Notes on Thermodynamic Formalism.

H. Bruin

October 5, 2017

Abstract

1 Historic Background

The program of statistical physics started with the aim of understanding how the microscopic considerations (say, the mechanics of single molecules) lead to macroscopic observations of the whole system. In particular, but not exclusively, the behaviour of an ideal gas (a liter of which roughly contains $2.7 \cdot 10^{22}$ molecules, *cf.* to the $\approx 2.710^{26}$ (the number of Avogadro) in a mol of substance.

- The macroscopic and experimental theory of the time included measurements of temperature, pressure, energy, etc. How should they be understood as averages over the movements of all the single particles? One is helped by conservation laws, such as the First Law of Thermodynamics:

  $$U = Q - W \quad \text{energy flow = heat flow minus net work done.}$$

- What exactly is **entropy**, *i.e.,* the measure of disorder of the system. It appears in the Second Law of Thermodynamics:

  $$dQ = TdS \quad \text{heat transfer = temperature times change of entropy}$$

  and the Third Law of Thermodynamics:

  the entropy of a perfect crystal at zero temperature is zero,

  but how to define such a quantity mathematically?

- Explain the macroscopic laws of thermodynamics, especially:

- Systems strive towards lowest energy.
- Systems strive towards largest entropy.

Especially the last was not without controversy: Newtonian reversibility doesn't reconcile with increase of entropy. Apart from this, there is Gibbs' paradox: Imagine a container divided into two parts A and B, which do communicate. The parts A and B are filled randomly with gases g_A and g_B. If g_A and g_B are different, the entropy increases by the mixture of the gases. If g_A and g_B are the same gas, then entropy is already maximal.

- Description of phase transitions:

$$\begin{array}{rcccc} \text{solid} & \leftrightarrow & \text{liquid} & \leftrightarrow & \text{gaseous} \\ \text{magnetized} & \leftrightarrow & \text{non-magnetized} & & \\ \text{super-conductive} & \leftrightarrow & \text{conductive} & \leftrightarrow & \text{non-conductive} \end{array}$$

Some of the protagonists of this theory are the following:

- James Clark Maxwell (1831 - 1879). Although his work of electromagnetism and color theory is much better known, he did publish works of kinetic theory of gas and also the "molecular" structure of conductors.
- Ludwig Boltzmann (1844 - 1906). He is basically the founder of this theory, but his ideas met with a lot of criticism. This was in a time when the existence of atoms was still a hypothesis rather than a fact. His theory didn't match empirical observations, and didn't yet fully explain the problems with entropy increase. Boltzmann moved between universities a lot (Graz, Vienna, Leipzig), being usually very unhappy if he shared his working place with scientific adversaries (*e.g.* Mach in Vienna, Oswald in Leipzig). Nonetheless he had success; for instance, his lectures on natural philosophy were so popular that the largest lecture hall in Vienna at the time wasn't big enough, and he had to move his lectures to a hall in the imperial palace.
- Josiah Willard Gibbs (1839 - 1903). Gibbs was the first to include probability theory in his analysis, and may be credited with coining the word "statistical mechanics". He published only late in his life. His 1902 book on the subject was praised by its mathematical rigor and elegance, but sometimes criticised for not really addressing the physical-philosophical problems of the structure of gases.
- Henri Poincaré (1854 - 1912). His theorem (now called Poincaré recurrence) gave a mathematical refutation of the principle of ever increasing entropy.
- Albert Einstein (1879 - 1955). Between 1902 to 1904, Einstein published several papers on statistical mechanics, which were definitely helpful in his treatise

of Brownian motion and of the photo-electric effect (1905), for which he would eventually receive the Nobel prize (*i.e.,* not for his relativity theory!).

- Paul and Tatiana Ehrenfest (1880 - 1933 and 1876 - 1964) published in 1910 an influential paper on the subject, in which they made the state of the art of the time accessible to a wide (German-speaking) audience. Among other things, they gave a more modest and mathematcally sound version of Boltzmann's famous but unrealistic Ergoden Hypothese.
- George Birkhoff (1884 – 1944) proved in 1931 the pointwise version of the Ergodic Theorem which was long anticipated, and was crucial for the development of the mathematical side of thermodynamic formalism.
- John von Neumann (1903 – 1957) proved the L^p version of the Ergodic Theorem. His result came before Birkhoff's but was published slightly later (in 1932). In this struggle for priority, Birkhoff was not entirely innocent.
- Wilhelm Lenz (1988 - 1957) was PhD supervisor of Ising (below), and suggested a now famous model of ferro-magnetism to Ising.
- Ernst Ising (1900 - 1998) treated this system in his thesis, found no "phase transitions" and but concluded that the model was insufficient to explain the magnetization of iron. He was in fact quite surprised to learn that 25 years after his thesis, people were still interested. In fact, the model doesn't explain magnetization in a one-dimensional model, but performs much better in dimension three.
- Andrej Kolmogorov (1903-1987), Russian probabilist and founding father of ergodic theory in Russia. His definiton of entropy paved the way for the current mathematical approach to thermodynamic formalism.

The introduction of thermodynamic formalism within the mathematical field of dynamical systems occurred in the 1970s, and was primarily due to the following people.

- Yakov Sinaĭ(1935 -): Mathematical physicists in moscow and Princeton, student of Kolmogorov. He proved the ergodicity of what is now called Sinai billiards, which is a model for bounding molecules in a gas.
- David Ruelle (1935 -): Mathematical physicists born in Belgium but worked mostly in France, mainly on statistical physics and turbulence. The Ruelle-Takens route to chaos, Ruelle's ζ-function and the Ruelle inequality are named after him.
- Rufus Bowen (1947 - 1978): Mathematician at Berkeley (USA), student of Field's medalist Stephen Smale, worked on Axiom A diffeomorphisms and symbolic dynamics.

1.1 Introductory example of the Ising model

This extended example is meant to give a feel for many of the ingredients in thermodynamic formalism. It is centred around a simplified Ising model, which can be computed completely.

We take the configuration space $\Omega = \{-1,+1\}^{\mathbb{Z}}$, that is the space of all bi-infinite sequences of $+1$'s and -1's. This give a rough model of ferro-magnetic atoms arranged on a line, having spin either upwards ($+1$) or downwards (-1). If all spins are upwards (or all downwards), then the material is fully magnetized, but usually the heat in the material means that atom rotate directing their spin in all directions over time, which we discretize to either up or down.

Of course, infinitely many atoms is unrealistic, and hence a configuration space $\{-1,+1\}^{[-n,n]}$ would be better (where $[-n,n]$ is our notation of the integer interval $\{-n,-n+1,\dots,n-1,n\}$), but for simplicity, let us look at the infinite line for the moment.

A **probability measure** μ indicates how likely it is to find a particular configuration, or rather a particular ensemble of configurations. For example, the fully magnetized states are expressed by the measures:

$$\delta_+(A) = \begin{cases} 1 & \text{if } A \ni (\dots,+1,+1,+1,+1,\dots) \\ 0 & \text{if } A \not\ni (\dots,+1,+1,+1,+1,\dots) \end{cases}$$

and δ_- with the analogous definition. For these two measures, only one configuration is likely to occur. Usually a single configuration occurs with probability zero, and we have to look at ensembles instead. Define **cylinder sets**

$$C_{m,n}(\omega) = \{\omega' \in \Omega : \omega'_i = \omega_i \text{ for } i \in [m,n]\}$$

as the set of all configurations that agree with configuration ω on sites i for $m \leqslant i \leqslant n$. Its **length** is $n-m+1$. Another notation would be $C_{m,n}(\omega) = [\omega_m\omega_{m+1}\dots\omega_n]$.

The Bernoulli measure (stationary product measure) μ_p is defined as[1]

$$\mu_p([\omega_m\omega_{m+1}\dots\omega_n]) = \prod_{i=m}^{n} p(\omega_i), \text{ where } p(+1) = p \text{ and } p(-1) = 1-p.$$

There is a Bernoulli measure μ_p for each $p \in [0,1]$ and $\mu_1 = \delta_+$, $\mu_0 = \delta_-$. However, for $p \in (0,1)$, every single configuration has measure 0. The Law of Large Numbers implies that the set of configurations in which the frequency of $+1$'s is anything else than p has zero measure.

Since physical problem is translation invariant. Define the **left-shift** as

$$\sigma(\omega)_i = \omega_{i+1}.$$

[1]This measure extends uniquely to all measurable sets by Kolmogorov's Extension Theorem.

Translation invariance of a measure then means shift-invariance: $\mu(A) = \mu(\sigma(A))$ for each ensemble $A \subset \Omega$. Many probability measures on Ω are not translation invariant, but fortunately, the examples μ_p above are.

Another example of shift-invariant measures are the Gibbs measures, associated to some potential function $\psi : \Omega \to \mathbb{R}$; the integral $\int_\Omega \psi \, d\mu$ is called the (potential) **energy** of μ.

Definition 1. *A measure μ is a* **Gibbs measure** *w.r.t. potential function $\psi : \Omega \to \mathbb{R}$ if there are constants $C > 0$ and $P \in \mathbb{R}$ such that for all cylinder sets $C_{m,n}$ and all $\omega \in C_{m,n}$,*

$$\frac{1}{C} \leqslant \frac{\mu(C_{m,n})}{\exp \sum_{i=m}^{n} (\psi \circ \sigma^i(\omega) - P)} \leqslant C. \tag{1}$$

The number P is called the **pressure***; in this setting it is a sort of normalizing constant, adjusting the exponential decrease of the denominator to the exponential decrease of the numerator*[2]

If we choose the potential to be

$$\psi(\omega) = \begin{cases} \log p & \text{if } \omega_0 = +1 \\ \log 1 - p & \text{if } \omega_0 = -1 \end{cases},$$

then the Bernoulli measure μ_p is actually a Gibbs measure, with pressure $P = 0$ and "distortion constant" $C = 1$. Indeed,

$$\mu(C_{m,n}(\omega)) = \prod_{i=m}^{n} p(\omega_i) = \prod_{i=m}^{n} e^{\psi(\sigma^i(\omega))} = \exp(\sum_{i=m}^{n} \psi(\sigma^i(\omega))),$$

and (1) follows.

The next ingredient is **entropy**. We postpone the precise definition, except for to say that there are different kinds. The system itself can have **topological entropy** $h_{top}(\sigma)$ which is independent of the measure, while each shift-invariant measure μ has its metric entropy or rather **measure theoretical entropy** $h_\mu(\sigma)$. For the Bernoulli measure μ_p, the measure theoretical entropy is

$$h_{\mu_p}(\sigma) = -(p \log p + (1-p)\log(1-p))$$

is the minus the expectation of ψ.

Exercise 1. *For $\varphi : [0,1] \to \mathbb{R}$ defined as $\varphi(x) = -(x \log x + (1-x)\log(1-x))$, we can write $h_{\mu_p}(\sigma) = \varphi(p)$. Compute the limits $\lim_{x \to 0} \varphi(x)$ and $\lim_{x \to 1} \varphi(x)$. Conclude that δ_+ and δ_- have zero entropy. (This agrees with the idea that entropy is suppose to measure disorder.) Where does φ assume its maximum? What does this suggest about the measure of maximal entropy?*

[2] This is the definition for one-dimensional lattices. For a d-dimensional lattice, we need to add an extra factor $(n - m + 1)^{d-1}$ in the lower and upper bounds in (1).

Exercise 2. *Compute its first and second derivative. Is φ (strictly) concave?*

Let us fix the potential

$$\psi(\omega) = \begin{cases} 0 & \text{if } \omega_0 = +1 \\ 1 & \text{if } \omega_0 = -1. \end{cases} \tag{2}$$

The potential energy $E(\mu) = \int_\Omega \psi \, d\mu$ becomes smaller for measures that favours configurations ω where many entries are $+1$. We can think of ψ as representing a fixed external magnetic field; the better the atoms align themselves to this field, the smaller the potential energy of their configuration. In extremo, $E(\delta_+) = 0$, but the entropy of δ_+ is zero, so we don't maximise entropy with this choice.

Pressure can also be defined by the Variational Principle. We introduce a weighing parameter $\beta \in \mathbb{R}$ between energy and entropy content of the measure. The physical interpretation of $\beta = 1/T$, where T stands for the absolute temperature (*i.e.,* degrees Kelvin normalised in some way), and thus it makes only physical sense to take $\beta \in (0, \infty)$, but we will frequently look at limit case $\beta \to 0$ and $\beta \to \infty$.

Now let the **(Variational) Pressure** be

$$P(\beta) = \sup\{h_\mu(\sigma) - \beta \int \psi \, d\mu \ : \ \mu \text{ is a shift-invariant probability measure}\} \tag{3}$$

A shift-invariant probability measure μ is called **equilibrium state** or **equilibrium measure**, if it assume the pressure in (3).

For the limit case $T \to \infty$, *i.e.,* $\beta \to 0$, the potential energy plays no role, and we are just maximising entropy. For the limit case $T \to 0$, *i.e.,* $\beta \to \infty$, the potential energy becomes all important, so in our example we expect δ_+ to be the limit equilibrium state. The physical interpretation of this statement is: as the temperature decreases to zero for some fixed external magnetic field (and also as the external magnetic field grows to infinity), the material becomes totally magnetized.

The question is now: do we find total magnetization (*i.e.,* the measure δ_+ as equilibrium state) also for some positive temperature (or finite external magnetic field)?

For each fixed measure, the function $\beta \mapsto h_\mu(\sigma) + \beta \int \psi \, d\mu$ is a straight line with slope $-\int \psi \, d\mu$ (non-positive because our potential ψ is non-negative) and abscissa $h_\mu(\sigma)$. If we look at (3) again, we can view the pressure function $\beta \mapsto P(\beta)$ as the envelope of all these straight lines. From this it follows immediately that $\beta \mapsto P(\beta)$ is continuous and convex (and non-increasing due to ψ being non-negative).

Once full magnetization is obtained, increasing β further will not change the equilibrium state anymore. Indeed, there is no measure that favours $\omega_i = +1$ more than δ_+. So if there is a finite β_0 such that δ_+ is equilibrium state, then $P(\beta) = 0$ for all $\beta \geqslant \beta_0$. We can call this a **freezing phase transition**, because at this parameter, the equilibrium state doesn't change anymore (as if the system is frozen in one configuration). The right-hand slope of the pressure function at β_0 is 0; how abrupt this **phase transition** is depends also on the left slope at β_0 which might be different from 0, but always $\geqslant 0$ because of convexity.

Let us now do the computation if there really is a phase transition at a finite β_0. For simplicity (and without justification at the moment) we will only compute the supremum in (3) over the Bernoulli measures μ_p. So then (3) simplifies to

$$P(\beta) = \sup_{p\in[0,1]} -(p\log p + (1-p)\log(1-p)) - \beta(1-p) =: \sup_{p\in[0,1]} F(\mu_p, \beta)$$

The quantity $F(\mu_p,\beta)$ is called the **free energy** of the measure μ_p. In our simplified case, it is a smooth curve in p, so to find the supremum (= maximum), we simply compute the derivative and put it equal to 0:

$$0 = \frac{\partial}{\partial p} F(\mu_p, \beta) = -(\log p - \log(1-p)) + \beta.$$

This is equivalent to $\log \frac{p}{1-p} = \beta$, *i.e.*,

$$p = \frac{e^\beta}{1+e^\beta}, \quad 1-p = \frac{1}{1+e^\beta}$$

Substituting in $P(\beta)$, we find that the pressure is

$$\begin{aligned}
P(\beta) &= -\left(\frac{e^\beta}{1+e^\beta}\log\frac{e^\beta}{1+e^\beta} + \frac{1}{1+e^\beta}\log\frac{1}{1+e^\beta}\right) - \beta\frac{1}{1+e^\beta} \\
&= -\left(\underbrace{\frac{e^\beta+1}{1+e^\beta}\log\frac{e^\beta}{1+e^\beta}} + \underbrace{\frac{1}{1+e^\beta}\log\frac{1}{1+e^\beta} - \frac{1}{1+e^\beta}\log\frac{e^\beta}{1+e^\beta}}\right) - \beta\frac{1}{1+e^\beta} \\
&= -\left(\log\frac{e^\beta}{1+e^\beta} - \frac{\beta}{1+e^\beta}\right) - \frac{\beta}{1+e^\beta} \\
&= \log(1+e^{-\beta}) \qquad \begin{cases} \to 0 & \text{as } \beta\to\infty \\ = \log 2 & \text{if } \beta = 0 \\ \sim -\beta & \text{as } \beta \to -\infty \end{cases}
\end{aligned}$$

So the pressure function is smooth (even real analytic) and never reaches the line $\beta \equiv 0$ for any finite β. Hence, there is no phase transition.

Exercise 3. *Verify that for potential* (2), μ_p *is indeed a Gibbs measure. For which value of the pressure? Here it is important to incorporate the factor* $-\beta$ *in the potential, so* $\psi_\beta(\omega) = 0$ *if* $\omega_0 = 1$ *and* $\psi_\beta(\omega) = -\beta$ *if* $\omega_0 = -1$.

In the proper Ising model, the potential also contains also a local interaction term between nearest neighbors:

$$\psi(\omega) = \sum_i J\omega_i\omega_{i+1} + \psi_{ext}(\omega),$$

where $J < 0$, so neighboring atomic magnets with the same spin have lower joint energy than neighboring atoms with opposite spin. The term $\psi_{ext}(\omega)$ still stands for the external magnetic field, and can be taken as ψ in (2). This gives a problem for the infinite lattice, because here all configurations have a divergent sum $\sum_i J\omega_i\omega_{i+1}$. Ising's solution to this problem lies in first dealing with a large lattice $[-n, n]$, so the configuration space is $\{-1, +1\}^{[-n,n]}$, and considering the Gibbs measures and/or equilibrium states projected to fixed finite lattice $[-m, m]$ (these projections are called **marginal measures**), and then letting n tend to infinity. Such limits are called **thermodynamic limits**. If there is no external magnetic field (*i.e.,* $\psi_{ext} \equiv 0$), then as $\beta \to \infty$, $n \to \infty$, there are two ergodic thermodynamic limits, namely δ_+ and δ_-. There is no preference from one over the other; this preference would arise if the is an external magnetic field of definite direction. However, no such magnetization takes place for a finite β. For this reason, Ising dismissed the model as a good explanation for magnetization of iron (and other substances). However, as was found much later, on higher dimensional lattices, the Ising model does produce phase transitions and magnetization at finite values of β (*i.e.,* positive temperature).

2 Configuration Spaces, Subshifts of Finite Type and Symbolic Dynamics

In this section we provide some examples of frequently occurring configuration spaces, and we want to give an indication of their size, which is directly related to the complexity of the maps we define on them. We start with symbolic spaces, which may seem the most abstract, but which are used to code dynamics on more concrete space symbolically.

