the $\psi^{(i)}$) tell us that, for n sufficiently large,

- all but at most $\frac{1}{200}\varepsilon_2$ of Ω is covered by cells of $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)})$ of size
$$\mathrm{e}^{-n(h_\mu(\tau,\theta)+\mathrm{o}_{n\to\infty}(\alpha))} = \mathrm{e}^{-n(H(\mathbf{p})+\mathrm{o}_{n\to\infty}(\alpha))}$$
– in particular, let us say, of size greater than $\mathrm{e}^{-n(H(\mathbf{p})+\alpha/3)}$;
- all but at most $\frac{1}{200}\varepsilon_2$ of Ω is covered by cells from $(\psi^{(0)}, \psi^{(1)}, \ldots, \psi^{(n-1)})$ of size
$$\mathrm{e}^{-n(H(\mathbf{q})+\mathrm{o}_{n\to\infty}(\alpha))}$$
– in particular, let us say, of size less than $\mathrm{e}^{-n(H(\mathbf{q})-\alpha/3)}$.

Now, since
$$\frac{\mathrm{e}^{-n(H(\mathbf{q})-\alpha/3)}}{\mathrm{e}^{-n(H(\mathbf{p})+\alpha/3)}} \to 0 \quad \text{as } n \to \infty$$
(this is why we needed a positive gap $H(\mathbf{q}) > H(\mathbf{p})$), it follows that we can also assume n so large that all but at most $\frac{1}{100}\varepsilon_2$ of Ω is covered, on the one hand, by cells of $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)})$ (call this set G_1), and on the other, by cells of $(\psi^{(0)}, \psi^{(1)}, \ldots, \psi^{(n-1)})$ that are at most $\frac{1}{4}$ of their size (call this set G_2). Restricting attention to those cells A from these families that individually intersect $G_1 \cap G_2$ in at least half $\mu(A)$, we can therefore apply the Marriage Theorem to the bipartite graph obtained by joining each of these former cells to all of the latter that overlap with it, and so choose $\tilde{\gamma}_2$.

We finish this step by observing that, by our construction of the $\psi^{(i)}$ as having entropy $H(\mathbf{q})$, by choosing ε_2 smaller still if necessary we can ensure $H_\mu(\phi_2) > H(\mathbf{p}) - \frac{1}{100}\delta_2$.

Notice that in the above we did not need to construct $\psi^{(i)}$ with reference to θ at all; so long as our independence and law conditions are satisfied, for large enough n these $\psi^{(i)}$ will have the properties we want. Later on, however, we *will* need to tie the $\psi^{(i)}$ more closely to ϕ_1. $\surd$

Step 0.5

It remains to select our suitable F. In addition to the conclusions of Rokhlin's Theorem, we need it *not* to overlap the union G of the various error sets that we have accrued so far, whose measure is at most $\frac{1}{10}\varepsilon_2$. However, if we have obtained F_0 simply from Rokhlin's Theorem with a small enough remainder
$$\varepsilon_3 = \mu\Big(\Omega \setminus \bigcup_{0\leq i\leq n-1} \tau^{-i}F_0\Big),$$
then if the overlap $F_0 \cap G$ is not small enough, we can just choose some image $\tau^i F_0$ in place of F for some $0 \leq i \leq n-1$; it follows quickly that one such image does have a sufficiently small overlap with G, if these images cover a big enough part of Ω. This completes the proof of our reduced version of the proposition. $\surd$

This completes the construction for the special case. Before moving on, less us summarize the dependencies among the various additional quantities that we have introduced so far:

- First, θ is chosen with reference to δ_2: $h_\mu(\tau,\theta) > h_\mu(\tau) - \frac{1}{100}\delta_2$;
- Next, ε_2 is chosen with reference to δ_2 and θ, such that for *any* $n \geq 1$ the approximate refinement
$$(\phi_2, \phi_2 \circ \tau, \ldots, \phi_2 \circ \tau^{n-1}) \Rightarrow^{\varepsilon_2} \theta$$
implies $h_\mu(\tau,\phi_2) > h_\mu(\tau,\theta) - \frac{1}{100}\delta_2$;