2.1 Symbolic spaces

Let $\mathcal{A} = \{0, \dots, N-1\}$ be some finite collection of symbols, called the **alphabet**. We can make finite words of these symbols by concatenation; the notation of the collection of these finite words is $\mathcal{A}^*$. More interesting are the infinite words of symbols, $\mathcal{A}^{\mathbb{N}_0}$ or $\mathcal{A}^{\mathbb{Z}}$, depending on we have one-sided (with $\mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$) or two-sided infinite words. If $\Omega = \mathcal{A}^{\mathbb{N}_0}$ or $\mathcal{A}^{\mathbb{Z}}$, then we can define the **left-shift** $\sigma : \Omega \to \Omega$ as

$$\sigma(\omega)_i = \omega_{i+1} \text{ for all } i \in \mathbb{N} \text{ or } \mathbb{Z}.$$

The space (Ω, σ) is called the one-sided and two-sided **full shift** on n letters. The left-shift is invertible (and σ^{-1} is the right-shift) on two-sided infinite words, *i.e.,* on $\Omega = \mathcal{A}^{\mathbb{Z}}$. Define a metric on Ω as

$$d(\omega, \omega') = \sum_n 2^{-|n|}(1 - \delta(\omega_n, \omega'_n)), \qquad \text{where the Dirac delta } \delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$$

In this metric, two words ω and ω' are close together if they agree on a large block around the zero-th coordinate.

Exercise 4. *Show that in the above topology, Ω is a Cantor set. That is: Ω is compact, has no isolated points and is totally disconnected (each of its connected components is a point).*

Exercise 5. *There is nothing special about the number 2. We could take any $\lambda > 1$ instead, obtaining a metric*

$$d_\lambda(\omega, \omega') = \sum_n \lambda^{-|n|}(1 - \delta(\omega_n, \omega'_n)).$$

Show that the metrics d and d_λ are not equivalent in the sense that there would be $C > 0$ so that

$$\frac{1}{C} d(\omega, \omega') \leqslant d_\lambda(\omega, \omega') \leqslant C d(\omega, \omega').$$

On the other hand, show that the identity map $I : (\Omega, d) \to (\Omega, d_\lambda)$ is uniformly continuous, with uniformly continuous inverse.

Exercise 6. *Take $\lambda > 2$ and show that the identity map $I : (\Omega, d) \to (\Omega, d_\lambda)$ is Hölder continuous,* i.e., *there are C and exponent α such that*

$$d_\lambda(\omega, \omega') \leqslant C d(\omega, \omega')^\alpha.$$

What is the largest value of α that we can take?

Definition 2. *A set $\Sigma \subset \Omega$ is called a* **subshift** *if it is closed and shift-invariant,* i.e., *$\sigma(\Sigma) \subset \Sigma$.*

The prime example of a subshift are the **subshifts of finite type (SFT)** in which the occurrence of a finite collection of words is forbidden. For example, the Fibonacci SFT

$$\Sigma_{\neg 11} = \{\omega \in \{0,1\}^{\mathbb{N}} \text{ or } \{0,1\}^{\mathbb{Z}} \ : \ \omega_i \omega_{i+1} \neq 11 \text{ for all } i\}.$$

Naturally, we can think of SFTs in which blocks of length > 2 are forbidden, but since there are only finitely many forbidden word, we can always recode the subshift (using a larger alphabet) so as to obtain a SFT in which only some words of length 2 are forbidden. This means that we can define a **transition matrix**:

$$A = (a_{i,j})_{i,j=0}^{N-1} \qquad a_{i,j} = \begin{cases} 1 & \text{if the word } ij \text{ is allowed,} \\ 0 & \text{if the word } ij \text{ is forbidden.} \end{cases}$$

For example, the transition matrix of $\Sigma_{\neg 11}$ is $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. The transition matrix A is a non-negative matrix, and hence (by the Perron-Frobenius Theorem) has a real eigenvalue $\rho(A)$ which is at least as large as any other eigenvalue of A. In fact, if A is irreducible and a-periodic (that is, there is $n_0 \in \mathbb{N}$ such that A^n is a strictly positive matrix for all $n \geqslant n_0$), then $\rho(A)$ has multiplicity one and is strictly larger than the absolute value of every other eigenvalue.

Exercise 7. *Examine the eigenvalues of*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

to see what lack of irreducibility and of non-periodicity can mean.

Exercise 8. *Recode the SFT $\Sigma_{\neg 101,111}$ to a SFT with forbidden words of length 2 only. Compute the corresponding transition matrix.*

Example 1. *One can consider configuration spaces over higher-diemnsional lattices, for instance*

$$\Omega = \{-1, +1\}^{\mathbb{Z}^2}$$

for which every $\omega \in \Omega$ is an infinite square patterns of -1's and $+1$s. There are now two shifts, the left-shift σ with σ^{-1} being the right-shift, but also the down-shift τ with τ^{-1} being the up-shift, and $\tau \circ \sigma = \sigma \circ \tau$. Also here you can consider subshifts of finite type, for instance $\Sigma_{\neg 11, {1 \atop 1}}$ would stand for all infinite square patterns of -1's and $+1$s without two $+1$s begin left-right or up-down neighbors, but diagonal neighbors is allowed. In this case, the word-complexity $p(m, n)$, here the number of different patterns in an $m \times n$-block, is an open problem. We don't know the precise value of $\lim_n \frac{1}{n} \log p(n, n)$.

2.2 Sizes of symbolic spaces

Recall that **cylinder sets** of length $n - m + 1$ are

$$C_{m,n}(\omega) = [\omega_m \omega_{m+1} \dots \omega_n] = \{\omega' \in \Omega : \omega'_i = \omega_i \text{ for } i \in [m, n]\}$$

as the set of all configurations that agree with configuration ω on sites i for $m \leqslant i \leqslant n$. If Ω is a one-sided shift-space, then it is convenient to abbreviate n-cylinders as $C_n(\omega) = [\omega_0 \dots \omega_{n-1}]$.

Definition 3. *The word-complexity of a subshift Σ is defined as*

$$p(n) = p(n, \Sigma) = \#\{\textit{different } n\textit{-cylinders in } \Sigma\}.$$

Clearly $p(n) = N^n$ for the full-shift on N letters. For the Fibonacci shift $\Sigma_{\neg 11}$ on two letters we have

$$p(1) = 2, p(2) = 3, p(3) = 5, \dots, p(n) \text{ is } n + 1\text{st Fibonacci number.}$$

To see why this is true, let $p_0(n)$ be the number of n-cylinders ending with 0 and $p_0(n)$ be the number of n-cylinders ending with 1. Then $p_1(n) = p_0(n-1)$, because every 1 must have been precede by a 0. On the other hand, $p_0(n) = p_1(n-1) + p_0(n-1) = p_0(n-2) + p_0(n-1)$. With initial values $p_0(1) = 1$ and $p_0(2) = 2$, it follows immediately that $p_0(n)$ is the n-th Fibonacci number. The step to $p(n)$ is now easy.

Theorem 1. *For a SFT with transition matrix A,*

$$\lim_{n\to\infty} \frac{1}{n} \log p(n) = \log \rho(A)$$

is the logarithm of the largest (Perron-Frobenius) eigenvalue of A.

Proof. Matrix multiplication $A^n = (a^n_{i,j})_{i,j=0}^{N-1}$ gives the number of allowed words of length $n+1$ that start with i and end with j. Hence the total number of words of length $n+1$ is $\sum_{i,j=0}^{N-1} a^n_{i,j}$, but the latter grows as $\rho(A)^n$ (disregarding polynomial factors if the Jordan block associated to $\rho(A)$ is non-trivial). Therefore $\lim_{n\to\infty} \frac{1}{n} \log p(n) = \lim_{n\to\infty} \frac{1}{n} \log \rho(A)^n = \log \rho(A)$, as required. □

2.3 Further configuration spaces

In this subsection, we review some configuration spaces and maps acting on them that frequently occur in thermodynamic formalism an dynamics as a whole. Without a physical interpretation as direct, maybe, as $\{-1,+1\}^{\mathbb{Z}}$ as the simple Ising model of Section 1, it is quite common within mathematics to use manifolds as configuration space Ω. In this setting, the word "configuration" seems less apt, so we tend to prefer the word **phase space** instead, even though this has nothing to do with phase transition in physics. Maybe **dynamical space** would yet be better, but that is not used so often, in fact only to distinguish it from parameter space.

Examples of phase spaces are: the unit interval $[0,1]$, the unit circle $\mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$, the d-dimensional torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$, the d-dimensional sphere $\mathbb{S}^d$, etc. In this case, there is no shift, but we have to specify the dynamics $f : \Omega \to \Omega$ explicitly, and it doesn't express translation invariance anymore.

Example 2. *An example is the angle doubling map $T_2 : \mathbb{S}^1 \to \mathbb{S}^1$, defined as*

$$T_2(x) = 2x \pmod 1,$$

*and the generalization $T_d(x) = dx \pmod 1$ for any $d \in \mathbb{Z}$ is easy to grasp. In fact, there is no reason to stick to integer d; we can define the β-**transformation** (see Figure 1) for any $\beta \in \mathbb{R}$ as*

$$T_\beta(x) = \beta x \pmod 1.$$

This is not continuous anymore for non-integer β, taking away the advantage of the circle $\mathbb{S}^1$. Therefore, the β transformation is usually defined on the unit interval: $T_\beta : [0,1] \to [0,1]$.

Example 3. *The* **rotation map** *$R_\gamma : \mathbb{S}^1 \to \mathbb{S}^1$ is defined as*

$$R_\gamma(x) = x + \gamma \pmod 1,$$

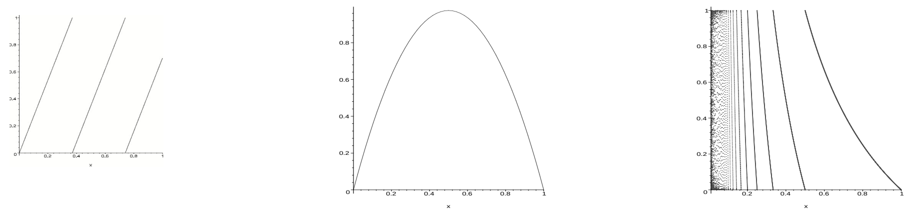

Figure 1: The β-transformation ($\beta = 2.7$), quadratic Chebyshev map and Gauss map.

and depending on whether the rotation angle γ is rational or not, every orbit is periodic or dense. One can easily construct higher dimensional analogs, i.e., *rotations $R_{\vec{\gamma}} : \mathbb{T}^d \to \mathbb{T}^d$ with a d-dimensional rotation vector.*

Example 4. *The integer matrix $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ acts as a linear transformation on $\mathbb{R}^2$, with eigenvalues $\lambda_\pm = \frac{3\pm\sqrt{5}}{2}$ and eigenvectors $\vec{v}_\pm = \begin{pmatrix} \frac{1\pm\sqrt{5}}{2} \\ 1 \end{pmatrix}$. It preserves the integer lattice $\mathbb{Z}^2$, so it is possible to defined the factor map $f : \mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{T}^2$ as*

$$f \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \pmod 1.$$

This is called **Arnol'd cat-map** *because Arnol'd in his book [2, 3] uses a picture of a cat's head to show what happens to shapes when iterated by f. The tangent space T_p of the fixed point $p = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ (and in fact near every point), decomposes in an unstable direction $E^u(p) = \operatorname{span} \vec{v}_+$ and a stable direction $E^s(p) = \operatorname{span} \vec{v}_-$. This means that shapes are stretched by a factor λ_+ in the $\vec{v}_+$-direction and contracted by a factor λ_- in the $\vec{v}_-$-direction. But since $\lambda_+ \cdot \lambda_- = 1$, the area doesn't change under f.*

Definition 4. *A map f on a d-dimensional manifold M is called* **Anosov** *if*

- *At each $p \in M$, there is a splitting of the tangent space $T_pM = E^u(p) \oplus E^s(p)$ which is invariant under the derivative map: $Df_p(E^{u/s}(p)) = E^{u/s}(f(p))$.*
- *The splitting $E^u(p) \oplus E^s(p)$ depends continuously on the point p.*
- *There is exponential expansion and contraction along $E^{u/s}(p)$,* i.e., *there is $\lambda > 1$ and $C > 0$ such that*

$$\begin{aligned} \|Df_p^n \cdot v\| &\geqslant C\lambda^n \|v\| \text{ for all } v \in E^u \text{ and } p \in M; \\ \|Df_p^n \cdot v\| &\leqslant \frac{1}{C}\lambda^{-n} \|v\| \text{ for all } v \in E^s \text{ and } p \in M. \end{aligned}$$

2.4 Symbolic dynamics

In this section we make the connection between dynamics on manifolds and symbolic spaces. The latter is a coded version of the first. For this coding, we need a partition $\mathcal{P} = (X_n)$ of the phase space X, and each partition element has X_n a label, which is a letter from the alphabet $\mathcal{A}$. The orbit $\mathrm{orb}(x) = (f^k(x))$ is associated to a code $i(x) = i_0(x)i_2(x)\dots$ (or $i(x) = \dots i_{-1}(x)i_0(x)i_2(x)\dots$ if f is invertible) defined as

$$i_k(x) = n \text{ if } f^k(x) \in X_n. \tag{4}$$

The **coding map** or **itinerary map** $i : X \to \Omega = \mathcal{A}^{\mathbb{N}_0}$ or $\mathcal{A}^{\mathbb{Z}}$ need not be injective or continuous, but hopefully, the points where i fails to be injective or continuous are so small as to be negligible in terms of the measures we are considering, see Section 3. The following commuting diagram holds:

$$\begin{array}{ccc} X & \xrightarrow{f} & X \\ {\scriptstyle i}\downarrow & & \downarrow{\scriptstyle i} \\ \Omega & \xrightarrow[\sigma]{} & \Omega \end{array} \qquad \text{In other words: } i \circ f = \sigma \circ i.$$

Example 5. *Let us look at the angle doubling map $T_2 : \mathbb{S}^1 \to \mathbb{S}^1$. It is natural to use the partition $I_0 = [0, \frac{1}{2})$ and $[\frac{1}{2}, 1)$ because $T_2 : I_k \to \mathbb{S}^1$ is bijective for both $k = 0$ and $k = 1$. Using the coding of (4), we find that $i(x)$ is just the binary expansion of x! Note however that $i : \mathbb{S}^1 \to \Omega =: \{0, 1\}^{\mathbb{N}_0}$ is not a bijection. For example, there is no $x \in \mathbb{S}^1$ such that $i(x) = 11111\dots$ Also i is not continuous. For example,*

$$\lim_{y \uparrow \frac{1}{2}} i(y) = 0111\cdots \neq 1000\cdots = \lim_{y \downarrow \frac{1}{2}} i(y).$$

The points $x \in \mathbb{S}^1$ where i is discontinuous is however countable (namely all dyadic rationals) and the points $\omega \in \Omega$ for which there is no x with $i(x) = \omega$ is also countable (namely those sequences ending in an infinite block of ones. These are so small sets of exceptions that we decide to neglect them. It is worth noting that in general no choice of itinerary map can be a homeomorphism, simply because the topology of a manifold is quite different from the topology of the symbol space Ω, i.e., *a Cantor set.*

Exercise 9. *Show that $i^{-1} : \Omega \to \mathbb{S}^1$ is continuous (wherever defined). Is it Hölder continuous or even Lipschitz?*

Example 6. *Coding for the cat-map.*

Definition 5. *Given a dynamical system $f : X \to X$, a partition $\{X_k\}$ of X is called a* **Markov partition** *if $f : X_k \to f(X_k)$ is a bijection for each k, and:*

- *If f is non-invertible: $f(X_k) \supset X_l$ whenever $f(X_k) \cap X_l \neq \emptyset$.*

- *If f is invertible: $f(X_k)$ stretches entirely across X_l in the expanding direction whenever $f(X_k) \cap X_l \neq \emptyset$, and $f^{-1}(X_k)$ stretches entirely across X_l in the contracting direction whenever $f^{-1}(X_k) \cap X_l \neq \emptyset$,*

Using a Markov partition for the coding, the resulting coding spaces is a subshift of finite type (two-sided or one-sided according to whether f is invertible or not.

Theorem 2. *Every Anosov diffeomorphism on a compact manifold has a finite Markov partition.*

We will not prove this theorem, cf. [4, Theorem 3.12]. In general the construction of such a Markov partition is very difficult and there doesn't seem to be a general practical method to create them. Therefore we restricted ourselves to the standard example of the cat-map.

Example 7. *Let $R_\gamma : \mathbb{S}^1 \to \mathbb{S}^1$ be a circle rotation over an irrational angle γ. Take the partition $I_0 = [0, \gamma)$ and $I_1 = [\gamma, 0)$. This is not a Markov partition, so the symbolic dynamics resulting from it is not a subshift of finite type. Yet, it gives another type of subshift, called* **Sturmian subshift** *Σ_γ, and is obtain as the closure of $i(\mathbb{S}^1)$, or equivalently (since every orbit of R_γ is dense in $\mathbb{S}^1$) the closure of $\{\sigma^n(i(x)) : n \in \mathbb{N}_0\}$.*

2.5 Complexity of maps on phase spaces

We have defined word-complexity in Definition 3. Now that we have introduced symbolic dynamics, this immediately gives a measure of the complexity of maps. Given partitions $\mathcal{P}$ and $\mathcal{Q}$ of X (Markov partition or not), call

$$\mathcal{P} \vee \mathcal{Q} = \{P \cap Q \ : \ P \in \mathcal{P} \text{ and } Q \in \mathcal{Q}\}$$

be the **joint** of $\mathcal{P}$ and $\mathcal{Q}$. Let $f : X \to X$ be the dynamics on X and let $f^{-1}\mathcal{P} = \{f^{-1}(P) \ : \ P \in \mathcal{P}\}$. Define

$$\mathcal{P}_n = \bigvee_{k=0}^{n-1} f^{-k}\mathcal{P}$$

Lemma 1. *Each $P_n \in \mathcal{P}_n$ corresponds to exacly one cylinder set of length n in the coding space Ω of (X, f) w.r.t $\mathcal{P}$. (For this reason, we call the elements of $\mathcal{P}_n$ cylinder sets as well.)*

Exercise 10. *Prove Lemma 1.*

It turns out (as we shall see in Section 4) that the exponential growth rate of $\#\mathcal{P}_n$ is largely independent of the finite partition we take[3]. Therefore we can define the

[3]But naturally, there are always partitions where it doesn't work, *e.g.* the trivial partition - it is important that $f : P \to f(P)$ is bijective on each $P \in \mathcal{P}$.

topological entropy

$$h_{top}(f) = \lim_{n\to\infty} \frac{1}{n} \log \#\mathcal{P}_n = \lim_{n\to\infty} \frac{1}{n} \log p(n).$$

Lemma 2. *Let H be the set of points determining a partition $\mathcal{P}$ of a (non-invertible) dynamical system $([0,1], f)$; we suppose $H \supset \{0,1\}$. Then $\#\mathcal{P}_n = \#\left(\bigcup_{k=0}^{n-1} f^{-k}(H)\right)$.*

Proof. Each element of $\mathcal{P}_n = \bigvee_{k=0}^{n} f^{-k}\mathcal{P}$ corresponds to exactly one component of the complement of $\bigcup_{k=0}^{n-1} f^{-k}(H)$. Since $\{0,1\} \subset \bigcup_{k=0}^{n-1} f^{-k}(H)$ by our choice of H, these points separate $[0,1]$ in exactly $\#\left(\bigcup_{k=0}^{n-1} f^{-k}(H)\right) - 1$ intervals. $\square$

Corollary 1. *The word-complexity of a Sturmian shift is $p(n) = n + 1$.*

Proof. The partition of $(\mathbb{S}^1, R_\gamma)$ to be used is $\mathcal{P} = \{[0,\gamma), [\gamma, 0)\}$, so $H = \{0, \gamma\}$ and $R_\gamma^{-1}(H) = \{-\gamma, 0\}$ adds only one point to H. Each next iterate add another point $-n\gamma$ to the set, so $\#\left(\bigcup_{k=0}^{n-1} f^{-k}(H)\right) = n + 1$. This $n+1$ points separate $\mathbb{S}^1$ in exactly $n+1$ intervals, so $p(n) = n + 1$. $\square$

Example 8. *For the map $T_{10}(x) = 10x \pmod 1$, say on $[0,1)$, the natural partition is $\mathcal{P} = \mathcal{P}_1 = \{[0, \frac{1}{10}), [\frac{1}{10}, \frac{2}{10}), \dots, [\frac{9}{10}, 1)\}$. At every iteration step, each interval $P \in \mathcal{P}_{n-1}$ splits into ten equal subintervals, so $\mathcal{P}_n = \{[\frac{a}{10^n}, \frac{a+1}{10^n}) \ : \ a = 0, \dots, 10^n - 1\}$ and $\#\mathcal{P}_n = 10^n$. Therefore the topological entropy $h_{top}(T_{10}) = \lim_n \frac{1}{n} \log 10^n = \log 10$.*

Exercise 11. *Take $\beta = \frac{1+\sqrt{5}}{2}$ the golden mean and consider the β-transformation T_β with this slope. Show that $\#P_n$ is the $n+1$st Fibonacci number, and hence compute the topological entropy.*

Remark: The fact that both in Example 8 and Exercise 11 the topological entropy is the logarithm of the (constant) slope of the map is no coincidence!