- Next, α is chosen with reference to δ_2: $0 < \alpha < \frac{1}{100}\delta_2$.
- Next, $\mathbf{q}$ — the distribution of our $\psi^{(i)}$ to come — is chosen with reference to $\mathbf{p}$ and α.
- Next, $n \geq 1$ is chosen with reference to α, δ_2, ε_2 and $\mathbf{q}$, so that our deductions from the Shannon-McMillan-Breiman Theorem and the Law of Large Numbers are sure to go through once we have chosen our $\psi^{(i)}$.
- Next, our auxiliary partitions $\psi^{(i)}$ for $0 \leq i \leq n-1$ are chosen with reference to $\mathbf{q}$ and, clearly, n.
- Finally, our set F and error probability ε_3 are chosen with reference to n, ε_2, θ and the $\psi^{(i)}$.

We now pass through the same five steps as above a little more briskly, showing the modifications needed to ensure also that ϕ_2 is not too far from ϕ_1 (that is, showing that some positive $\delta_1(\varepsilon_1)$ can be found as in the statement of the theorem). Let us suppose without loss of generality that $\delta_2 < \frac{1}{100}\varepsilon_1$.

Step 1.1

Once again we introduce an auxiliary partition $\theta : \Omega \to [K]$ of nearly full entropy, except now we ask also that it refine ϕ_1 (by forming a composite with ϕ_1 if necessary). So, as before, assume $h_\mu(\tau, \theta) > h_\mu(\tau) - \frac{1}{100}\delta_2$. Now ϕ_1 factors through θ; write $\phi_1 = \gamma_1 \circ \theta$.

We are again going to assemble ϕ_2 by introducing some much finer partition and assigning its cells into families that will form the cells of ϕ_2. However, now we will perform this construction in such a way that most cells of ϕ_2 can be obtained as only a small modification of the corresponding cell of ϕ_1, by cutting out and re-assigning some small piece. Our goal will still be that for some sufficiently large N and some sufficiently small ε_2 we have

$$(\phi_2, \phi_2 \circ \tau, \ldots, \phi_2 \circ \tau^N) \Rightarrow^{\varepsilon_2} \theta;$$

provided once again that we can force $H_\mu(\phi_2)$ to be not much larger than $H(\mathbf{p})$, say again not by more than $\frac{1}{100}\delta_2$, this will give us the desired very high entropic efficiency just as in Step 0.1 above. $\surd$

Step 1.2

Here we run through the Rokhlin's Theorem-trick again for our current more delicate situation, bearing in mind now that we need to treat θ as a refinement of ϕ_1 in particular, not just some high-entropy partition.

We will select which small pieces of the cells of ϕ_1 to re-assign by chopping up its cells into a refinement and then moving some of the cells of that finer partition. Just as before, Rokhlin's Theorem will give us a way to introduce more freedom into how we do this. Suppose again that for some $n \geq 1$ and some $\varepsilon_3 > 0$ we have used Rokhlin's Theorem to find some $F \subseteq \Omega$ with F, $\tau^{-1}F$, ..., $\tau^{-(n-1)}F$ disjoint and filling up all but $1 - \varepsilon_3$ of Ω.

Now both ϕ_1 and θ induce partitions on each $\tau^{-i}F$, $0 \leq i \leq n-1$ by restriction, and so can be approximately encoded in the corresponding n-tuples $(\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi_1^{(n-1)})$ and $(\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)})$ of functions defined on F.

So long as we have chosen ε_3 sufficiently small, it will suffice for us to find a suitable re-definition ϕ_2 of ϕ_1 restricted to the set $\bigcup_{0 \leq i \leq n-1} \tau^{-i}F$, and then define ϕ_2 arbitrarily on the remainder of Ω. Therefore, by the above identification, our problem can be considered instead as one of seeking a suitable modification $(\phi_2^{(0)}, \phi_2^{(0)}, \ldots, \phi_2^{(n-1)})$ of the n-tuple $(\phi_1^{(0)}, \phi_1^{(0)}, \ldots, \phi_1^{(n-1)})$ of functions on F.