3 Invariant Measures

Definition 6. *Given a dynamical system $T : X \to X$, a measure μ is called invariant if $\mu(B) = \mu(T^{-1}(B))$ for every measurable set B. We denote the set of T-invariant measures by $\mathcal{M}(T)$.*

Example 9. *Examples of shift-invariant measures are Bernoulli measures on $\mathcal{A}^{\mathbb{N}_0}$ or $\mathcal{A}^{\mathbb{Z}}$. Dirac measures δ_p are invariant if and only if p is a fixed point. If p is periodic under the shift, say of period n, then $\delta_{orb(p)} = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{\sigma^i(p)}$ is invariant.*

Example 10. *For interval maps such as $T(x) = nx \pmod 1$ (where $n \in \mathbb{Z} \setminus \{0\}$ is a fixed integer, Lebesgue measure is T-invariant. Lebesgue measure is also invariant for circle rotations: $R_\gamma : \mathbb{S}^1 \to \mathbb{S}^1$, $x \mapsto x + \gamma \pmod 1$.*

Theorem 3 (Poincaré's Recurrence Theorem)**.** *If* (X, T, μ) *is a measure preserving system with* $\mu(X) = 1$*, then for every measurable set* $U \subset X$ *of positive measure,* μ*-a.e.* $x \in U$ *returns to* U*,* i.e., *there is* $n = n(x)$ *such that* $T^n(x) \in U$*.*

Naturally, reapplying this theorem shows that μ-a.e. $x \in U$ returns to U infinitely often. Because this result uses very few assumptions, it posed a problem for the perceived "Law of Increasing Entropy" in thermodynamics. If the movement of a gas, say, is to be explained purely mechanical, namely as the combination of many particles moving and bouncing against each other according to Newton's laws of mechanics, and hence preserving energy, then in principle it has an invariant measure. This is Liouville measure[4] on the huge phase space containing the six position and momentum components of every particle in the system. Assuming that we start with a containing of gas in which all molecules are bunch together in a tiny corner of the container. This is a state of low entropy, and we expect the particles to fill the entire container rather evenly, thus hugely increasing the entropy. However, Poincaré Recurrence Theorem predicts that at some time t, the system returns arbitrarily closely to the original system, so again with small entropy. Ergo, entropy cannot increase monotonically throughout all time.

Proof of Theorem 3. Let U be an arbitrary measurable set of positive measure. As μ is invariant, $\mu(T^{-i}(U)) = \mu(U) > 0$ for all $i \geqslant 0$. On the other hand, $1 = \mu(X) \geqslant \mu(\cup_i T^{-i}(U))$, so there must be overlap in the backward iterates of U, *i.e.,* there are $0 \leqslant i < j$ such that $\mu(T^{-i}(U) \cap T^{-j}(U)) > 0$. Take the j-th iterate and find $\mu(T^{j-i}(U) \cap U) \geqslant \mu(T^{-i}(U) \cap T^{-j}(U)) > 0$. This means that a positive measure part of the set U returns to itself after $n := j - i$ iterates.

For the part U' of U that didn't return after n step, assuming this part has positive measure, we repeat the argument. That is, there is n' such that $\mu(T^{n'}(U') \cap U') > 0$ and then also $\mu(T^{n'}(U') \cap U) > 0$.

Repeating this argument, we can exhaust the set U up to a set of measure zero, and this proves the theorem. □

Definition 7. *A measure* μ *is* **ergodic** *if for every set* A *such that the inverse* $T^{-1}(A) = A$ *holds:* $\mu(A) = 0$ *or* $\mu(A^c) = 0$*.*

Ergodic measure cannot be decomposed into "smaller elements", whereas non-ergodic measure are mixtures of ergodic measures. For example, the angle doubling map T_2 of Example 2, taken on the interval $[0, 1]$, has fixed points 0 and 1. The Dirac measures δ_0 and δ_1 are both invariant, and therefore every convex combination $\mu_\lambda = (1-\lambda)\delta_0 + \lambda\delta_1$ are invariant too. But μ_λ is not ergodic for $\lambda \in (0, 1)$, whereas $\delta_0 = \mu_0$ and $\delta_1 = \mu_1$

[4] This is an invariant measure over continuous time $t \in \mathbb{R}$, rather than discrete time as stated in Theorem 3, but this doesn't matter for the argument

are. Lebesgue measure is another ergodic invariant measure for T_2, but its ergodicity is more elaborate to prove.

Every invariant measure μ can be decomposed into ergodic components, but in general there are so many ergodic measures, that this decomposition is not finite (as in the example above) but infinite, and it is expressed as an integral. Let $\mathcal{M}_{erg}(T)$ be the collection of ergodic T-invariant measures. Then the **ergodic decomposition** of a (non-ergodic) T-invariant measure μ requires a probability measure τ on the space $\mathcal{M}_{erg}(T)$:

$$\mu(A) = \int_{\mathcal{M}_{erg}} \nu(A)\ d\tau(\nu) \qquad \text{for all measurable subsets } A \subset X. \tag{5}$$

Every invariant measure has such an ergodic decomposition, and because of this, it suffices in many cases to consider only the ergodic invariant measures instead of all invariant measures.

Theorem 4 (Birkhoff's Ergodic Theorem)**.** *If (X, T, μ) is a dynamical system with T-invariant measure μ, and $\psi : X \to \mathbb{R}$ is integrable w.r.t. μ. Then*

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

exists μ-a.e., and the function $\bar{\psi}$ is T-invariant, i.e., *$\bar{\psi}(x) = \bar{\psi} \circ T(x)$.*

If in addition, μ is ergodic, then $\bar{\psi}(x)$ is constant, and $\int \bar{\psi}\ d\mu = \int \psi\ d\mu$. In other words, the **space average** *of ψ over an ergodic invariant measure is the same as the* **time average** *of the ergodic sums of μ-a.e. starting point:*

$$\int_X \psi\ d\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \psi \circ T^i(x) \quad \mu\text{-a.e.} \tag{6}$$

Birkhoff's Ergodic Theory (proved in 1931) is a milestone, yet preceded by a short while by Von Neumann's L^2 Ergodic Theorem, in which the convergence of ergodic averages is in the L^2-norm.

Example 11. *Let $T : [0, 1] \to [0, 1]$ be defined as $x \mapsto 10x \pmod 1$, and write $x_k := T^k(x)$. Then it is easy to see that the integer part of $10x_{k-1}$ is the k-th decimal digit of x.*

Lebesgue measure is T-variant and ergodic[5]*, so Birkhoff's Ergodic Theorem applies as follows: For Lebesgue-a.e. $x \in [0, 1]$, the frequency of decimal digit a is exactly $\int_{a/10}^{(a+1)/10} dx = \frac{1}{10}$. In fact, the frequency in the decimal expansion of Lebesgue-a.e. x of*

[5] ergodicity you will have to believe; we won't prove it here

a block of digits $a_1 \dots a_n$ is exactly 10^{-n}. This property is known in probability theory as **normality**. *Before Birkhoff's Ergodic Theorem, proving normality of Lebesgue-a.e. $x \in [0,1]$ was a lengthy exercise in Probability Theory. With Birkhoff's Ergodic Theorem it is a two-line proof.*

Theorem 5 (Krylov-Bogol'ubov). *If $T : X \to X$ is a continuous map on a nonempty compact metric space X, then $\mathcal{M}(T) \neq \emptyset$.*

Proof. The proof relies on the Let ν be any probability measure and define Cesaro means:

$$\nu_n(A) = \frac{1}{n}\sum_{j=0}^{n-1} \nu(T^j A),$$

these are all probability measures. The collection of probability measures on a compact metric space is known to be compact in the weak topology, *i.e.*, there is limit probability measure μ and a subsequence $(n_i)_{i\in\mathbb{N}}$ such that for every continuous function $\psi : X \to \mathbb{R}$:

$$\int_X \psi \; d\nu_{n_i} \to \int \psi \; d\mu \text{ as } i \to \infty.$$

On a metric space, we can, for any $\varepsilon > 0$ and set A, find a continuous function $\psi_A : X \to [0,1]$ such that $\psi_A(x) = 1$ if $x \in A$ and $\mu(A) \leqslant \int_X \psi_A d\mu \leqslant \mu(A) + \varepsilon$. Now

$$\begin{aligned}
|\mu(T^{-1}(A)) - \mu(A)| &\leqslant \left| \int \psi_A \circ T \; d\mu - \int \psi_A \; d\mu \right| + 2\varepsilon \\
&= \lim_{i\to\infty} \left| \int \psi_A \circ T \; d\nu_{n_i} - \int \psi_A \; d\nu_{n_i} \right| + 2\varepsilon \\
&= \lim_{i\to\infty} \frac{1}{n_i} \left| \sum_{j=0}^{n_i-1} \left(\int \psi_A \circ T^{j+1} \; d\nu - \int \psi_A \circ T^j \; d\nu \right) \right| + 2\varepsilon \\
&\leqslant \lim_{i\to\infty} \frac{1}{n_i} \left| \int \psi_A \circ T^{n_i} \; d\nu - \int \psi_A \; d\nu \right| + 2\varepsilon \\
&\leqslant \lim_{i\to\infty} \frac{1}{n_i} 2\|\psi_A\|_\infty + 2\varepsilon = 2\varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we find that $\mu(T^{-1}(A)) = \mu(A)$ as required. □

Definition 8. *We call (X, f)* **uniquely ergodic** *if there is only one f-invariant probability measure.*

Examples of uniquely ergodic systems are circle rotations R_γ over rotation angle $\gamma \in \mathbb{R} \setminus \mathbb{Q}$. Here, Lebesgue measure is the only invariant measure.

Exercise 12. *Why is $(\mathbb{S}^1, R_\gamma)$ not uniquely ergodic for rational angles γ?*

In general, a system (X, f) can have many invariant measures, and it is worth thinking about what may be useful invariant measures (*e.g.* for the application of Birkhoff's Ergodic Theorem).

Definition 9. *A measure μ is* **absolutely continuous** *w.r.t. ν (notation: $\mu \ll \nu$) if $\nu(A) = 0$ implies $\mu(A) = 0$. If both $\mu \ll \nu$ and $\nu \ll \mu$, then we say that μ and ν are* **equivalent***. If μ is a probability measure and $\mu \ll \nu$ then the Radon-Nikodym Theorem asserts that there is a function $h \in L^1(\nu)$ (called* **Radon-Nikodym derivative** *or* **density***) such that $\mu(A) = \int_A h(x)\ d\nu(x)$ for every measurable set A. Sometimes we use the notation: $h = \frac{d\mu}{d\nu}$.*

The advantage of knowing that an invariant measure μ absolutely continuous w.r.t. a given "reference" measure ν (such as Lebesgue measure), is that instead of μ-a.e. x, we can say that Birkhoff's Ergodic Theorem applies to ν-a.e. x, and ν-a.e. x may be much easier to handle.

Suppose that $T : [0, 1] \to [0, 1]$ is some (piecewise) differentiable interval map. If $\mu \ll Leb$ is an T-invariant measure, then this can be expressed in terms of the density h, namely:

$$h(x) = \sum_{y, T(y)=x} \frac{1}{|T'(y)|} h(y). \tag{7}$$

Example 12. *The Gauss map $G : [0, 1] \to [0, 1]$ (see Figure 1) is defined as*

$$G(x) = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor,$$

where $\lfloor y \rfloor$ denotes rounding down to the nearest integer below y. It is related to continued fractions by the following algorithm with starting point $x \in [0, 1)$. Define

$$x_k = G^k(x), \quad a_k = \lfloor \frac{1}{x_{k-1}} \rfloor,$$

then

$$x = \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots}}} =: [0; a_1, a_2, a_3, \dots]$$

is the standard continued fraction expansion of x. If $x \in \mathbb{Q}$, then this algorithm terminates at some $G^k(x) = 0$, and we cannot iterate G any further. In this case, $x = [0; a_1, a_2, a_3, \dots a_k]$ has a finite continued fraction expansion. For $x \in [0, 1] \setminus \mathbb{Q}$, the continued fraction expansion is infinite.

Gauss discovered (without revealing how) that G has an invariant density

$$h(x) = \frac{1}{\log 2} \frac{1}{1 + x},$$

where $\frac{1}{\log 2}$ *is a normalizing constant, making* $\int_0^1 h(x)\,dx = 1$. *We check* (7). *Note that* $G(y) = x$ *means that* $y = \frac{1}{x+n}$ *for some* $n \in \mathbb{N}$, *and also* $G'(y) = -1/y^2$. *Therefore we can compute:*

$$\begin{aligned}\sum_{G(y)=x} \frac{1}{|G'(y)|} h(y) &= \sum_{n=1}^{\infty} \left| \frac{-1}{(x+n)^2} \right| \frac{1}{\log 2} \frac{1}{1+\frac{1}{x+n}} = \frac{1}{\log 2} \sum_{n=1}^{\infty} \frac{1}{x+n} \frac{1}{x+n+1} \\ &= \frac{1}{\log 2} \sum_{n=1}^{\infty} \frac{1}{x+n} - \frac{1}{x+n+1} = \frac{1}{\log 2} \frac{1}{x+1} = h(x).\end{aligned}$$

Using the Ergodic Theorem, we can estimate the frequency of digits $a_k = N$ *for typical points points* $x \in [0,1]$ *as*

$$\lim_{n\to\infty} \frac{1}{n} \{1 \leqslant k \leqslant n : a_k = N\} = \int_{1/(N+1)}^{1/N} h(x)\,dx = \left[\frac{\log(1+x)}{\log 2} \right]_{1/(N+1)}^{1/N} = \frac{\log(1+\frac{1}{N(N+2)})}{\log 2}.$$

Exercise 13. *Let* $T : [0,1] \to [0,1]$ *be a piecewise affine map such that each branch of* T *is onto* $[0,1]$. *That is, there is a partition* J_k *of* $[0,1]$ *such that* $T|_{J_k}$ *is an affine map so that* $T(J_k) = [0,1]$. *Show that* T *preserves Lebesgue measure.*

Exercise 14. *Let the quadratic Chebyshev polynomial* $T : [0,1] \to [0,1]$ *(see Figure 1) be defined as* $T(x) = 4x(1-x)$. *Verify that the density* $h(x) = \frac{1}{\pi} \frac{1}{\sqrt{x(1-x)}}$ *is* T*-invariant and that* $\int h(x)\,dx = 1$.

Exercise 15. *If* T *is defined on a subset of* d*-dimensional Euclidean space, then* (7) *needs to be replace by*

$$h(x) = \sum_{y, T(y)=x} \underbrace{|\det DT(y)|}_{J(y)}{}^{-1} h(y).$$

Show that the cat-map of Example 4 preserves Lebesgue measure.

4 Entropy

4.1 Measure theoretic entropy

Entropy is a measure for the complexity of a dynamical system (X, T). In the previous sections, we related this (or rather topological entropy) to the exponential growth rate of the cardinality of $\mathcal{P}_n = \bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ for some partition of the space X. In this section, we look at the measure theoretic entropy $h_\mu(T)$ of an T-invariant measure μ, and this amounts to, instead of just counting $\mathcal{P}_n$, taking a particular weighted sum of the elements $Z_n \in \mathcal{P}_n$. However, if the mass of μ is equally distributed over the all the $Z_n \in \mathcal{P}_n$, then the outcome of this sum is largest; then μ would be the measure of maximal entropy. In "good" systems (X,T) is indeed the supremum over the measure theoretic entropies of all the T-invariant probability measures. This is called the **Variational Principle**:

$$h_{top}(T) = \sup\{h_\mu(T) \ : \ \mu \text{ is } T\text{-invariant probability measure}\}. \tag{8}$$

In this section, we will skip some of the more technical aspect, such as **conditional entropy** (however, see Proposition 1) and σ-algebras (completing a set of partitions), and this means that at some points we cannot give full proofs. Rather than presenting more philosophy what entropy should signify, let us first give the mathematical definition.

Define

$$\varphi : [0,1] \to \mathbb{R} \quad \varphi(x) = -x \log x$$

with $\varphi(0) := \lim_{x \downarrow 0} \varphi(x) = 0$. Clearly $\varphi'(x) = -(1+\log x)$ so $\varphi(x)$ assume its maximum at $1/e$ and $\varphi(1/e) = 1/e$. Also $\varphi''(x) = -1/x < 0$, so that φ is **strictly concave**:

$$\alpha\varphi(x) + \beta\varphi(y) \leqslant \varphi(\alpha x + \beta y) \qquad \text{for all } \alpha + \beta = 1, \alpha, \beta \geqslant 0, \tag{9}$$

with equality if and only if $x = y$.

Theorem 6. *For every strictly concave function $\varphi : [0,\infty) \to \mathbb{R}$ we have*

$$\sum_i \alpha_i \varphi(x_i) \leqslant \varphi(\sum_i \alpha_i x_i) \textit{ for } \alpha_i > 0, \sum_i \alpha_i = 1 \textit{ and } x_i \in [0,\infty), \tag{10}$$

with equality if and only if all the x_i are the same.

Proof. We prove this by induction on n. For $n = 2$ it is simply (9). So assume that (10) holds for some n, and we treat the case $n+1$. Assume $\alpha_i > 0$ and $\sum_{i=1}^{n+1} \alpha_i = 1$ and write $B = \sum_{i=1}^n \alpha_i$.

$$\begin{aligned}
\varphi(\sum_{i=1}^{n+1} \alpha_i x_i) &= \varphi(B \sum_{i=1}^{n} \frac{\alpha_i}{B} x_i + \alpha_{n+1} x_{n+1}) \\
&\geqslant B\varphi(\sum_{i=1}^{n} \frac{\alpha_i}{B} x_i) + \varphi(\alpha_{n+1} x_{n+1}) \qquad \text{by (9)} \\
&\geqslant B \sum_{i=1}^{n} \frac{\alpha_i}{B} \varphi(x_i) + \varphi(\alpha_{n+1} x_{n+1}) \qquad \text{by (10) for } n \\
&= \sum_{i=1}^{n+1} \alpha_i \varphi(x_i)
\end{aligned}$$

as required. Equality also carries over by induction, because if x_i are all equal for $1 \leqslant i \leqslant n$, (9) only preserves equality if $x_{n+1} = \sum_{i=1}^n \frac{\alpha_i}{B} x_i = x_1$. $\square$

This proof doesn't use the specific form of φ, only its (strict) concavity. Applying it to $\varphi(x) = -x \log x$, we obtain:

Corollary 2. *For $p_1 + \cdots + p_n = 1$, $p_i > 0$, then $\sum_{i=1}^n \varphi(p_i) \leqslant \log n$ with equality if and only if all p_i are equal,* i.e., $p_i \equiv \frac{1}{n}$.