Now, in terms of these n-tuples, our factorizing map γ_1 is still playing its original part, but for each i separately:

$$\phi_1^{(i)} = \gamma_1 \circ \theta^{(i)}$$

(this follows at once from the definitions). We re-write this relationship as

$$(\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi^{(n-1)}) = \tilde{\gamma_1} \circ (\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)});$$

that is, replacing the separate action of $\gamma_1 : [K] \to [k]$ for each $i = 0, 1, \ldots, n-1$ with the action of a single function $\tilde{\gamma_1} : [K]^n \to [k]^n$. Of course, $\tilde{\gamma_1}$ has the property that the $i^{\rm th}$ coordinate of the image depends only on the $i^{\rm th}$ coordinate of the domain, and each of these dependencies is given by the same function γ_1, which is the simplistic behaviour we were forced to move away from in our previous construction of $\tilde{\gamma_2}$.

So, by analogy with Step 0.2, rather than just choosing a modification $\gamma_2 : [K] \to [k]$ of γ_1 from which to make ϕ_2, we can choose some more general modification $\tilde{\gamma_2} : [K]^n \to [k]^n$ of $\tilde{\gamma_1}$ and define ϕ_2 via the relationships

$$(\phi_2^{(0)}, \phi_2^{(1)}, \ldots, \phi_2^{(n-1)}) = \tilde{\gamma_2} \circ (\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(n-1)}).$$

Once again, we will find it easier to choose our $\tilde{\gamma_2}$ by referring to the behaviour of ϕ_1, θ, and some other partitions on the *whole* of Ω; then, having done this, we will return to the Rokhlin's Theorem-trick by using that construction of $\tilde{\gamma_2}$ to choose some F with certain additional properties, and then build ϕ_2 from the restrictions of $\phi^{(0)}$, $\phi^{(1)}$, ..., $\phi^{(n-1)}$ to that particular F. $\surd$

Step 1.3

The desired properties of ϕ_2 can be translated into properties of $\phi_2^{(0)}$, $\phi_2^{(1)}, \ldots, \phi_2^{(n-1)}$, and thence of $\tilde{\gamma_2}$, as in Step 0.3. The first two appear the same, but now we must worry about ensuring also the third property: closeness of ϕ_2 to ϕ_1. We want $\phi_2(\omega) = \phi_1(\omega)$ except possibly for ω coming from some error set of size at most ε_1. Let us in this case insist on $\varepsilon_3 < \frac{1}{100}\varepsilon_1$; now our conclusion will follow if we ask further that, for all but some proportionately small set of $\omega \in F$ (measure at most $\frac{1}{100n}\delta_2 < \frac{1}{10000n}\varepsilon_1$ will do), the string $(\phi_2^{(0)}(\omega), \phi_2^{(1)}(\omega), \ldots, \phi_2^{(n-1)}(\omega))$ disagrees with $(\phi_1^{(0)}(\omega), \phi_1^{(1)}(\omega), \ldots, \phi_1^{(n-1)}(\omega))$ in at most $\frac{n}{100}\varepsilon_1$ coordinates. This, in turn, will follow if for some points $\boldsymbol{\ell}$ of $[K]^n$ for which the corresponding cells cover all but our small part of F, the images $\tilde{\gamma_2}(\boldsymbol{\ell})$ and $\tilde{\gamma_1}(\boldsymbol{\ell})$ disagree in at most $\frac{n}{100}\varepsilon_1$ coordinates. $\surd$

Step 1.4

The choice of $\tilde{\gamma_2}$ proceeds using a sequence of independent auxiliary partitions $\psi^{(i)}$ and an application of Hall's Marriage Lemma just as in Step 0.4, except now we must choose these auxiliary partitions to be close also to the $\phi_1^{(i)}$.

As previously, let $\mathbf{q}$ be some probability distribution on $[k]$ which is a little closer to uniform than $\mathbf{p}$, in that $H(\mathbf{q}) > H(\mathbf{p}) + \alpha$ for some $\alpha > 0$ (recall that we assumed early on that $\mathbf{p}$ is not the uniform distribution), but with $\alpha < \frac{1}{100}\delta_2$. Now we will choose $\psi^{(0)}$, $\psi^{(1)}$, ..., $\psi^{(n-1)}$ to be a strictly *independent* sequence of partitions on Ω which have law $\mathbf{q}$, *and which approximate* $\phi_1^{(0)}$*,* $\phi_1^{(1)}$*, ...,* $\phi_1^{(n-1)}$ *in the sense that* $\mu\{\omega \in \Omega : \phi_1^{(i)}(\omega) \neq \psi^{(i)}(\omega)\} < \varepsilon_1$ *for* $0 \leq i \leq n-1$.