Proof. Take $\alpha_i = \frac{1}{n}$, then by Theorem 6,

$$\frac{1}{n} \sum_{i=1}^n \varphi(p_i) = \sum_{i=1}^n \alpha_i \varphi(p_i) \leqslant \varphi(\sum_{i=1}^n \frac{1}{n} p_i) = \varphi(\frac{1}{n}) = \frac{1}{n} \log n.$$

Now multiply by n. $\square$

Corollary 3. *For real numbers a_i and $p_1 + \cdots + p_n = 1$, $p_i > 0$, $\sum_{i=1}^n p_i(a_i - \log p_i) \leqslant \log \sum_{i=1}^n e^{a_i}$ with equality if and only if $p_i = e^{a_i} / \sum_{i=1}^n e^{a_i}$ for each i.*

Proof. Write $\mathcal{Z} = \sum_{i=1}^n e^{a_i}$. Put $\alpha_i = e^{a_i}/\mathcal{Z}$ and $x_i = p_i \mathcal{Z}/e^{a_i}$ in Theorem 6. Then

$$\begin{aligned} \sum_{i=1}^n p_i(a_i - \log \mathcal{Z} - \log p_i) &= -\sum_{i=1}^n \frac{e^{a_i}}{\mathcal{Z}} \left(\frac{p_i \mathcal{Z}}{e^{a_i}} \log \frac{p_i \mathcal{Z}}{e^{a_i}} \right) \\ &\leqslant -\sum_{i=1}^n \frac{e^{a_i}}{\mathcal{Z}} \frac{p_i \mathcal{Z}}{e^{a_i}} \log \sum_{i=1}^n \frac{e^{a_i}}{\mathcal{Z}} \frac{p_i \mathcal{Z}}{e^{a_i}} = \varphi(1) = 0. \end{aligned}$$

Rearranging gives $\sum_{i=1}^n p_i(a_i - \log p_i) \leqslant \log Z$, with equality only if $x_i = p_i Z/e^{a_i}$ are all the same, *i.e.*, $p_i = e^{a_i}/Z$. $\square$

Exercise 16. *Reprove Corollaries 2 and 3 using Lagrange multipliers.*

Given a finite partition $\mathcal{P}$ of a probability space (X, μ), let

$$H_\mu(\mathcal{P}) = \sum_{P \in \mathcal{P}} \varphi(\mu(P)) = -\sum_{P \in \mathcal{P}} \mu(P) \log(\mu(P)), \tag{11}$$

where we can ignore the partition elements with $\mu(P) = 0$ because $\varphi(0) = 0$. For a T-invariant probability measure μ on $(X, \mathcal{B}, T)$, and a partition $\mathcal{P}$, define the **entropy of μ w.r.t. $\mathcal{P}$** as

$$H_\mu(T, \mathcal{P}) = \lim_{n \to \infty} \frac{1}{n} H_\mu(\bigvee_{k=0}^{n-1} T^{-k} \mathcal{P}). \tag{12}$$

Finally, the **measure theoretic entropy** of μ is

$$h_\mu(T) = \sup\{H_\mu(T, \mathcal{P}) \;:\; \mathcal{P} \text{ is a finite partition of } X\}. \tag{13}$$

Naturally, this raises the questions:

Does the limit exist in (12)?

How can one possibly consider **all** partitions of X?

We come to this later; first we want to argue that entropy is a characteristic of a measure preserving system. That is, two measure preserving systems $(X, \mathcal{B}, T, \mu)$ and $(Y, \mathcal{C}, S, \nu)$ that are **isomorphic**, *i.e.,* there is a bi-measurable invertible measure-preserving map π (called **isomorphism**) such that the diagram

$$\begin{array}{ccc} (X, \mathcal{B}, \mu) & \xrightarrow{T} & (X, \mathcal{B}, \mu) \\ \pi \downarrow & & \downarrow \pi \\ (Y, \mathcal{C}, \nu) & \xrightarrow{S} & (Y, \mathcal{C}, \nu) \end{array}$$

commutes, then $h_\mu(T) = h_\nu(S)$. This holds, because the bi-measurable measure-preserving map π preserves all the quantities involved in (11)-(13), including the class of partitions for both systems.

A major class of systems where this is very important are the Bernoulli shifts. These are the standard probability space to measure a sequence of i.i.d. events each with outcomes in $\{0, \dots, N-1\}$ with probabilities $p_0, \dots, p_{N-1}$ respectively. That is: $X = \{0, \dots, N-1\}^{\mathbb{N}_0}$ or $\{0, \dots, N-1\}^{\mathbb{Z}}$, σ is the left-shift, and μ the Bernoulli measure that assigns to every cylinder set $[x_m \dots x_n]$ the mass

$$\mu([x_m \dots x_n]) = \prod_{k=m}^{n} \rho(x_k) \qquad \text{where } \rho(x_k) = p_i \text{ if } x_k = i.$$

For such a Bernoulli shift, the entropy is

$$h_\mu(\sigma) = -\sum_i p_i \log p_i, \tag{14}$$

so two Bernoulli shifts (X, p, μ_p) and $(X', p', \mu_{p'})$ can only be isomorphic if $-\sum_i p_i \log p_i = -\sum_i p'_i \log(p'_i)$. The famous theorem of Ornstein showed that entropy is a complete invariant for Bernoulli shifts:

Theorem 7 (Ornstein 1974 [7], cf. page 105 of [9])**.** *Two Bernoulli shifts (X, p, μ_p) and $(X', p', \mu_{p'})$ are isomorphic if and only if $-\sum_i p_i \log p_i = -\sum_i p'_i \log p'_i$.*

Exercise 17. *Conclude that the Bernoulli shift $\mu_{(\frac14,\frac14,\frac14,\frac14)}$ is isomorphic to $\mu_{(\frac18,\frac18,\frac18,\frac18,\frac12)}$, but that no Bernoulli measure on four symbols can be isomorhic to $\mu_{(\frac15,\frac15,\frac15,\frac15,\frac15)}$*

Let us go back to the definition of entropy, and try to answer the outstanding questions.

Definition 10. *We call a real sequence $(a_n)_{n\geqslant 1}$* **subadditive** *if*

$$a_{m+n} \leqslant a_m + a_n \quad \text{for all } m, n \in \mathbb{N}.$$

Theorem 8. *If $(a_n)_{n\geqslant 1}$ is subadditive, then* $\lim_n \frac{a_n}{n} = \inf_{r\geqslant 1} \frac{a_r}{r}$.

Proof. Every integer n can be written uniquely as $n = i \cdot r + j$ for $0 \leqslant j < r$. Therefore

$$\limsup_{n\to\infty} \frac{a_n}{n} = \limsup_{i\to\infty} \frac{a_{i\cdot r+j}}{i\cdot r+j} \leqslant \limsup_{i\to\infty} \frac{ia_r + a_j}{i\cdot r+j} = \frac{a_r}{r}.$$

This holds for all $r \in \mathbb{N}$, so we obtain

$$\inf_r \frac{a_r}{r} \leqslant \liminf_n \frac{a_n}{n} \leqslant \limsup_n \frac{a_n}{n} \leqslant \inf_r \frac{a_r}{r},$$

as required. □

Definition 11. *Motivated by the conditional measure* $\mu(P|Q) = \frac{\mu(P\cap Q)}{\mu(Q)}$, *we define* **conditional entropy** *of a measure μ as*

$$H_\mu(\mathcal{P}|\mathcal{Q}) = -\sum_j \mu(Q_j) \sum_i \frac{\mu(P_i \cap Q_j)}{\mu(Q_j)} \log \frac{\mu(P_i \cap Q_j)}{\mu(Q_j)}, \tag{15}$$

where i runs over all elements $P_i \in \mathcal{P}$ and j runs over all elements $Q_j \in \mathcal{Q}$.

Avoiding philosophical discussions how to interpret this notion, we just list the main properties that are needed in this course that rely of condition entropy:

Proposition 1. *Given measures μ, μ_i and two partitions $\mathcal{P}$ and $\mathcal{Q}$, we have*

1. $H_\mu(\mathcal{P} \vee \mathcal{Q}) \leqslant H_\mu(\mathcal{P}) + H_\mu(\mathcal{Q})$;
2. $H_\mu(T, \mathcal{P}) \leqslant H_\mu(T, \mathcal{Q}) + H_\mu(\mathcal{P} \mid \mathcal{Q})$.
3. $\sum_{i=1}^n p_i H_{\mu_i}(\mathcal{P}) \leqslant H_{\sum_{i=1}^n p_i\mu_i}(\mathcal{P})$ *whenever* $\sum_{i=1}^n p_1 = 1$, $p_i \geqslant 0$,

Subadditivity is the key to the convergence in (12). Call $a_n = H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$. Then

$$\begin{aligned}
a_{m+n} &= H_\mu(\bigvee_{k=0}^{m+n-1} T^{-k}\mathcal{P}) && \text{use Proposition 1, part 1.}\\
&\leqslant H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{P}) + H_\mu(\bigvee_{k=m}^{m+n-1} T^{-k}\mathcal{P}) && \text{use } T\text{-invariance of } \mu\\
&= H_\mu(\bigvee_{k=0}^{m-1} T^{-k}\mathcal{P}) + H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})\\
&= a_m + a_n.
\end{aligned}$$

Therefore $H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$ is subadditive, and the existence of the limit of $\frac{1}{n}H_\mu(\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P})$ follows.

Proposition 2. *Entropy has the following properties:*

1. *The identity map has entropy* 0*;*
2. $h_\mu(T^R) = R \cdot h_\mu(T)$ *and for invertible systems* $h_\mu(T^{-R}) = R \cdot h_\mu(T)$*.*

Proof. Statement 1. follows simply because $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P} = \mathcal{P}$ if T is the identity map, so the cardinality of $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ doesn't increase with n.

For statement 2. set $\mathcal{Q} = \bigvee_{j=0}^{R-1} T^{-j}\mathcal{P}$. Then for $k \geqslant 1$,

$$\begin{aligned} R \cdot H_\mu(T, \mathcal{P}) &= \lim_{n\to\infty} R \cdot \frac{1}{nR} H_\mu(\bigvee_{j=0}^{nR-1} T^{-k}\mathcal{P}) \\ &= \lim_{n\to\infty} \frac{1}{n} H_\mu(\bigvee_{j=0}^{n-1} (T^R)^{-j}\mathcal{Q}) \\ &= H_\mu(T^R, \mathcal{Q}). \end{aligned}$$

Taking the supremum over all $\mathcal{P}$ or $\mathcal{Q}$ has the same effect. □

The next theorem is the key to really computing entropy, as it shows that a single well-chosen partition $\mathcal{P}$ suffices to compute the entropy as $h_\mu(T) = H_\mu(T, \mathcal{P})$.

Theorem 9. *Let* $(X, \mathcal{B}, T, \mu)$ *be a measure-preserving dynamical system. If partition* $\mathcal{P}$ *is such that*

$$\begin{cases} \bigvee_{j=0}^{\infty} T^{-k}\mathcal{P} \text{ generates } \mathcal{B} & \text{if } T \text{ is non-invertible,} \\ \bigvee_{j=-\infty}^{\infty} T^{-k}\mathcal{P} \text{ generates } \mathcal{B} & \text{if } T \text{ is invertible,} \end{cases}$$

then $h_\mu(T) = H_\mu(T, \mathcal{P})$*.*

We haven't explained properly what "generates $\mathcal{B}$ means, but the idea you should have in mind is that (up to measure 0), every two points in X should be in different elements of $\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ (if T is non-invertible), or of $\bigvee_{k=-n}^{n-1} T^{-k}\mathcal{P}$ (if T is invertible) for some sufficiently large n. The partition $\mathcal{B} = \{X\}$ fails miserably here, because $\bigvee_{j=-n}^{n} T^{-k}\mathcal{P} = \mathcal{P}$ for all n and no two points are ever separated in $\mathcal{P}$. A more subtle example can be created for the doubling map $T_2 : \mathbb{S}^1 \to \mathbb{S}^1$, $T_2(x) = 2x \pmod 1$. The partition $\mathcal{P} = \{[0, \frac{1}{2}), [\frac{1}{2}, 1)\}$. is separating every two points, because if $x \neq y$, say $2^{-(n+1)} < |x - y| \leqslant 2^{-n}$, then there is $k \leqslant n$ such that $T_2^k x$ and $T_2^k y$ belong to different partition elements.

On the other hand, $\mathcal{Q} = \{[\frac{1}{4}, \frac{3}{4}), [0, \frac{1}{4}) \cup [\frac{3}{4}, 1)\}$ does **not** separate points. Indeed, if $y = 1 - x$, then $T_2^k(y) = 1 - T_2^k(x)$ for all $k \geqslant 0$, so x and y belong to the same partition element, $T_2^k(y)$ and $T_2^k(x)$ will also belong to the same partition element!

In this case, $\mathcal{P}$ can be used to compute $h_\mu(T)$, while $\mathcal{Q}$ in principle cannot (although here, for all Bernoulli measure $\mu = \mu_{p,1-p}$, we have $h_\mu(T_2) = H_\mu(T, \mathcal{P}) = H_\mu(T, \mathcal{Q})$).

We finish this section with computing the entropy for a Bernoulli shift on two symbols, *i.e.,* we will prove (14) for two-letter alphabets and any probability $\mu([0]) =: p \in [0,1]$. The space is thus $X = \{0,1\}^{\mathbb{N}_0}$ and each $x \in X$ represents an infinite sequence of coin-flips with an unfair coin that gives head probability p (if head has the symbol 0). Recall from probability theory

$$\mathbb{P}(k \text{ heads in } n \text{ flips}) = \binom{n}{k} p^k (1-p)^{n-k},$$

so by full probability:

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Here $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ are the binomial coefficients, and we can compute

$$\begin{cases} k\binom{n}{k} = \frac{n!}{(k-1)!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = n\binom{n-1}{k-1} \\ (n-k)\binom{n}{k} = \frac{n!}{(k)!(n-k-1)!} = n\frac{(n-1)!}{k!(n-k-1)!} = n\binom{n-1}{k} \end{cases} \tag{16}$$

This gives all the ingredients necessary for the computation.

$$\begin{aligned} H_\mu(\bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}) &= -\sum_{x_0,\dots,x_{n-1}=0}^{1} \mu([x_0,\dots,x_{n-1}]) \log \mu([x_0,\dots,x_{n-1}]) \\ &= -\sum_{x_0,\dots,x_{n-1}=0}^{1} \prod_{j=0}^{n-1} \rho(x_j) \log \prod_{j=0}^{n-1} \rho(x_j) \\ &= -\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \log \left(p^k (1-p)^{n-k}\right) \\ &= -\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} k \log p \\ &\qquad -\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (n-k) \log(1-p) \end{aligned}$$

In the first sum, the term $k=0$ gives zero, as does the term $k=n$ for the second sum.

Thus we leave out these terms and rearrange by (16):

$$\begin{aligned}
&= -p\log p \sum_{k=1}^{n} k \binom{n-1}{k} p^{k-1}(1-p)^{n-k} \\
&\qquad -(1-p)\log(1-p)\sum_{k=0}^{n-1}(n-k)\binom{n}{k} p^{k}(1-p)^{n-k-1} \\
&= -p\log p \sum_{k=1}^{n} n \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \\
&\qquad -(1-p)\log(1-p)\sum_{k=0}^{n-1} n\binom{n-1}{k} p^{k}(1-p)^{n-k-1} \\
&= n\left(-p\log p-(1-p)\log(1-p)\right).
\end{aligned}$$

The partition $\mathcal{P} = \{[0], [1]\}$ is generating, so by Theorem 9,

$$h_\mu(\sigma) = H_\mu(\sigma, \mathcal{P}) = \lim_n \frac{1}{n} H_\mu(\bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}) = -p\log p - (1-p)\log(1-p)$$

as required.

4.2 Topological entropy and the topological pressure

Topological entropy was first defined in 1965 by Adler et al. [1], but the form that Bowen [4] redressed it in is commonly used nowadays, and Bowen's approach readily generalises to topological pressure as well.

Let T be map of a compact metric space (X, d). If my eyesight is not so good, I cannot distinguish two points $x, y \in X$ if they are at a distance $d(x, y) < \varepsilon$ from one another. I may still be able to distinguish there orbits, if $d(T^k x, T^k y) > \varepsilon$ for some $k \geqslant 0$. Hence, if I'm willing to wait n iterations, I can distinguish x and y if

$$d_n(x, y) := \min\{d(T^k x, T^k y) : 0 \leqslant k < n\} > \varepsilon.$$

If this holds, then x and y are said to be (n, ε)**-separated**. Among all the subsets of X of which all points are mutually (n, ε)-separated, choose one, say $E_n(\varepsilon)$, of maximal cardinality. Then $s_n(\varepsilon) := \#E_n(\varepsilon)$ is the maximal number of n-orbits I can distinguish with my ε-poor eyesight.

The **topological entropy** is defined as the limit (as $\varepsilon \to 0$) of the exponential growth-rate of $s_n(\varepsilon)$:

$$h_{top}(T) = \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log s_n(\varepsilon).$$

Note that $s_n(\varepsilon_1) \geqslant s_n(\varepsilon_2)$ if $\varepsilon_1 \leqslant \varepsilon_2$, so $\limsup_n \frac{1}{n} \log s_n(\varepsilon)$ is a decreasing function in ε, and the limit as $\varepsilon \to 0$ indeed exists.

Instead of (n,ε)-separated sets, we can also work with (n,ε)**-spanning** sets, that is, sets that contain, for every $x \in X$, a y such that $d_n(x,y) \leqslant \varepsilon$. Note that, due to its maximality, $E_n(\varepsilon)$ is always (n,ε)-spanning, and no proper subset of $E_n(\varepsilon)$ is (n,ε)-spanning. Each $y \in E_n(\varepsilon)$ must have a point of an $(n,\varepsilon/2)$-spanning set within an $\varepsilon/2$-ball (in d_n-metric) around it, and by the triangle inequality, this $\varepsilon/2$-ball is disjoint from $\varepsilon/2$-ball centred around all other points in $E_n(\varepsilon)$. Therefore, if $r_n(\varepsilon)$ denotes the minimal cardinality among all (n,ε)-spanning sets, then

$$r_n(\varepsilon) \leqslant s_n(\varepsilon) \leqslant r_n(\varepsilon/2).$$

Thus we can equally well define

$$h_{top}(T) = \lim_{\varepsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log r_n(\varepsilon).$$

Examples: Consider the β-transformation $T_\beta : [0,) \to [0,1)$, $x \mapsto \beta x \pmod 1$ for some $\beta > 1$. Take $\varepsilon < 1/(2\beta^2)$, and $G_n = \{\frac{k}{\beta^n} \ : \ 0 \leqslant k < \beta^n\}$. Then G_n is (n,ε)-separating, so $s_n(\varepsilon) \geqslant \beta^n$. On the other hand, $G_n' = \{\frac{2k\varepsilon}{\beta^n} \ : \ 0 \leqslant k < \beta^n/(2\varepsilon)\}$ is (n,ε)-spanning, so $r_n(\varepsilon) \leqslant \beta^n/(2\varepsilon)$. Therefore

$$\log\beta = \limsup_n \frac{1}{n} \log \beta^n \leqslant h_{top}(T_\beta) \leqslant \limsup_n \log \beta^n/(2\varepsilon) = \log\beta.$$

Circle rotations, or in general isometries, T have zero topological entropy. Indeed, if $E(\varepsilon)$ is an ε-separated set (or ε-spanning set), it will also be (n,ε)-separated (or (n,ε)-spanning) for every $n \geqslant 1$. Hence $s_n(\varepsilon)$ and $r_n(\varepsilon)$ are bounded in n, and their exponential growth rates are equal to zero.

Finally, let (X,σ) be the full shifts on N symbols. Let $\varepsilon > 0$ be arbitrary, and take m such that $2^{-m} < \varepsilon$. If we select a point from each $n+m$-cylinder, this gives an (n,ε)-spanning set, whereas selecting a point from each n-cylinder gives an (n,ε)-separated set. Therefore

$$\begin{aligned}
\log N = \limsup_n \frac{1}{n} \log N^n &\leqslant \limsup_n \frac{1}{n} \log s_n(\varepsilon) \leqslant h_{top}(T_\beta) \\
&\leqslant \limsup_n \frac{1}{n} \log r_n(\varepsilon) \leqslant \limsup_n \log N^{n+m} = \log N.
\end{aligned}$$

The topological pressure $P_{top}(T,\psi)$ combines entropy with a potential function $\psi : X \to \mathbb{R}$. Its definition is so much analogous to topological entropy that we immediately get $h_{top}(T) = P_{top}(T,\psi)$ if $\psi(x) \equiv 0$. Denote the n-th ergodic sum of ψ by

$$S_n\psi(x) = \sum_{k=0}^{n-1} \psi \circ T^k(x).$$

Next set

$$\begin{cases} P_n(T,\psi,\varepsilon) = \sup\{\sum_{x\in E} e^{S_n\psi(x)} \ : \ E \text{ is } (n,\varepsilon)\text{-separated}\}, \\ Q_n(T,\psi,\varepsilon) = \inf\{\sum_{x\in E} e^{S_n\psi(x)} \ : \ E \text{ is } (n,\varepsilon)\text{-spanning}\}. \end{cases} \tag{17}$$

For reasonable choices of potentials, the quantities $\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log P_n(T,\psi,\varepsilon)$ and $\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log Q_n(T,\psi,\varepsilon)$ are the same, and this quantity is called the **topological pressure**. To give an example of an unreasonable potential, take X_0 be a dense subset of X such that $X\setminus X_0$ is also dense. Let

$$\psi(x) = \begin{cases} 100 & \text{if } x\in X_0, \\ 0 & \text{if } x\notin X_0. \end{cases}$$

Then $Q_n(T,\psi,\varepsilon) = r_n(\varepsilon)$ whilst $P_n(T,\psi,\varepsilon) = e^{100n}s_n(\varepsilon)$, and their exponential growth rates differ by a factor 100. Hence, some amount of continuity of ψ is necessary to make it work.

Lemma 3. *If $\varepsilon > 0$ is such that $d(x,y) < \varepsilon$ implies that $|\psi(x)-\psi(y)| < \delta/2$, then*

$$e^{-n\delta}P_n(T,\psi,\varepsilon) \leqslant Q_n(T,\psi,\varepsilon/2) \leqslant P_n(T,\psi,\varepsilon/2).$$

Exercise 18. *Prove Lemma 3. In fact, the second inequality holds regardless of what ψ is.*

Theorem 10. *If $T : X\to X$ and $\psi : X\to\mathbb{R}$ are continuous on a compact metric space, then the topological pressure is well-defined by*

$$P_{top}(T,\psi) := \lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log P_n(T,\psi,\varepsilon) = \lim_{\varepsilon\to 0}\limsup_{n\to\infty}\frac{1}{n}\log Q_n(T,\psi,\varepsilon).$$

Exercise 19. *Show that $P_{top}(T^R, S_R\psi) = R\cdot P_{top}(T,\psi)$.*

4.3 The Variational Principle

The Variational Principle as mentioned in (3) and (8) claims that topological entropy (or pressure) is achieved by taking the supremum of the measure-theoretic entropies over all invariant probability measures. But in the course of these notes, topological entropy has seen various definitions. Even $\sup\{h_\mu(T) : \mu \text{ is a } T\text{-invariant probability measure}\}$ is sometimes used as definition of topological entropy. So it is high time to be more definite.

We will do this by immediately passing to topological pressure, which we will base on the definition in terms of (n,δ)-spanning sets and/or (n,ε)-separated sets. Topological entropy then simply emerges as $h_{top}(T) = P_{top}(T,0)$.

Theorem 11 (The Variational Principle)**.** *Let (X,d) be a compact metric space, $T : X \to X$ a continuous map and $\psi : X \to \mathbb{R}$ as continuous potential. Then*

$$P_{top}(T,\psi) = \sup\{h_\mu(T) + \int_X \psi \, d\mu \ : \ \mu \text{ is a } T\text{-invariant probability measure}\}. \tag{18}$$

Remark 1. *By the ergodic decomposition, every T-invariant probability measure can be written as convex combination (sometimes in the form of an integral) of ergodic T-invariant probability measures. Therefore, it suffices to take the supremum over all ergodic T-invariant probability measures in* (18).

Proof. First we show that for every T-invariant probability measure, $h_\mu(T)+\int_X \psi \, d\mu \leqslant P_{top}(T,\psi)$. Let $\mathcal{P} = \{P_0, \dots, P_{N-1}\}$ be an arbitrary partition with $N \geqslant 2$ (if $\mathcal{P} = \{X\}$, then $h_\mu(T,\mathcal{P}) = 0$ and there is not much to prove). Let $\eta > 0$ be arbitrary, and choose $\varepsilon > 0$ so that $\varepsilon N \log N < \eta$.

By "regularity of μ", there are compact sets $Q_i \subset P_i$ such that $\mu(P_i \setminus Q_i) < \varepsilon$ for each $0 \leqslant i < N$. Take $Q_N = X \setminus \cup_{i=0}^{N-1} Q_i$. Then $\mathcal{Q} = \{Q_0, \dots, Q_N\}$ is a new partition of X, with $\mu(Q_N) \leqslant N\varepsilon$. Furthermore

$$\frac{\mu(P_i \cap Q_j)}{\mu(Q_j)} = \begin{cases} 0 & \text{if } i \neq j < N, \\ 1 & \text{if } i = j < N. \end{cases}$$

whereas $\sum_{i=0}^{N-1} \frac{\mu(P_i \cap Q_N)}{\mu(Q_N)} = 1$. Therefore the conditional entropy

$$\begin{aligned} H_\mu(\mathcal{P}|\mathcal{Q}) &= \sum_{j=0}^{N} \sum_{i=0}^{N-1} \mu(Q_j) \underbrace{\varphi\left(\frac{\mu(P_i \cap Q_j)}{\mu(Q_j)}\right)}_{= \ 0 \text{ if } j<N} \\ &= -\mu(Q_N) \sum_{i=0}^{N-1} \frac{\mu(P_i \cap Q_N)}{\mu(Q_N)} \log(\frac{\mu(P_i \cap Q_N)}{\mu(Q_N)}) \\ &\leqslant \mu(Q_N) \log N \qquad \text{by Corollary 2} \\ &\leqslant \varepsilon N \log N < \eta. \end{aligned}$$

Choose $0 < \delta < \frac{1}{2} \min_{0 \leqslant i < j < N} d(Q_i, Q_j)$ so that

$$d(x,y) < \delta \text{ implies } |\psi(x) - \psi(y)| < \varepsilon/2. \tag{19}$$

Here we use uniform continuity of ψ on the compact space X. Fix n and let $E_n(\delta)$ be an (n,δ)-spanning set. For $Z \in \mathcal{Q}_n := \bigvee_{k=0}^{n-1} T^{-k}\mathcal{Q}$, let $\alpha(Z) = \sup\{S_n\psi(x) \ : \ x \in Z\}$. For each such Z, also choose $x_Z \in \overline{Z}$ such that $S_n\psi(x) = \alpha(Z)$ (again we use continuity of ψ here), and $y_Z \in E_n(\delta)$ such that $d_n(x_Z, y_Z) < \delta$. Hence

$$\alpha(Z) - n\varepsilon \leqslant S_n\psi(y_Z) \leqslant \alpha(Z).$$

This gives

$$H_\mu(\mathcal{Q}_n) + \int_X S_n\psi \, d\mu \leqslant \sum_{Z\in\mathcal{Q}_n} \mu(Z)(\alpha(Z) - \log\mu(Z)) \leqslant \log \sum_{Z\in\mathcal{Q}_n} e^{\alpha(Z)} \tag{20}$$

by Corollary 3.

Each δ-ball intersects the closure of at most two elements of $\mathcal{Q}$. Hence, for each y, the cardinality $\#\{Z \in \mathcal{Q}_n : y_Z = y\} \leqslant 2^n$. Therefore

$$\sum_{Z\in\mathcal{Q}_n} e^{\alpha(Z)-n\varepsilon} \leqslant \sum_{Z\in\mathcal{Q}_n} e^{S_n\psi(y_Z)} \leqslant 2^n \sum_{y\in E_n(\delta)} e^{S_n\psi(y)}.$$

Take the logarithm and rearrange to

$$\log \sum_{Z\in\mathcal{Q}_n} e^{\alpha(Z)} \leqslant n(\varepsilon + \log 2) + \log \sum_{y\in E_n(\delta)} e^{S_n\varphi(y)}.$$

By T-invariance of μ we have $\int S_n\psi \, d\mu = n\int \psi \, d\mu$. Therefore

$$\begin{aligned}
\frac{1}{n}H_\mu(\mathcal{Q}_n) + \int_X \psi \, d\mu &\leqslant \frac{1}{n}H_\mu(\mathcal{Q}_n) + \frac{1}{n}\int_X S_n\psi \, d\mu \\
&\leqslant \frac{1}{n}\log \sum_{Z\in\mathcal{Q}_n} e^{\alpha(Z)} \\
&\leqslant \varepsilon + \log 2 + \frac{1}{n}\log \sum_{y\in E_n(\delta)} e^{S_n\varphi(y)}.
\end{aligned}$$

Taking the limit $n \to \infty$ gives

$$H_\mu(T, \mathcal{Q}) + \int_X \psi \, d\mu \leqslant \varepsilon + \log 2 + P_{top}(T, \psi).$$

By Proposition 1, part 2., and recalling that $\varepsilon < \eta$, we get

$$H_\mu(T, \mathcal{P}) + \int_X \psi \, d\mu = H_\mu(T, \mathcal{Q}) + H_\mu(\mathcal{P}|\mathcal{Q}) + \int_X \psi \, d\mu \leqslant 2\eta + \log 2 + P_{top}(T, \psi).$$

We can apply the same reasoning to T^R and $S_R\psi$ instead of T and ψ. This gives

$$\begin{aligned}
R \cdot \left(H_\mu(T, \mathcal{P}) + \int_X \psi \, d\mu \right) &= H_\mu(T^R, \mathcal{P}) + \int_X S_R\psi \, d\mu \\
&\leqslant 2\eta + \log 2 + P_{top}(T^R, S_R\psi) \\
&= 2\eta + \log 2 + R \cdot P_{top}(T, \psi).
\end{aligned}$$

Divide by R and take $R \to \infty$ to find $H_\mu(T, \mathcal{P}) + \int_X \psi \, d\mu \leqslant P_{top}(T, \psi)$. Finally take the supremum over all partitions $\mathcal{P}$.

Now the other direction, we will work with (n, ε)-separated sets. After choosing $\varepsilon > 0$ arbitrary, we need to find a T-invariant probability measure μ such that

$$h_\mu(T) + \int_X \psi \, d\mu \geqslant \limsup_{n\to\infty} \frac{1}{n} \log P_n(T, \psi, \varepsilon) := P(T, \psi, \varepsilon).$$

Let $E_n(\varepsilon)$ be an (n, ε)-separated set such that

$$\log \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)} \geqslant \log P_n(T, \psi, \varepsilon) - 1. \tag{21}$$

Define Δ_n as weighted sum of Dirac measures:

$$\Delta_n = \frac{1}{\mathcal{Z}} \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)} \delta_y,$$

where $\mathcal{Z} = \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)}$ is the normalising constant. Take a new probability measure

$$\mu_n = \frac{1}{n} \sum_{k=0}^{n-1} \Delta_n \circ T^{-k}.$$

Therefore

$$\begin{aligned}
\int_X \psi \, d\mu_n &= \frac{1}{n} \sum_{k=0}^{n-1} \int_X \psi \, d(\Delta_n \circ T^{-k}) = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{y \in E_n(\varepsilon)} \psi \circ T^k(y) \frac{1}{\mathcal{Z}} e^{S_n\psi(y)} \\
&= \frac{1}{n} \sum_{y \in E_n(\varepsilon)} S_n\psi(y) \frac{1}{\mathcal{Z}} e^{S_n\psi(y)} = \frac{1}{n} \int_X S_n\psi \, d\Delta_n.
\end{aligned} \tag{22}$$

Since the space of probability measures on X is compact in the weak topology, we can find a sequence $(n_j)_{j\geqslant 1}$ such that for every continuous function $f : X \to \mathbb{R}$

$$\int_X f \, d\mu_{n_j} \to \int_X f \, d\mu \quad \text{as } j \to \infty.$$

Choose a partition $\mathcal{P} = \{P_0, \dots, P_{N-1}\}$ with $\operatorname{diam}(P_i) < \varepsilon$ and $\mu(\partial P_i) = 0$ for all $0 \leqslant i < N$. Since $Z \in \mathcal{P}_n := \bigvee_{k=0}^{n-1} T^{-k}\mathcal{P}$ contains at most one element of an (n, ε)-separated set, we have

$$\begin{aligned}
H_{\Delta_n}(\mathcal{P}_n) + \int_X S_n\psi \, d\Delta_n &= \sum_{y \in E_n(\varepsilon)} \Delta_n(\{y\}) \left(S_n\psi(y) - \log \Delta_n(\{y\})\right) \\
&= \log \sum_{y \in E_n(\varepsilon)} e^{S_n\psi(y)} = \log \mathcal{Z}.
\end{aligned}$$

by Corollary 3

Take $0 < q < n$ arbitrary, and for $0 \leqslant j < q$, let

$$U_j = \{j, j+1, \ldots, a_j q + j - 1\} \qquad \text{where } a_j = \lfloor \frac{n-j}{q} \rfloor.$$

Then

$$\{0, 1, \ldots, n-1\} = U_j \cup \underbrace{\{0, 1, \ldots, j-1\} \cup a_j q + j, a_j q + j + 1, \ldots, n-1\}}_{V_j}$$

where V_j has at most $2q$ elements. We split

$$\begin{aligned}
\bigvee_{k=0}^{n-1} T^{-k}\mathcal{P} &= \left(\bigvee_{r=0}^{a_j-1} \bigvee_{i=0}^{q-1} T^{-(rq+j+i)}\mathcal{P} \right) \vee \bigvee_{l \in V_j} T^{-l}\mathcal{P} \\
&= \left(\bigvee_{r=0}^{a_j-1} T^{-(rq+j)} \bigvee_{i=0}^{q-1} T^{-i}\mathcal{P} \right) \vee \bigvee_{l \in V_j} T^{-l}\mathcal{P}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\log \mathcal{Z} &= H_{\Delta_n}(\mathcal{P}_n) + \int_X S_n\psi \; d\Delta_n \\
&\leqslant \sum_{r=0}^{a_j-1} H_{\Delta_n}(T^{-(rq+j)} \bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + H_{\Delta_n}(\bigvee_{l \in V_j} T^{-l}\mathcal{P}) + \int_X S_n\psi \; d\Delta_n \\
&\leqslant \sum_{r=0}^{a_j-1} H_{\Delta_n \circ T^{-(rq+j)}}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q \log N + \int_X S_n\psi \; d\Delta_n,
\end{aligned}$$

because $\bigvee_{l \in V_j} T^{-l}\mathcal{P}$ has at most N^{2q} elements and using Corollary 2. Summing the above inequality over $j = 0, \ldots, q-1$, gives

$$\begin{aligned}
q \log \mathcal{Z} &= \sum_{j=0}^{q-1} \sum_{r=0}^{a_j-1} H_{\Delta_n \circ T^{-rq+j}}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q^2 \log N + q \int_X S_n\psi \; d\Delta_n \\
&\leqslant n \sum_{k=0}^{n-1} \frac{1}{n} \, H_{\Delta_n \circ T^{-k}}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q^2 \log N + q \int_X S_n\psi \; d\Delta_n.
\end{aligned}$$

Proposition 1, part 3., allows us to swap the weighted average and the operation H:

$$q \log \mathcal{Z} \leqslant n H_{\mu_n}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + 2q^2 \log N + q \int_X S_n\psi \; d\Delta_n.$$

Dividing by n and recalling (21) for the left hand side, and (22) to replace Δ_n by μ_n, we find

$$\frac{q}{n}\log P_n(T,\psi,\varepsilon) - \frac{q}{n} \leqslant H_{\mu_n}(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + \frac{2q^2}{n}\log N + q\int_X \psi\ d\mu_n.$$

Because $\mu(\partial P_i) = 0$ for all i, we can replace n by n_j and take the weak limit as $j \to \infty$. This gives

$$qP(T,\psi,\varepsilon) \leqslant H_\mu(\bigvee_{i=0}^{q-1} T^{-i}\mathcal{P}) + q\int_X \psi\ d\mu.$$

Finally divide by q and let $q \to \infty$:

$$P(T,\psi,\varepsilon) \leqslant h_\mu(T) + \int_X \psi\ d\mu.$$

This concludes the proof. $\square$

4.4 Measures of maximal entropy

For the full shift (Ω,σ) with $\Omega = \{0,\dots,N-1\}^{\mathbb{N}_0}$ or $\Omega = \{0,\dots,N-1\}^{\mathbb{Z}}$, we have $h_{top}(\sigma) = \log N$, and the $(\frac{1}{N},\dots,\frac{1}{N})$-Bernoulli measure μ indeed achieves this maximum: $h_\mu(\sigma) = h_{top}(\sigma)$. Hence μ is a (and in this case unique) **measure of maximal entropy**. The intuition to have here is that for a measure to achieve maximal entropy, it should distribute its mass as evenly over the space as possible. But how does this work for subshifts, where it is not immediately obvious how to distribute mass evenly?