That we can choose such $\psi^{(i)}$ with this additional property is not instantly obvious; here we see an explicit manifestation of the rule-of-thumb that, in trying to construct measurable maps by hand, one quickly finds that one can call on as many 'degrees of freedom' as one needs. Indeed, first let $\psi^{(n-1)}$ be any map of distribution $\mathbf{q}$ approximating $\phi^{(n-1)}$ as stated; certainly this is possible. Now we establish the other $\psi^{(i)}$ by backwards induction. Suppose we already know $\psi^{(i+1)}$, $\psi^{(i+2)}$, ..., $\psi^{(n-1)}$. These together induce a large composite partition $(\psi^{(i+1)}, \psi^{(i+2)}, \ldots, \psi^{(n-1)})$, and similarly so do the predecessors of $\phi_1^{(i)}$, which give $(\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi_1^{(i-1)})$. However, since, by assumption,

$$h_\mu(\tau, \phi_1) > H_\mu(\phi_1) - \delta_1,$$

we know from Lemma 2.14 that, *uniformly in* n, there is a suitable choice of δ_1 (depending only on ε_1) so that each $\phi_1^{(i)}$ is forced to be $\frac{1}{100}\varepsilon_1$-independent from $(\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi_1^{(i-1)})$, which implies that its conditional distribution on most individual cells of $(\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi_1^{(i-1)})$ are close to its unconditional distribution $\mu \circ \phi_1^{-1}$. Hence for all but a few cells of $(\phi_1^{(0)}, \phi_1^{(1)}, \ldots, \phi_1^{(i-1)})$ (together covering little of Ω), we can modify $\phi_1^{(i)}$ to form $\psi^{(i)}$ cell-by-cell, so that $\psi^{(i)}$ has the desired distribution cell-wise, and differs from $\phi_1^{(i)}$ on only at most a proportion $\frac{1}{100}\varepsilon_1$ of the whole cell. Furthermore, we can chop up these cells further using the partition $(\psi^{(i+1)}, \psi^{(i+2)}, \dots, \psi^{(n-1)})$, and work cell-by-cell in that too ('we can always call on as many degrees of freedom as we need'), so guaranteeing also the independence of $\psi^{(i)}$ from $\psi^{(i+1)}$, $\psi^{(i+2)}$, $\dots$, $\psi^{(n-1)}$.

(It is worth taking a moment to consider the order of the quantifiers in the above: α depends on our uncontrolled very small quantity δ_2 and may also depend on ε_1; but our initial choice of δ_1 depends only on ε_1, since in order to construct the independent partitions $\psi^{(i)}$ with the desired law $\mathbf{q}$, so long as we took $\mathbf{q}$ close enough to $\mathbf{p}$, we need only that the $\phi^{(i)}$ satisfy the above approximate independence property with parameter a small constant multiple of ε_1.)

Having obtained these $\psi^{(i)}$ we construct $\tilde{\gamma}_2$ exactly as before; this time, since the $\psi^{(i)}$ are sufficiently close to the $\phi^{(i)}$, we can ensure additionally that for a set of points $\boldsymbol{\ell} \in [K]^n$ corresponding to cells covering all but a small enough part of Ω, the images $\tilde{\gamma}_1(\boldsymbol{\ell})$ and $\tilde{\gamma}_2(\boldsymbol{\ell})$ differ in at most $\frac{n}{100}\varepsilon_1$ coordinates, which is what we need from the previous Step. $\surd$

Step 1.5

It remains to select our suitable F; this is done from some F_0 originally guaranteed by Rokhlin's Theorem just as before. This completes the proof. $\surd$

□

Having obtained our upgrading procedure, we can complete the proof of Sinai's Theorem.

Proof of Theorem 6.1 Suppose first that we actually have equality: $h_\mu(\tau) = H(\mathbf{p})$. Select any summable sequence $(\varepsilon_i)_{i\geq 1}$ of strictly positive numbers. Applying the weaker, $\varepsilon_1 \geq 1$ case of Proposition 6.2 once we can obtain a partition $\phi_1 : \Omega \to [k]$ which we can ensure has distribution within $\delta_1(\varepsilon_1)$ of $\mathbf{p}$ and satisfies $h_\mu(\tau, \phi_1) > H_\mu(\phi_1) - \delta_1(\varepsilon_1)$ (where the function $\delta_1(\varepsilon_1)$ is as in Proposition 6.2).