For subshifts of finite type, Parry [8] demonstrated how to construct the measure of maximal entropy, which is now called after him. Let (Σ_A,σ) be a subshift of finite type on alphabet $\{0,\dots,N-1\}$ with transition matrix $A = (a_{i,j})_{i,j=0}^{N-1}$, so $x = (x_n) \in \Sigma_n$ if and only if $a_{x_n,x_{n+1}} = 1$ for all n. Let us assume that A is aperiodic and irreducible. Then there is a unique real eigenvalue, of multiplicity one, which is larger in absolute value than every other eigenvalue, and $h_{top}(\sigma) = \log\lambda$. Furthermore, by irreducibility of A, the left and right eigenvectors $u = (u_0,\dots,u_{N-1})$ and $v = (v_0,\dots,v_{N-1})^T$ associated to λ are unique up to a multiplicative factor, and they can be chosen to be strictly positive. We will scale them such that

$$\sum_{i=0}^{N-1} u_i v_i = 1.$$

Now define the **Parry measure** by

$$\begin{aligned} p_i &:= u_i v_i = \mu([i]), \\ p_{i,j} &:= \frac{a_{i,j}v_j}{\lambda v_i} = \mu([ij] \mid [i]), \end{aligned}$$

so $p_{i,j}$ indicates the conditional probability that $x_{n+1} = j$ knowing that $x_n = i$. Therefore $\mu([ij]) = \mu([i])\mu([ij] \mid [i]) = p_i p_{i,j}$. It is stationary (*i.e.*, shift-invariant) but not quite a product measure, but $\mu([i_m \dots i_n]) = p_{i_m} \cdot p_{i_m,i_{m+1}} \cdots p_{i_{n-1},i_n}$.

Theorem 12. *The Parry measure μ is the unique measure of maximal entropy for a subshift of finite type with irreducible transition matrix.*

Proof. In this proof, we will only show that $h_\mu(\sigma) = h_{top}(\sigma) = \log \lambda$, and skip the (more complicated) uniqueness part.

The definitions of mass of 1-cylinders and 2-cylinders are compatible, because (since v is a right eigenvector)

$$\sum_{j=0}^{N-1} \mu([ij]) = \sum_{j=0}^{N-1} p_i p_{i,j} = p_i \sum_{j=0}^{N-1} \frac{a_{i,j} v_j}{\lambda v_i} = p_i \frac{\lambda v_i}{\lambda v_i} = p_i = \mu([i]).$$

Summing over i, we get $\sum_{i=0}^{N-1} \mu([i]) = \sum_{i=0}^{N-1} u_i v_i = 1$, due to the our scaling.

To show that μ is shift-invariant, we take any cylinder set $Z = [i_m \dots i_n]$ and compute

$$\begin{aligned}
\mu(\sigma^{-1} Z) &= \sum_{i=0}^{N-1} \mu([i i_m \dots i_n]) = \sum_{i=0}^{N-1} \frac{p_i p_{i,i_m}}{p_{i_m}} \mu([i_m \dots i_n]) \\
&= \mu([i_m \dots i_n]) \sum_{i=0}^{N-1} \frac{u_i v_i a_{i,i_m} v_{i_m}}{\lambda v_i u_{i_m} v_{i_m}} \\
&= \mu(Z) \sum_{i=0}^{N-1} \frac{u_i a_{i,i_m}}{\lambda u_{i_m}} = \mu(Z) \frac{\lambda u_{i_m}}{\lambda u_{i_m}} = \mu(Z).
\end{aligned}$$

This invariance carries over to all sets in the σ-algebra $\mathcal{B}$ generated by the cylinder sets.

Based on the interpretation of conditional probabilities, the identity

$$\sum_{\substack{i_{m+1},\dots,i_n=0 \\ a_{i_k,i_{k+1}}=1}}^{N-1} p_{i_m} p_{i_m,i_{m+1}} \cdots p_{i_{n-1},i_n} = p_{i_m} \text{ and } \sum_{\substack{i_m,\dots,i_{n-1}=0 \\ a_{i_k,i_{k+1}}=1}}^{N-1} p_{i_m} p_{i_m,i_{m+1}} \cdots p_{i_{n-1},i_n} = p_{i_n} \quad (23)$$

follows because the left hand side indicates the total probability of starting in state i_m and reach some state after $n-m$ steps, respectively start at some state and reach state n after $n-m$ steps.

To compute $h_\mu(\sigma)$, we will confine ourselves to the partition $\mathcal{P}$ of 1-cylinder sets; this partition is generating, so this restriction is justified by Theorem 9.

$$\begin{aligned}
H(\bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}) &= -\sum_{\substack{i_0,\dots,i_{n-1}=0 \\ a_{i_k,i_{k+1}}=1}}^{N-1} \mu([i_0 \dots i_{n-1}]) \log \mu([i_0 \dots i_{n-1}]) \\
&= -\sum_{\substack{i_0,\dots,i_{n-1}=0 \\ a_{i_k,i_{k+1}}=1}}^{N-1} p_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n} \left(\log p_{i_0} + \log p_{i_0,i_1} + \dots + \log p_{i_{n-2},i_{n-1}}\right) \\
&= -\sum_{i_0=0}^{N-1} p_{i_0} \log p_{i_0} - (n-1) \sum_{i,j=0}^{N-1} p_i p_{i,j} \log p_{i,j},
\end{aligned}$$

by (23) used repeatedly. Hence

$$\begin{aligned}
h_\mu(\sigma) &= \lim_{n\to\infty} \frac{1}{n} H(\bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}) \\
&= -\sum_{i,j=0}^{N-1} p_i p_{i,j} \log p_{i,j} \\
&= -\sum_{i,j=0}^{N-1} \frac{u_i a_{i,j} v_j}{\lambda} \left(\log a_{i,j} + \log v_j - \log v_i - \log \lambda\right).
\end{aligned}$$

The first term is zero because $a_{i,j} \in \{0,1\}$. The second term (summing first over i) simplifies to

$$-\sum_{j=0}^{N-1} \frac{\lambda u_j v_j}{\lambda} \log v_j = -\sum_{j=0}^{N-1} u_j v_j \log v_j,$$

whereas the third term (summing first over j) simplifies to

$$\sum_{i=0}^{N-1} \frac{u_i \lambda v_i}{\lambda} \log v_i = \sum_{i=0}^{N-1} u_i v_i \log v_i.$$

Hence these two terms cancel each other. The remaining term is

$$\sum_{i,j=0}^{N-1} \frac{u_i a_{i,j} v_j}{\lambda} \log \lambda = \sum_{i=0}^{N-1} \frac{u_i \lambda v_i}{\lambda} \log \lambda = \sum_{i=0}^{N-1} u_i v_i \log \lambda = \log \lambda.$$

$\square$

Remark 2. *There are systems without maximising measure, for example among the "shifts of finite type" on infinite alphabets. To give an example (without proof!), if $\mathbb{N}$ is the alphabet, and the infinite transition matrix $A = (a_{i,j})_{i,j\in\mathbb{N}}$ is given by*

$$a_{i,j} = \begin{cases} 1 & \text{if } j \geqslant i-1, \\ 0 & \text{if } j < i-1, \end{cases}$$

then $h_{top}(\sigma) = \log 4$, but there is no measure of maximal entropy.

Exercise 20. *Find the maximal measure for the Fibonacci subshift of finite type. What is the limit frequency of the symbol zero in μ-typical sequences x?*

5 Equilibrium states and Gibbs measures

5.1 The Griffith-Ruelle Theorem

Let (Ω, σ) now be a one-sided or two-sided subshift of finite type. Throughout we will assume that the transition matrix is aperiodic and irreducible, so the Perron-Frobenius Theorem applies in its full force. Let $\psi : \Omega \to \mathbb{R}$ be a potential function, which we will assume to be Hölder continuous, *i.e.,* there is $C > 0$ and $\alpha \in (0, 1)$ such that if x_k and y_k agree for $|k| < n$, then $|\psi(x) - \psi(y)| \leqslant C\alpha^n$. The Hölder property can be applied to ergodic sums on n-cylinders Z:

$$\begin{aligned}\sup\{S_n\psi(x) : x \in Z\} &\geqslant \inf\{S_n\psi(x) : x \in Z\} \\ &\geqslant \sup\{S_n\psi(x) : x \in Z\} - \underbrace{\sum_{k=0}^{n-1} C\alpha^k}_{=C\frac{1-\alpha^n}{1-\alpha} < \frac{C}{1-\alpha}}\end{aligned} \tag{24}$$

Definition 12. *We say that a shift-invariant probability measure μ satisfies the* **Gibbs property** *if there are constants $C_2 \geqslant C_1 > 0$ such that for all n, all n-cylinders Z and all $x \in Z$,*

$$C_1 \leqslant \frac{\mu(Z)}{e^{S_n\psi(x) - Pn}} \leqslant C_2. \tag{25}$$

Here P is some constant, which, as we will see later, coincides with the topological pressure of the system. It is the number by which we need to translate the potential such that the measure of an n-cylinder scales as $e^{S_n(\psi - P)}$.

The main theorem of this section is sometimes called, in physics the Griffith-Ruelle Theorem (which actually also include analyticity of the pressure function):

Theorem 13. *If ψ is Hölder continuous potential function on an aperiodic irreducible subshift of finite type, then there is a unique Gibbs measure μ; this measure is the unique equilibrium state for (Ω, σ, ψ).*

We will prove this theorem in various steps. We start by a trick to reduce the potentially two-sided shift space to a one-sided shift.

Definition 13. *Two potential functions ψ and χ on Ω are called* **cohomologous** *if there is a function u such that*

$$\psi = \chi + u - u \circ \sigma. \tag{26}$$

From this definition, the following consequence are immediate for σ-invariant measure:

$$\begin{aligned} S_n\psi(x) &= S_n\chi(x) + u(x) - u \circ \sigma^n(x), \\ \lim_{n\to\infty} \frac{1}{n} S_n\psi(x) &= \lim_{n\to\infty} \frac{1}{n} S_n\chi(x) \quad \mu\text{-a.e.}, \\ \int \psi \, d\mu &= \int \chi \, d\mu. \end{aligned}$$

From this it is easy to derive that cohomologous potentials have the same equilibrium states. This will be used, in the next proposition, to reduce our task from two-sided shifts spaces to one-sided shift spaces.

Proposition 3. *If (Ω, σ) is a two-sided subshift of finite type and ψ a Hölder potential, then there is a potential χ which is also Hölder continuous but depending only on forward coordinates $(x_k)_{k \geqslant 0}$ of $x \in \Omega$, such that ψ and χ are cohomologous.*

Proof. For each symbol $n \in \{0, \dots, N-1\}$ pick a fix sequence $a^n \in \Omega$ such that $a^n_0 = n$. For $x \in \Omega$, let x^* be the sequence with $x^*_k = x_k$ if $k \geqslant 0$ and $x^*_k = a^n_k$ if $k < 0$ and $x_0 = n$. Next choose

$$u(x) = \sum_{j=0}^{\infty} \psi \circ \sigma^j(x) - \psi \circ \sigma^j(x^*).$$

Note that $|\psi \circ \sigma^j(x) - \psi \circ \sigma^j(x^*)| < C\alpha^j$, so the sum $u(x)$ converges and is continuous in x. Let $m = \lfloor n/2 \rfloor$. If x_k and y_k coincide for $|k| < n$, then

$$\begin{aligned} |u(x) - u(y)| &\leqslant \sum_{j=0}^{m} |\psi \circ \sigma^j(x) - \psi \circ \sigma^j(y)| + |\psi \circ \sigma^j(x^*) - \psi \circ \sigma^j(y^*)| \\ &\quad + \sum_{j>m} |\psi \circ \sigma^j(x) - \psi \circ \sigma^j(x^*)| + |\psi \circ \sigma^j(y) - \psi \circ \sigma^j(y^*)| \\ &\leqslant 2\sum_{j=0}^{m} C\alpha^{n-j} + 2\sum_{j>m} C\alpha^j \leqslant 4C\frac{\alpha^m}{1-\alpha}. \end{aligned}$$

Hence u is Hölder continuous with even a better Hölder exponent $\sqrt{\alpha}$ instead of α.

Now for $\chi = \psi - u + u \circ \sigma$, which is also Hölder, we have

$$\begin{aligned} \chi(x) &= \psi(x) + \sum_{j=0}^{\infty} \underbrace{\psi \circ \sigma^j(x) - \psi \circ \sigma^j(x^*)}_{\text{separate term } j=0} - \sum_{j=0}^{\infty} \psi \circ \sigma^j(\sigma x) - \psi \circ \sigma^j((\sigma x)^*) \\ &= \psi(x^*) + \sum_{j=1}^{\infty} \psi \circ \sigma^j(x) - \psi \circ \sigma^j(x^*) - \sum_{j=0}^{\infty} \psi \circ \sigma^j(\sigma x) - \psi \circ \sigma^j((\sigma x)^*) \\ &= \psi(x^*) + \sum_{j=0}^{\infty} \psi \circ \sigma^j((\sigma x)^*) - \psi \circ \sigma^j(\sigma x^*). \end{aligned}$$

This depends only on the forward coordinates of x. $\square$

Now that we can work on one-sided shift spaces, it is instructive to see why:

Lemma 4. *Gibbs measures of Hölder potentials are equilibrium states (*i.e., *measures that achieve the supremum in the Variational Principle).*

Proof. Let $\mathcal{P}$ is the partition into 1-cylinders, and recall that $\mathcal{P}_n = \bigvee_{k=0}^{n-1} \sigma^{-k}\mathcal{P}$ is the partition into n-cylinders. Write

$$\mathcal{Z}_n = \sum_{Z \in \mathcal{P}_n} e^{\sup\{S_n\psi(x) \,:\, x \in Z\}}.$$

be the n-th **partition function**. For Hölder continuous ψ, due to (24), whether we choose sup or inf, the result only changes by a multiplicative factor $e^{\frac{C}{1-\alpha}}$, independently of n.

Now suppose that μ satisfies the Gibbs property (25). Summing over all n-cylinders gives

$$C_1 \frac{\mathcal{Z}_n}{e^{Pn}} \leqslant \sum_{Z \in \mathcal{P}_n} \mu(Z) = 1 \leqslant C_2 \frac{\mathcal{Z}_n}{e^{Pn}}.$$

therefore $P = \lim_n \frac{1}{n} \log \mathcal{Z}_n$. Combining this with (20) in the proof of the Variational Principle, with $\mathcal{P}_n$ instead of $\mathcal{Q}_n$, we can write

$$H_\mu(\mathcal{P}_n) + \int_\Omega S_n\psi \; d\mu \leqslant \log \mathcal{Z}_n$$

Now we divide by n and take the limit $n \to \infty$ to obtain $h_\mu(\sigma) + \int \psi \; d\mu \leqslant P$.

For any x in an n-cylinder Z, we have

$$\begin{aligned} -\mu(Z)\log\mu(Z) + \int_Z S_n\psi \; d\mu &\geqslant -\mu(Z)\left[\log\mu(Z) - S_n\psi(x) + \frac{C}{1-\alpha}\right] \\ &\geqslant -\mu(Z)\left[\log C_2 e^{-Pn+S_n\psi(x)} - S_n\psi(x) + \frac{C}{1-\alpha}\right] \\ &= \mu(Z)\left[Pn - \log C_2 - \frac{C}{1-\alpha}\right]. \end{aligned}$$

Summing over all n-cylinders $Z \in \mathcal{P}_n$ gives

$$\begin{aligned} H_\mu(\mathcal{P}_n) + \int_\Omega S_n\psi \; d\mu &\geqslant \sum_{Z \in \mathcal{P}_n} \mu(Z)\left[Pn - \log C_2 - \frac{C}{1-\alpha}\right] \\ &= Pn - \log C_2 - \frac{C}{1-\alpha}. \end{aligned}$$

Dividing by n and letting $n \to \infty$, we find $h_\mu(\sigma) + \int_\Omega \psi \; d\mu \geqslant P$. Therefore we have equality $h_\mu(\sigma) + \int_\Omega \psi \; d\mu = P$.

To show that $P = P_{top}(\sigma, \psi)$, take $\varepsilon > 0$ arbitrary and M such that $2^{-(M+1)} \leqslant \varepsilon < 2^{-M}$. Taking a point x in each $n+M$-cylinder then produces an (n, ε)-separated set $E_n(\varepsilon)$ of maximal cardinality. Therefore, as in (17), we find

$$\mathcal{Z}_{n+M} = \sup \left\{ \sum_{x \in E} e^{S_n \psi(x)} \ : \ E \text{ is } (n, \varepsilon)\text{-separated} \right\} =: P_n(\sigma, \psi, \varepsilon).$$

The ε-dependence of the left hand side is only in the choice of M. This dependence disappears when we take the limit $\lim_n \frac{1}{n} \log \mathcal{Z}_n = \lim_n \frac{1}{n} \log P_n(\sigma, \psi, \varepsilon)$, and therefore taking the limit $\varepsilon \to 0$ gives

$$P = \lim_n \frac{1}{n} \log \mathcal{Z}_n = \lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log P_n(\sigma, \psi, \varepsilon) = P_{top}(\sigma, \psi).$$

This completes the proof. □

Next we give somewhat abstract results from functional analysis to find a candidate Gibbs measure as the combination of the eigenfunction and eigenmeasure of a particular operator and its dual.

Definition 14. *The Ruelle-Perron-Frobenius operator acting on functions $f : \Omega \to \mathbb{R}$ is defined as*

$$\mathcal{L}_\psi f(x) = \sum_{\sigma y = x} e^{\psi(y)} f(y). \tag{27}$$

The dual operator $\mathcal{L}_\psi^*$ acts on measures: $\int f \ d(\mathcal{L}_\psi^* \nu) = \int \mathcal{L}_\psi f \ d\nu$ for all $f \in L^1(\nu)$.

This operator describes how densities are transformed by the dynamics. For instance, if instead of σ we had a differentiable transformation $T : [0,1] \to [0,1]$ and $\psi = -\log |T'|$, then $\mathcal{L}_\psi f(x) = \sum_{Ty=x} \frac{1}{|T'(y)|} f(y)$ which, when integrated over $[0,1]$, we can recognise as the integral formula for a change of coordinates $x = T(y)$.

The following theorem can be seen as the operator-version of the Perron-Frobenius Theorem for matrices:

Theorem 14. *If Ω is a one-sided subshift of finite type, with aperiodic irreducible transition matrix, then there is a unique $\lambda > 0$ and continuous positive (or more precisely: bounded away from zero) function h and a probability measure ν such that*

$$\mathcal{L}_\psi h = \lambda \ h \qquad\qquad \mathcal{L}_\psi^* \nu = \lambda \ \nu.$$

The Ruelle-Perron-Frobenius operator has the properties:

1. $\mathcal{L}_\psi$ is positive: $f \geqslant 0$ implies $\mathcal{L}_\psi f \geqslant 0$.

2. $\mathcal{L}_\psi^n f(x) = \sum_{\sigma^n y = x} e^{S_n\psi(y)} f(y)$.

3. ν is in general **not** σ-invariant. Instead it satisfies

$$\nu(\sigma A) = \lambda \int_A e^{-\psi}\, d\nu \tag{28}$$

whenever $\sigma : A \to \sigma(A)$ is one-to-one and A is measurable. Measures with this property are called $\lambda e^{-\psi}$**-conformal**.

4. Instead, the measure $d\mu = h\, d\nu$ is σ-invariant. We can always scale h such that μ is a probability measure too.

5. We will see later that $\lambda = e^P$ where P is the topological pressure.

Proof. Property 1. is obvious, since $e^{\psi(y)}$ is always positive. Property 2. follows by direct computation. For Property 3., we have

$$\begin{aligned}
\lambda \int_A e^{-\psi}\, d\nu &= \lambda \int_\Omega e^{-\psi} 1\!\!1_A\, d\nu = \int_\Omega e^{-\psi} 1\!\!1_A\, d(\lambda\nu) \\
&= \int_\Omega e^{-\psi(x)} 1\!\!1_A(x)\, d(\mathcal{L}_\psi^* \nu) = \int_\Omega \mathcal{L}_\psi(e^{-\psi(x)} 1\!\!1_A(x))\, d\nu \\
&= \int_\Omega \sum_{\sigma y = x} e^{\psi(y)} e^{-\psi(y)} 1\!\!1_A(y)\, d\nu = \int_\Omega \sum_{\sigma y = x} 1\!\!1_A(y)\, d\nu
\end{aligned}$$

Since $\sigma : A \to \sigma(A)$ is one-to-one, $\sum_{\sigma y = x} 1\!\!1_A(y) = 1$ if $x \in \sigma(A)$ and $= 0$ otherwise. Hence the integral $\int \sum_{\sigma y = x} 1\!\!1_A(y)\, d\nu = \nu(\sigma A)$ as required.

For Property 4., first check that

$$\mathcal{L}_\psi f(x) \cdot g(x) = \sum_{\sigma y = x} e^{\psi(y)} f(y) g(x) = \sum_{\sigma y = x} e^{\psi(y)} f(y) g(\sigma y) = \mathcal{L}_\psi(f \cdot g \circ \sigma)(x).$$

This gives

$$\begin{aligned}
\int_\Omega f\, d\mu &= \int_\Omega f \cdot h\, d\nu = \frac{1}{\lambda} \int_\Omega f \cdot \mathcal{L}_\psi h\, d\nu \\
&= \frac{1}{\lambda} \int_\Omega \mathcal{L}_\psi (h \cdot f \circ \sigma)\, d\nu = \frac{1}{\lambda} \int_\Omega h \cdot f \circ \sigma\, d(\mathcal{L}_\psi^* \nu) \\
&= \int_\Omega f \circ \sigma \cdot h\, d\nu = \int_\Omega f \circ \sigma\, d\mu
\end{aligned}$$

Property 5. will follow from the next proposition. $\square$

Proposition 4. *For Hölder potential ψ, the measure $d\mu = h\, d\nu$ satisfies the Gibbs property with $P = \log \lambda$.*

Proof. For each $z \in \Omega$ and n-cylinder Z, there is at most one $y \in Z$ with $\sigma^n y = z$. Take $x \in Z$ arbitrary. Then

$$\mathcal{L}_\psi^n(h \cdot 1\!\!1_Z) = \sum_{\sigma^n y = z} e^{S_n\psi(y)} h(y) 1\!\!1_Z(y) \leqslant \underbrace{e^{\frac{C}{1-\alpha}} \|h\|_\infty}_{C_2} e^{S_n\psi(x)}.$$

Hence

$$\begin{aligned}\mu(Z) &= \int_Z h \, d\nu = \int_\Omega h \cdot 1\!\!1_Z \, d\nu = \lambda^{-n} \int_\Omega h \cdot 1\!\!1_Z \, d(\mathcal{L}_\psi^{*n}\nu) \\ &= \lambda^{-n} \int_\Omega \mathcal{L}_\psi^n(h \cdot 1\!\!1_Z) \, d\nu \leqslant C_2 \lambda^{-n} e^{S_n\psi(x)}. \end{aligned} \tag{29}$$

On the other hand, since the subshift of finite type is irreducible, there is some uniform integer M and $y \in Z$ such that $\sigma^{n+M}(y) = z$. Therefore

$$\mathcal{L}_\psi^n(h \cdot 1\!\!1_Z) \geqslant e^{S_{n+M}\psi(y)} h(y) \geqslant \underbrace{e^{-M\|\psi\|_\infty} e^{-\frac{C}{1-\alpha}} \cdot \inf h}_{C_1} e^{S_n\psi(x)}.$$

Integrating over Z gives us $\mu(Z) \geqslant C_1 \lambda^{-n} e^{S_n\psi(x)}$ by the same reasoning as in (29). Therefore

$$C_1 \leqslant \frac{\mu(Z)}{\lambda^{-n} e^{S_n\psi(x)}} \leqslant C_2,$$

for all n-cylinders and thus if we choose $e^P = \lambda$, we obtain the Gibbs property. $\square$

Lemma 5. *The Gibbs measure is unique.*

Proof. If both μ and μ' satisfy (25) for some constants C_1, C_2, P and C_1', C_2', P' then we can first take (1) for μ' and sum over all n-cylinders. This gives

$$C_1' e^{-P'n} \sum_{Z \in \mathcal{P}_n} e^{S_n\psi(x)} \leqslant 1 \leqslant C_2' e^{-P'n} \sum_{Z \in \mathcal{P}_n} e^{S_n\psi(x)},$$

so that $P' = \lim_n \frac{1}{n} \log \sum_{Z \in \mathcal{P}_n} e^{S_n\psi(x)}$, independently of μ'. Therefore $P' = P$.

Now divide (25) for μ' by the same expression for μ. This gives

$$\frac{C_1'}{C_2} \leqslant \frac{\mu'(Z)}{\mu(Z)} \leqslant \frac{C_2'}{C_1}$$

independently of Z. Therefore μ' and μ are equivalent: they have the same null-sets. In particular, for each continuous f, the set of points $x \in \Omega$ for which the Birkhoff Ergodic Theorem holds for μ' and μ differs by at most a nullset. For any point which is typical for both, we find $\int f \, d\mu' = \lim_n \frac{1}{n} S_n f(x) = \int f \, d\mu$. Therefore $\mu = \mu'$. $\square$

5.2 Upper semicontinuity of entropy

For a continuous potential $\psi : X \to \mathbb{R}$, and a sequence of measures $(\mu_n)_{n\in\mathbb{N}}$ such that $\mu_n \to \mu$ in the weak* topology, we always have $\int \psi \, d\mu_n \to \int \psi \, d\mu$, simply because that is the definition of weak* convergence. However, entropy isn't continuous in this sense. For example, if (Σ, σ) is the full shift on two symbols, then the $\frac{1}{2}$-$\frac{1}{2}$ Bernoulli measure μ is the measure of maximal entropy $\log 2$. If $x \in \Sigma$ is a typical point (in the sense of the Birkhoff Ergodic Theorem), then we can create a sequence of measure μ_n by

$$\mu_n = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\sigma^j y}$$

where $y = \overline{x_0 x_1 \dots x_{n-1}}$ is the n-periodic point in the same n-cylinder as x. For these measure $\mu_n \to \mu$ in the weak* topology, but since μ_n is supported on a single periodic orbit, the entropy $h_{\mu_n}(\sigma) = 0$ for every n. Therefore

$$\lim_{n\to\infty} h_{\mu_n}(\sigma) = 0 < \log 2 = h_\mu(\sigma).$$

Lacking continuity, the best we can hope for is **upper semicontinuity** (USC) of the entropy function, *i.e.,*

$$\mu_n \to \mu \text{ implies } h_\mu(\sigma) \geqslant \limsup_{n\to\infty} h_{\mu_n}(\sigma).$$

In other words, the value of h can make a jump **upwards** at the limit measure, but not downwards. Fortunately, the entropy function $\mu \mapsto h_\mu(\sigma)$ is indeed USC for subshifts on a finite alphabet, and USC is enough to guarantee the existence of equilibrium states.

Proposition 5. *Let (X, T) be a continuous dynamical system on a compact metric space X. Assume that potential $\psi : X \to \mathbb{R}$ is continuous. If the entropy function is USC, then there is an equilibrium state,*

Proof. We use the Variation Principle

$$P(\psi) = \sup\{h_\nu(T) + \int \psi \, d\nu \ : \ \nu \text{ is } T\text{-invariant probability measure}\}. \tag{30}$$

Hence there exists a sequence $(\mu_n)_{n\in\mathbb{N}}$ such that $P(\psi) = \lim_n h_{\mu_n}(T) + \int \psi \, d\mu_n$. Passing to a subsequence (n_k) if necessary, we can assume that $\mu_{n_k} \to \mu$ as $k \to \infty$ in the weak* topology, and therefore $\int \psi \, d\mu_{n_k} \to \int \psi \, d\mu$ as $k \to \infty$. Due to upper semicontinuity,

$$P(\psi) = \limsup_{k\to\infty} h_{\mu_{n_k}}(T) + \int \psi \, d\mu_{n_k} \leqslant h_\mu(T) + \int \psi \, d\mu,$$

but also $h_\mu(T) + \int \psi \, d\mu \leqslant P(\psi)$ by (30). Hence μ is an equilibrium state. $\square$

The following corollary follows in the same way.

Corollary 4. *Let (X, T) be a continuous dynamical system on a compact metric space X, and suppose that the entropy function is USC. Let ψ_β be a family (continuous in β) of continuous potentials and $\beta \to \beta^*$. If μ_β are equilibrium states for ψ_β and $\mu_\beta \to \mu_{\beta^*}$ in the weak* topology as $\beta \to \beta^*$, then μ_{β^*} is an equilibrium state for ψ_{β^*}.*

Upper semicontinuity of entropy also gives us another way of characterizing entropy:

Lemma 6 (Dual Variational Principle)**.** *Let (X, T) be a continuous dynamical system on a compact metric space. Assume that the entropy function is upper semi-continuous and that $P(0) < \infty$. Then*

$$h_\mu(T) = \inf\{P(\psi) - \int \psi \, d\mu \; : \; \psi : X \to \mathbb{R} \text{ continuous}\}.$$

Proof. See [6, Theorem 4.2.9] or [9, Theorem 9.12]. □

5.3 Smoothness of the pressure function

In Section 5.1 we have given conditions under which a Gibbs measure is unique. Gibbs measures are equilibrium states, but that doesn't prove uniqueness of equilibrium states. There could in principle be equilibrium states that are not Gibbs measures. In this section we will connect uniqueness of equilibrium states of a parametrised family ψ_β of potentials to smoothness of the pressure function $\beta \mapsto P(\psi_\beta)$. In fact, the remaining part of the Griffith-Ruelle Theorem is about smoothness, more precisely analyticity, of pressure function when $\psi_\beta = \beta \cdot \psi$, for inverse temperature $\beta \in \mathbb{R}$.

Theorem 15 (Griffith-Ruelle Theorem (continued))**.** *If ψ is Hölder continuous potential function on an aperiodic irreducible subshift of finite type, then the pressure function*

$$\beta \mapsto P(\beta \cdot \psi)$$

is real analytic.

We will not prove this here, but rather focus on how differentiability of $\beta \mapsto P(\psi_\beta)$ is related to equilibrium states. In the simplest case when $\psi_\beta = \beta \cdot \psi$, then the graph of

$$\beta \mapsto P(\beta \cdot \psi) := \sup\{h_\nu(T) + \beta \int \psi \, d\nu : \nu \text{ is } T\text{-invariant probability measure } \}$$

is the envelope of straight lines $\beta \mapsto h_\nu(T) + \beta \int \psi \, d\nu$, and therefore continuous. We think of ψ (or at least $\int \psi \, d\nu$) as non-positive, so that maximising $P(\beta)$ corresponds to maximising entropy and minimising energy in agreement with the Laws of Thermodynamics. Hence the graph $\beta \mapsto P(\beta)$, as the envelop of non-increasing lines, is non-increasing and convex.

Furthermore, if μ_0 is an equilibrium state for β_0, and $\beta \mapsto P(\beta)$ is differentiable at $\beta = \beta_0$, then $P'(\beta_0) = \int \psi \, d\mu_0$. Hence if μ_0 and μ_0' are two different equilibrium states for β_0 with $\int \psi \, d\mu_0 \neq \int \psi \, d\mu_0'$, then $\beta \mapsto P(\beta)$ cannot be differentiable at $\beta = \beta_0$.

Definition 15. *Given a continuous potential $\psi : X \to \mathbb{R}$, we say that:*

- *a measure ν on X is a* **tangent measure** *if*

$$P(\psi + \phi) \geqslant P(\psi) + \int \phi \, d\nu \text{ for all continuous } \phi : X \to \mathbb{R}. \tag{31}$$

- *P is* **differentiable** *at ψ if there is a unique tangent measure.*

It would be more correct to speak of tangent functional since a priori, we just have $\nu \in C^*(X)$, but in all cases ν turns out to be indeed an "unsigned" probability measure.

So compared to differentiability of $\beta \mapsto P(\beta \cdot \psi)$, differentiability in the above sense requires (31) not just for $\phi = (\beta - 1) \cdot \psi$ (which follows from convexity of $\beta \mapsto P(\beta \cdot \psi)$), but for all continuous $\phi : X \to \mathbb{R}$.

Theorem 16. *Let (X, T) be a continuous dynamical system on a compact metric space X, and suppose that the entropy function is USC. Let $\psi : X \to \mathbb{R}$ be a continuous potential. Then P is differentiable at ψ with derivative μ if and only if*

$$\lim_{\varepsilon \to 0} \frac{P(\psi + \varepsilon\phi) - P(\psi)}{\varepsilon} = \int \phi \, d\mu \tag{32}$$

for all continuous $\phi : X \to \mathbb{R}$. In this case, μ is the unique equilibrium state for ψ.

Proof. We start by proving that the tangent measures are exactly the equilibrium states. Assume that μ is an equilibrium state for ψ. Then

$$\begin{aligned} P(\psi + \phi) &= \sup_\nu \{ h_\nu(T) + \int \psi \, d\nu + \int \phi \, d\nu \} \\ &\geqslant h_\mu(T) + \int \psi \, d\mu + \int \phi \, d\mu = P(\psi) + \int \phi \, d\mu \end{aligned}$$

for all continuous $\phi : X \to \mathbb{R}$, so μ is a tangent measure.

For the converse, assume that ν satisfies (31). Since $\psi : X \to \mathbb{R}$ is continuous on a compact space, we have

$$-\infty < \inf \psi \leqslant P(\psi) \leqslant P(0) + \sup \psi < \infty,$$

because by the Variational Principle and upper semicontinuity, $P(0) = \sup_\nu h_\nu(T) < \infty$. Since ν satisfies (31), we have

$$\begin{cases} P(\psi) + 1 = P(\psi + 1) \geqslant P(\psi) + \int d\nu, \\ P(\psi) - 1 = P(\psi - 1) \geqslant P(\psi) - \int d\nu, \end{cases}$$

so $\int d\nu = 1$ follows. Furthermore, if $\phi \geqslant 0$, we have

$$P(\psi) \geqslant P(\psi - \phi) \geqslant P(\psi) - \int \phi \, d\nu,$$

so $\int \phi \, d\nu \geqslant 0$. This shows that ν is an "unsigned" probability measure. To prove T-invariance, recall about cohomologous functions that

$$P(\psi) = P(\psi + \eta \cdot (\phi \circ T - \phi)) \geqslant P(\psi) + \eta \int \phi \circ T - \phi \, d\nu,$$

hence $0 \geqslant \eta \int \phi \circ T - \phi \, d\nu$. Since η can be both positive or negative, there is only one possibility: $0 = \int \phi \circ T - \phi \, d\nu$, and so ν is indeed T-invariant.

Finally, by Lemma 6,

$$\begin{aligned} h_\nu(T) &= \inf\{P(\psi + \phi) - \int \psi + \phi \, d\nu \ : \ \phi : X \to \mathbb{R} \text{ continuous}\} \\ &\geqslant \inf\{P(\psi) + \int \phi \, d\nu - \int \psi + \phi \, d\nu \ : \ \phi : X \to \mathbb{R} \text{ continuous}\} \\ &= P(\psi) - \int \phi \, d\nu \geqslant h_\nu(T). \end{aligned}$$

Therefore ν is indeed an equilibrium state.

Now for the second half of the proof, assume that P is differentiable at ψ with derivative μ, so μ is the only tangent measure, and therefore only equilibrium state. We need to establish (32). For $\varepsilon \neq 0$ and given ϕ, let μ_ε be an equilibrium state for $\psi + \varepsilon\phi$. Then $\mu_\varepsilon \to \mu$ as $\varepsilon \to 0$ by Corollary 4. Since μ is a tangent measure

$$P(\psi + \varepsilon\phi) - P(\psi) \geqslant \varepsilon \int \phi \, d\mu,$$

and since μ_ε are also tangent measures,

$$-\left(P(\psi + \varepsilon\phi) - P(\psi)\right) = P(\psi + \varepsilon\phi - \varepsilon\phi) - P(\psi + \varepsilon\phi) \geqslant -\varepsilon \int \phi \, d\mu_\varepsilon.$$

Combining the two, we find

$$\int \phi \, d\mu \leqslant \frac{P(\psi + \varepsilon\phi) - P(\psi)}{\varepsilon} \leqslant \int \phi \, d\mu_\varepsilon$$

if $\varepsilon > 0$ or with reversed inequalities if $\varepsilon < 0$. Now $\int \phi \, d\mu_\varepsilon \to \int \psi \, d\mu$ as $\varepsilon \to 0$, so (32) follows.

Conversely, if (32) holds for all continuous $\phi : X \to \mathbb{R}$ and ν is an arbitrary tangent measure, then

$$\int \phi \, d\mu = \lim_{\varepsilon \searrow 0} \frac{P(\psi + \varepsilon\phi) - P(\psi)}{\varepsilon} \geqslant \int \phi \, d\nu,$$

and also

$$\int \phi \, d\mu = \lim_{\varepsilon \nearrow 0} \frac{P(\psi + \varepsilon\phi) - P(\psi)}{\varepsilon} \leqslant \int \phi \, d\nu.$$

Hence $\int \phi \, d\mu = \int \phi \, d\nu$ for all continuous $\phi : X \to \mathbb{R}$, whence $\mu = \nu$, and P is indeed differentiable with single derivative. $\square$

In view of the Griffith-Ruelle Theorem, this motivates the definition:

Definition 16. *The system (X, T) with potential $\psi : X \to \mathbb{R}$ undergoes a* **phase transition** *at parameter β_0 if $\beta \mapsto P(\beta \cdot \psi)$ fails to be analytic at β_0.*

It is where pressure fails to be analytic, that equilibrium states may be non-existent (possible, if the potential is non-continuous), non-unique (possible, if the potential is non-Hölder) and/or discontinuous under change of parameters.

6 Hausdorff Dimension of Repellors

Let $f : D \subset [0,1] \to [0,1]$ be defined on a domain $D = \cup_{k=0}^{N-1} D_k$, where each D_k is a closed interval and $f : D_k \to [0,1]$ is surjective, C^2-smooth and expanding, *i.e.,* $\inf\{|f'(x)| \ : \ x \in D\} > 1$. Recall that $f^n = f \circ \cdots \circ f$ is the n-fold composition of a map and define

$$X = \{x \in [0,1] \ : \ f^n(x) \in D \text{ for all } n \geqslant 0\}.$$

This set X is sometimes called the **repellor** of f, and is usually a Cantor set, *i.e.,* compact, totally disconnected and without isolated points.

Example 13. *If*

$$f(x) = \begin{cases} 3x & \text{if } x \in [0, \frac{1}{3}] = D_0, \\ 3x - 2 & \text{if } x \in [\frac{2}{3}, 1] = D_1, \end{cases}$$

then X becomes the middle third Cantor set.

Example 14. *The full tent-map is defined as*

$$T(x) = \begin{cases} 2x & \text{if } x \in [0, \frac{1}{2}] = D_0, \\ 2(1 - x) & \text{if } x \in [\frac{1}{2}, 1] = D_1. \end{cases}$$

Here $X = [0,1]$, so not a Cantor set. (In this case, D_0 and D_1 overlap at one point, and that explains the difference.)