Now we can apply the full strength of Proposition 6.2 repeatedly by induction to obtain the remaining terms of a sequence $\phi_1, \phi_2, \phi_3, \dots$ such that, for each $i \geq 1$,

- (convergence to $\mathbf{p}$)
$$\|\mu \circ \phi_i^{-1} - \mathbf{p}\| < \delta_1(\varepsilon_i);$$
- (convergence to perfect entropic efficiency)
$$h_\mu(\tau, \phi_i) > H_\mu(\phi_i) - \delta_1(\varepsilon_i);$$
- (convergence of the partitions themselves)
$$\mu\{\omega \in \Omega : \phi_{i+1}(\omega) \neq \phi_i(\omega)\} < \varepsilon_i.$$

By the third of the above properties, the partitions ϕ_i actually converge to some limit ϕ (unique up to negligible sets), and now first and second properties, together with the continuity result Proposition 3.5, guarantee that ϕ has law $\mathbf{p}$ and has perfect entropic efficiency: that is, it is a Bernoulli partition.

Finally, if $h_\mu(\tau) > H(\mathbf{p})$ then we can recover the above case by letting $\mathbf{p}_1$ be some refinement of $\mathbf{p}$ on a larger finite set such that $h_\mu(\tau) = H(\mathbf{p}_1)$. Now we can apply the theorem to find our Bernoulli factor map with law $\mathbf{p}_1$; composing with a map giving collapsing $\mathbf{p}_1$ back down to $\mathbf{p}$ gives the desired Bernoulli factor. □

7 Epilogue: Ornstein's Theorem

Although originally regarded as a major achievement in its own right, Sinai's Theorem is perhaps now best remembered as a staging post on the way to Ornstein's Theorem. This 1969 gave a complete answer to one of the longest-standing open problems in structural ergodic theory: When are two Bernoulli shift isomorphic? The introduction of entropy gave one invariant that could be used to distinguish non-isomorphic shifts; by using (and adding to) the ideas underlying Sinai's Theorem, Ornstein showed that this is enough: any two Bernoulli shifts of equal entropy are isomorphic.

Later still, in 1975, Keane and Smorodinsky were able to complete a different proof of Ornstein's Theorem. Their approach gives a much more explicit construction of the isomorphism in question; in particular, it allows the construction of such an isomorphism which has the additional property of being **finitary**: any one coordinate of the image of a sequence in the domain space actually depends on only finitely many coordinates of that original sequence. On the other hand, their approach does not help in the question of finding isomorphisms between other kinds of measurable dynamical system and Bernoulli shifts, as can the more abstract Sinai/Ornstein approach. A readable treatment of the Keane-Smorodinsky proof can be found in Petersen's book [4].

References

[1] Deuschel J-D. & Stroock D.W., *Large Deviations*, American Mathematical Society, Providence, Rhode Island, 1989;

[2] Fremlin D.H., *Measure Theory*, "Volume 3: Measure Algebras", Torres Fremlin, Colchester, 2002 (available online at `http://www.essex.ac.uk/maths/staff/fremlin/mt.htm`);

[3] Kolmogorov A.N., "A new invariant for transitive dynamical systems", *D.A.N. SSSR* **119** (1958), 861 – 869;

[4] Petersen K.E., *Ergodic Theory*, Cambridge University Press, Cambridge, 1983;

[5] Ruelle D., "Thermodynamic Formalism: The mathematical structures of classical equilibrium statistical mechanics", *Encyclopedia of Mathematics and Its Applications*, Vol. 5, Addison-Wesley, Reading, Massachusetts, 1978;

[6] Shannon C., "The mathematical theory of communication", *Bell Syst. Tech. J.* **27** (1948), 379 – 423;

[7] Smorodinsky M., *Ergodic Theory, Entropy*, Springer Lecture Notes in Mathematics, Berlin, 1971.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA AT LOS ANGELES, LOS ANGELES, CA 90095, USA
Email: `timaustin@math.ucla.edu`
Web: `www.math.ucla.edu/~timaustin`

October, 2006