Definition 17. *Given some set A, an* **(open) ε-cover** $\mathcal{U} = \{U_j\}_{j\in\mathbb{N}}$ *of A is a collection of open sets such that $A \subset \cup_j U_j$ and the diameters* $\mathrm{diam}(U_j) < \varepsilon$ *for all j.*[6]

The **δ-dimensional Hausdorff measure** *is defined as*

$$\mu_\delta(A) = \lim_{\varepsilon\to 0} \inf\{\sum_j (\mathrm{diam}(U_j))^\delta \ : \ \mathcal{U} \text{ is an open } \varepsilon\text{-cover of } A\}.$$

It turns out that there is a unique δ_0 such that

$$\mu_\delta(A) = \begin{cases} \infty & \text{if } \delta < \delta_0, \\ 0 & \text{if } \delta > \delta_0. \end{cases}$$

This δ_0 is called the **Hausdorff dimension** *of A, and it is denoted as* $\dim_H(A)$.

Lebesgue measure on the unit cube $[0,1]^n$ coincides, up to a multiplicative constant, with n-dimensional Hausdorff measure. However, for "fractal" sets such as the middle third Cantor sets, the "correct" value of δ_0 can be non-integer, as we will argue in the next example.

Example 15. *Let X be the middle third Cantor set. For each n, we can cover X with 2^n closed intervals of length 3^{-n}, namely*

$$[0,3^{-n}] \cup [2\cdot 3^{-n}, 3\cdot 3^{-n}] \cup [6\cdot 3^{-n}, 7\cdot 3^{-n}] \cup [8\cdot 3^{-n}, 9\cdot 3^{-n}] \cup \cdots \cup [(3^n-1)\cdot 3^{-n}, 1].$$

We can make this into an open cover $\mathcal{U}_\varepsilon$ (with $\varepsilon = 3^{-n}(1+2\cdot 3^{-n})$) by thickening these intervals a little bit, i.e., *replacing $[m\cdot 3^{-n}, (m+1)\cdot 3^{-n}]$ by $(m\cdot 3^{-n} - 3^{-2n}, (m+1)\cdot 3^{-n} + 3^{-2n})$. Then*

$$\mu_\delta(X) \leqslant 2^n \cdot (3^{-n} + 2\cdot 3^{-2n})^\delta = 2^n \cdot 3^{-\delta n} \cdot (1 + 2\cdot 3^{-n})^\delta =: E_n.$$

Then

$$\lim_{n\to\infty} E_n = \begin{cases} \infty & \text{if } \delta < \frac{\log 2}{\log 3}, \\ 1 & \text{if } \delta = \frac{\log 2}{\log 3}, \\ 0 & \text{if } \delta > \frac{\log 2}{\log 3}. \end{cases}$$

This shows that $\dim_H(X) \leqslant \frac{\log 2}{\log 3}$. In fact, $\dim_H(X) = \frac{\log 2}{\log 3}$, but showing that covers $\mathcal{U}_\varepsilon$ are "optimal" is a bit messy, and we will skip this part.

[6] We can include $U_j = \emptyset$ for some j, so finite covers $\{U_j\}_{j=1}^R$ can always be extended to countable covers $\{U_j\}_{j\in\mathbb{N}}$ if necessary.

Coming back to our expanding interval map f, we choose the potential

$$\psi_\beta(x) = -\beta \log |f'(x)|,$$

which is C^1-smooth on each D_k, and negative for $\beta > 0$. The ergodic sum

$$\begin{aligned} S_n\psi_\beta(x) &= -\beta \sum_{k=0}^{n-1} \log |f' \circ f^k(x)| \\ &= -\beta \log \prod_{k=0}^{n-1} |f' \circ f^k(x)| = \log |(f^n)'(x)|^{-\beta} \end{aligned} \tag{33}$$

by the Chain Rule.

Theorem 17. *Let* $([0,1], f)$ *with repellor* $X = \{x \in [0,1] \ : \ f^n(x) \in D = \cup_k D_k$ *for all* $n \geqslant 0\}$ *and potential* $\psi_\beta = -\beta \log |f'|$ *be as above. Then there is a unique* β_0 *at which the pressure* $P(\psi_\beta)$ *vanishes, and* $\dim_H(X) = \beta_0$*.*

Sketch of Proof. We use symbolic dynamics on X by setting

$$e(x) = y_0y_1y_2 \dots \text{ with } y_n = k \in \{0, \dots, N-1\} \text{ if } f^n(x) \in D_k.$$

This uniquely associates a code $y \in \Sigma := \{0, \dots, N-1\}^{\mathbb{N}_0}$ to x provided the D_k's don't overlap, as in Example 14. If some D_k's overlap at one point, this affects only countably many points, and therefore we can neglect them. Conversely, since f is expanding, each code $y \in \Sigma$ is associated to no more than one $x \in X$.

To each n-cylinder set $[y_0y_1 \dots y_{n-1}] = Z \subset \Sigma$, we can associate a closed interval J such that $f^k(J) \subset D_{y_k}$ for $0 \leqslant k < n$, and in fact $f^{n-1}(J) = D_{y_{n-1}}$ and $f^n(J) = [0,1]$.

The C^2-smoothness of f guarantees that ψ_β transfers to a Hölder potential $\tilde{\psi}_\beta(y) := \psi_\beta \circ e^{-1}(y)$ on Σ, and therefore, for each β, we can apply the Griffith-Ruelle Theorem and obtain a unique equilibrium state which is also a Gibbs measure. Use the coding map $e : X \to \Sigma$ to transfer this to (X, f, ψ_β): For each $\beta \in \mathbb{R}$, there is a unique equilibrium state μ_β which is also a Gibbs measure, for ψ_β.

Therefore, there are $C_1, C_2 > 0$ depending only on f and β, such that for all n, all interval J associated to n-cylinders and all $x \in J \cap X$,

$$C_1 \leqslant \frac{\mu_\beta(J \cap X)}{e^{S_n(\psi_\beta(x) - P)}} \leqslant C_2, \tag{34}$$

where $P = P(\psi_\beta)$ is the pressure.

Recall from (33) that $e^{S_n(\psi_\beta(x)-P)} = e^{-nP}|(f^n)'(x)|^{-\beta}$ for $x \in J \cap X$; in fact the same holds for all $x \in J$. By the Mean Value Theorem, and since $f^n(J) = [0,1]$, there is

$x_J \in J$ such that $|(f^n)'(x_J)| = 1/\text{diam}(J)$. Now we don't know if $x_J \in X$, but we use a distortion argument[7] to rewrite (34) to

$$\frac{\mu_\beta(J)}{C_2} \leqslant e^{-Pn}\text{diam}(J)^\beta \leqslant \frac{\mu_\beta(J)}{C_1}$$

and summing over all cylinder sets, we arrive at

$$\frac{1}{C_2} \leqslant e^{-Pn} \sum_J \text{diam}(J)^\beta \leqslant \frac{1}{C_1}. \tag{35}$$

Now for $\beta = 0$, this gives $\frac{1}{C_2} \leqslant e^{-Pn}\#\{ \text{ intervals } J\} \leqslant \frac{1}{C_1}$, and since there are N^n intervals, we get $P(\psi_0) = \lim_n \frac{1}{n} \log N^n = \log N > 0$, which is indeed the topological entropy of the map f.

We have $\sum_J \text{diam}(J) \leqslant 1$, and therefore, for $\beta > 1$, $\sum_J \text{diam}(J)^\beta \to 0$ exponentially in n. Hence (35) implies that $P(\psi_\beta) < 0$ for all $\beta > 1$. Now since $\beta \mapsto P(\psi_\beta)$ is non-increasing and convex, this means that there is a unique β_0 such that $P(\psi_\beta) = 0$ for $\beta = \beta_0$.

For this β_0, we find

$$\frac{1}{C_2} \leqslant \sum_J \text{diam}(J)^{\beta_0} \leqslant \frac{1}{C_1}.$$

The sets J can be thickened a bit to produce an open ε-cover $\mathcal{U}_\varepsilon$ (with $\varepsilon < 2(\inf |f'|)^{-n}) \to 0$ as $n \to \infty$). This gives $\dim_H(X) \leqslant \beta_0$. To show that also $\dim_H(X) \geqslant \beta_0$, we need a similar argument that covers $\mathcal{U}_\varepsilon$ are "optimal" that we skipped in Example 15, and which we will omit here as well. □

Exercise 21. *Assume that $\cup_k D_k = [0, 1]$ as in Example 14. Show that $\beta_0 = 1$ and that the unique equilibrium state μ_1 is equivalent to Lebesgue measure.*

7 Gibbs distributions and large deviations - an example

The following is an adaptation of Example 1.2.1. from Keller's book [6]. Assume first that the entire system consists of a single particle that can assume states in alphabet $\mathcal{A} = \{0, \dots, N-1\}$, with energies $-\beta\psi_0(a)$ (where parameter $\beta \in \mathbb{R}$ denotes the inverse temperature). We call

$$\mathbb{P}(x = a) = q_\beta(a) := \frac{e^{-\beta\psi_0(a)}}{\sum_{a' \in \mathcal{A}} e^{-\beta\psi_0(a')}} \tag{36}$$

[7] which we will sweep under the carpet

a **Gibbs distribution**. (The Gibbs distribution in this section should not be confused with a Gibbs measure that satisfies the Gibbs property (1).) Note that a Gibbs distribution isn't a fixed state the particle is in, it is a probability distribution indicating (presumably) what proportion of time the particle assumes state $a \in \mathcal{A}$.

In this simple case, the configuration space $\Omega = \mathcal{A}$ and as there is no dynamics, entropy is just

$$H(q_\beta) = -\sum_{p \in \mathcal{P}} q_\beta(p) \log q_\beta(p)$$

with respect to the only sensible partition, namely into single symbols: $\mathcal{P} = \{\omega = a\}_{a \in \mathcal{A}}$. We know from Corollary 3 that

$$H(q_\beta) - \beta \int \psi_0 \, dq_\beta \geqslant H(\pi) - \beta \int \psi_0 \, d\pi$$

for every probability measure π on $\mathcal{A}$ with equality if and only if $\pi = q_\beta$. Hence the Gibbs measure is the equilibrium state for ψ_0. We take this as inspiration to measure how far π is from the "optimal" measure q_β by defining

$$d_\beta(\pi) = \left(H(q_\beta) - \beta \int \psi_0 \, dq_\beta \right) - \left(H(\pi) - \beta \int \psi_0 \, d\pi \right). \tag{37}$$

Let us now replace the single site by a finite lattice or any finite collection G of sites, say $n = \#G$, with particles at every site assuming states in $\mathcal{A}$. Thus now the configuration space is $\Omega = \mathcal{A}^G$ of cardinality $\#\Omega = N^n$, where we think of n as huge (number of Avogadro or like).

Assume that the energy $\psi(\omega)$ of configuration $\omega \in \Omega$ is just the sum of the energies of the separate particles: $\psi(\omega) = \sum_{g \in G} \psi_0(\omega_g)$. So there is no interaction between particles whatsoever; no coherence in the set G.

We can still define the Gibbs measure (and hence equilibrium state for ψ) as before; it becomes the product measure of the Gibbs measures at each site:

$$\mu_\beta(\omega) = \frac{e^{-\beta\psi(\omega)}}{\sum_{\omega' \in \Omega} e^{-\beta\psi(\omega')}} = \prod_{g \in G} \frac{e^{-\beta\psi_0(\omega_g)}}{\sum_{a' \in \mathcal{A}} e^{-\beta\psi_0(a')}}.$$

It is convenient to denote the denominator, *i.e.,* **partition function**, as $\mathcal{Z}(\beta) = \sum_{\omega' \in \Omega} e^{-\beta\psi(\omega')}$.

The measures $\mu_\beta(\omega)$ for each singular configuration are minute, even if ω minimises energy. Note however, that for small temperature (large β), configurations with minimal energies are extremely more likely to occur than those with large energies. For high temperature (small β), this relative difference is much smaller. As argued by Boltzmann, see the Ehrenfest paper [5], the vast majority of configurations (measure by μ_β) has the property that if you count proportions at which states $a \in \mathcal{A}$ occur, *i.e.,*

$$\pi_\omega(a) = \frac{1}{n}\#\{g \in G \ : \ \omega_g = a\}$$

you find that π_ω is extremely close to q_β. So without interactions, the effect of many particles averages out to q_β.

We can quantify "large majority" using distance d_β of (37). Write

$$U_{\beta,r} = \{\omega \in \Omega \ : \ d_\beta(\pi_\omega) < r\}$$

as the collection of configurations whose emperical distributions π_ω (*i.e.,* frequencies of particles taking the respective states in $\mathcal{A}$) are r-close to q_β.

Theorem 18. *For $0 < r < H(q_\beta) - \beta \int \psi_0 \ dq_\beta$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mu_\beta(\Omega \setminus U_{\beta,r}) = -r,$$

so $\mu_\beta(\Omega \setminus U_{\beta,r}) \sim e^{-nr}$ as $n = \#G$ grows large.

Proof. It is an exercise to check that $H(\mu_\beta) = nH(q_\beta)$. Next, for some configuration $\omega \in \Omega$, we have

$$\begin{aligned}
\log \mu_\beta(\omega) &= -\beta \sum_{g\in G} \psi_0(\omega_g) - \log \mathcal{Z}(\beta) \qquad \text{rewrite } \mathcal{Z}(\beta) \text{ by Corollary 3} \\
&= -\beta n \int \psi_0 \ d\pi_\omega - \left(H(\mu_\beta) - \beta \int \psi \ d\mu_\beta \right) \\
&= -\beta n \int \psi_0 \ d\pi_\omega - n \left(H(q_\beta) - \beta \int \psi_0 \ dq_\beta \right) \\
&= -n \left((H(q_\beta) - \beta \int \psi_0 \ dq_\beta) - (H(\pi_\omega) - \beta \int \psi_0 \ d\pi_\omega) \right) - nH(\pi_\omega) \\
&= nd_\beta(\pi_\omega) - nH(\pi_\omega).
\end{aligned}$$

Every π_ω represents a way to choose $n = \#G$ times from $N = \#\mathcal{A}$ boxes. The order of choosing is not important, only how many are drawn from each box. This can be indicated by a non-negative integer vector $v = (v_a)_{a\in\mathcal{A}}$ where $\sum_{a\in\mathcal{A}} v_a = n$. In fact, $\frac{v}{n}$ indicates the same probability distribution on $\mathcal{A}$ as π_ω. We can compute

$$M(v) := \#\{\omega \in \Omega \ : \ \pi_\omega \text{ leads to } v\} = \frac{n!}{\prod_{a\in\mathcal{A}} v_a!}.$$

Stirling's formula gives $n! \sim \sqrt{2\pi n} \ n^n e^{-n}$, neglecting an error factor that tends to 1 as $n \to \infty$. Thus

$$M(v) \sim \frac{\sqrt{2\pi n} \ n^n e^{-n}}{\prod_{a\in\mathcal{A}} \sqrt{2\pi v_a} \ v_a^{v_a} e^{-v_a}} \sim \sqrt{\frac{2\pi n}{\prod_{a\in\mathcal{A}} 2\pi v_a}} \prod_{a\in\mathcal{A}} \left(\frac{v_a}{n}\right)^{-v_a},$$

and

$$\log M(v) \sim \frac{1}{2}\log\frac{2\pi n}{\prod_{a\in\mathcal{A}} 2\pi v_a} + n\sum_{a\in\mathcal{A}} -\frac{v_a}{n}\log\frac{v_a}{n}.$$

Note that $\frac{v_a}{n} = \pi_\omega(x=a)$, so the dominating term in $\log M(v)$ is just $nH(\pi_\omega)$! The remaining terms, including the one we neglected in our version of Stirling's formula, are $O(\log n)$.

Therefore

$$\begin{aligned}
\frac{1}{n}\log\mu_\beta(\Omega\setminus U_{\beta,r}) &= \frac{1}{n}\log\sum_{\omega\in\Omega\setminus U_{\beta,r}}\mu_\beta(\omega) = \frac{1}{n}\log\sum_{\omega\in\Omega\setminus U_{\beta,r}} e^{-nd_\beta(\pi_\omega)-nH(\pi_\omega)}\\
&\leqslant \frac{1}{n}\log\sum_{v=(v_a)_{a\in\mathcal{A}}} M(v)\cdot e^{-nr-nH(\frac{v}{n})}\\
&= \frac{1}{n}\log\sum_{v=(v_a)_{a\in\mathcal{A}}} e^{nH(\frac{v}{n})+O(\log n)}\cdot e^{-nr-nH(\frac{v}{n})}\\
&= \frac{1}{n}\log\sum_{v=(v_a)_{a\in\mathcal{A}}} e^{-nr+O(\log n)} \leqslant \frac{1}{n}\log n^N e^{-nr+O(\log n)} \to -r
\end{aligned}$$

as $n\to\infty$, where we used in the last line that there are no more than n^N ways of choosing non-negative integer vectors $v=(v_a)_{a\in\mathcal{A}}$ with $\sum_{a\in\mathcal{A}} v_a = n$.

Now for the lower bound, take $r' > r$. For sufficiently large n, we can find some vector $v=(v_a)_{a\in\mathcal{A}}$ such that $r < d_\beta(\frac{v}{n}) < r'$. Therefore

$$\begin{aligned}
\frac{1}{n}\log\mu_\beta(\Omega\setminus U_{\beta,r}) &\geqslant \frac{1}{n}\log\sum_{\omega,\pi_\omega=\frac{v}{n}}\mu_\beta(\omega) = \frac{1}{n}\log\sum_{\omega,\pi_\omega=\frac{v}{n}} e^{-nd_\beta(\pi_\omega)-nH(\pi_\omega)}\\
&\geqslant \frac{1}{n}\log\Big(M(v)\cdot e^{-nr'-nH(\frac{v}{n})}\Big)\\
&\geqslant \frac{1}{n}\Big(nH(\frac{v}{n}) + O(\log n) - nr' - nH(\frac{v}{n})\Big) \to -r'
\end{aligned}$$

as $n\to\infty$. Since $r' > r$ is arbitrary, $\lim_n \frac{1}{n}\log\mu_\beta(\Omega\setminus U_{\beta,r}) = -r$ as claimed. □

References

[1] R.L. Adler, A.G. Konheim, and M.H. McAndrew, *Topological entropy,* Trans. Amer. Math. Soc. **114** (1965), 309-319.

[2] V. I. Arnol'd, A. Avez, *Ergodic Problems in Classical Mechanics,* New York: Benjamin (1968).

[3] V. I. Arnold, *Mathematical Methods of Classical Mechanics,* Springer-Verlag (1989).

[4] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms,* Lect. Math. **470** Spring 1974 and second revised edition 2008.

[5] P. and T. Ehrenfest, *Begriffliche Grundlagen der statistischen Auffassung in der Mechanik,* Enzy. d. Mathem. Wiss. IV, 2, II (1912), 3-90. English translation: *The conceptual foundations of the statistical approach in mechanics,* Cornell Univ. Press (1959).

[6] G. Keller, *Equilibrium states in Ergodic Theory,* London Math. Society, Student Texts **42** Cambridge Univ. Press, Cambridge 1998.

[7] D. Ornstein, *Ergodic theory, randomness and dynamical systems,* Yale Math. Monographs, **5** Yale Univ. New Haven, 1974.

[8] W. Parry, *Intrinsic Markov chains*, Trans. AMS. **112** (1965), 55-65.

[9] P. Walters, *An introduction to ergodic theory,* Springer Verlag (1982).