

Specification and the measure of maximal entropy

Vaughn Climenhaga

June 22, 2020

These are notes for a talk I am giving in Jon Chaika's online working seminar in ergodic theory. The purpose of the talk is to outline [Bowen's proof of uniqueness of the measure of maximal entropy for shift spaces with the specification property](#). Bowen's approach extends much more broadly than this: to non-symbolic systems (assuming expansivity); to equilibrium states for non-zero potential functions (assuming a bounded distortion property); and to non-uniformly hyperbolic systems using the notion of "obstructions to specification and expansivity" developed by Dan Thompson and myself (see some notes [here](#) and [here](#), and videos [here](#)). In these notes, though, I want to give the bare bones of the argument in the simplest possible setting, to make the essential structure as clear as possible. I gave an alternate argument in a [previous post](#); here I am giving Bowen's original argument, although I do not necessarily follow his presentation. I should also point out that this argument differs from the construction of the MME, or more generally the equilibrium state, in [Bowen's monograph](#), which uses the Ruelle operator.

1 Setting and result

Let A be a finite set; the *alphabet*. Then the *full shift* $A^{\mathbb{N}}$ is the set of infinite sequences of symbols from A ; this is a compact metric space with $d(x, y) = e^{-n(x, y)}$, where $n(x, y) = \min\{n : x_n \neq y_n\}$. The *shift map* $\sigma: A^{\mathbb{N}} \rightarrow A^{\mathbb{N}}$ is defined by $\sigma(x)_n = x_{n+1}$. A *shift space* is a closed set $X \subset A^{\mathbb{N}}$ with $\sigma(X) = X$.

Example 1 *The best example to keep in mind through this talk is a topological Markov shift: fix $d \in \mathbb{N}$ and put $A = \{1, \dots, d\}$; then fix a $d \times d$ transition*

matrix T with entries in $\{0, 1\}$, and write

$$i \rightarrow j \text{ if } T_{ij} = 1, \quad i \not\rightarrow j \text{ if } T_{ij} = 0. \quad (1)$$

Define $X = \{x \in A^{\mathbb{N}} : x_n \rightarrow x_{n+1} \text{ for all } n\}$. This is a topological Markov shift (TMS). It can be viewed in terms of a directed graph with vertex set A and edges given by (1): X consists of all sequences that label infinite paths on the graph.

The TMS is mixing or primitive if there is $N \in \mathbb{N}$ such that $(T^N)_{ij} > 0$ for all i, j . Equivalently, the graph is strongly connected and the set of loop lengths on the graph has gcd 1.

Given a shift space X , consider

$$\begin{aligned} \mathcal{M} &= \{\text{Borel probability measures on } X\}, \\ \mathcal{M}_\sigma &= \{\mu \in \mathcal{M} : \sigma_*\mu := \mu \circ \sigma^{-1} = \mu\}, \\ \mathcal{M}_\sigma^e &= \{\mu \in \mathcal{M}_\sigma : \mu \text{ is ergodic}\}. \end{aligned}$$

The set \mathcal{M}_σ of invariant measures is extremely large for a mixing TMS (and more generally for systems with some sort of hyperbolic behavior), and it is important to identify “distinguished” invariant measures. One way of doing this is via the *variational principle*

$$h_{\text{top}}(X) = \sup\{h(\mu) : \mu \in \mathcal{M}_\sigma\}.$$

The next section recalls the definitions of the topological and measure-theoretic entropies in this setting. A measure achieving the supremum is a *measure of maximal entropy* (MME).

We will see that every mixing TMS has a unique MME, via a more general result. Given $n \in \mathbb{N}_0$ and $w \in A^n$, let $[w] = wX \cap X$ be the set of sequences in X that start with the word w (juxtaposition denotes concatenation); call this the *cylinder* of w . Define the *language* of X by

$$\mathcal{L}_n = \{w \in A^n : [w] \neq \emptyset\}, \quad \mathcal{L} = \bigcup_{n \in \mathbb{N}_0} \mathcal{L}_n.$$

Definition 2 X has specification if there is $\tau \in \mathbb{N}_0$ such that for all $v, w \in \mathcal{L}$ there is $u \in \mathcal{L}_\tau$ such that $vuw \in \mathcal{L}$.

Exercise 1 Prove that every mixing TMS has specification.

Theorem 3 (Bowen) If X has specification, then it has a unique MME.

2 Entropies

2.1 Topological entropy

Every word in \mathcal{L}_{m+n} is of the form vw for some $v \in \mathcal{L}_m$ and $w \in \mathcal{L}_n$; thus

$$\mathcal{L}_{m+n} \subset \mathcal{L}_m \mathcal{L}_n \quad \Rightarrow \quad \#\mathcal{L}_{m+n} \leq \#\mathcal{L}_m \#\mathcal{L}_n.$$

This means that the sequence $c_n := \log \#\mathcal{L}_n$ has the *sub-additivity* property

$$c_{m+n} \leq c_m + c_n. \tag{2}$$

Exercise 2 *Prove Fekete's lemma: for any sequence satisfying (2), $\lim_{n \rightarrow \infty} \frac{1}{n} c_n$ exists and is equal to $\inf_n \frac{1}{n} c_n$ (a priori it could be $-\infty$).*

We conclude that the *topological entropy* $h_{\text{top}}(X) := \lim_{n \rightarrow \infty} \frac{1}{n} \log \#\mathcal{L}_n$ exists for every shift space. This quantifies the growth rate of the total complexity of the system.

Exercise 3 *Prove that $h_{\text{top}}(X)$ is the box dimension of X (easy). Then prove that it is also the Hausdorff dimension of X (harder). (Both of these facts rely very strongly on the fact that $d(\sigma x, \sigma y)/d(x, y)$ is globally constant whenever x, y are close; for general systems where the amount of expansion may vary, the definition of topological entropy is more involved and the relationship to dimension is more subtle, although it is worth noting that a [1973 paper of Bowen in TAMS](#) gives a definition of topological entropy that is analogous to Hausdorff dimension.)*

2.2 Measure-theoretic entropy

Recall the motivation from information theory for the definition of entropy: given $p \in (0, 1]$, let $I(p) = -\log p$. This can be interpreted as the information associated to an event with probability p ; note that it is monotonic (the less likely an event is, the more information we gain by learning that it happened) and that $I(pq) = I(p) + I(q)$.

Now define ϕ on $[0, 1]$ by $\phi(p) = pI(p) = -p \log p$ (and $\phi(0) = 0$); this can be interpreted as the *expected* amount of information associated to an event with probability p .

Exercise 4 *Show that ϕ is concave.*

Given $N \in \mathbb{N}$, let $\Delta = \Delta_N = \{\bar{p} = (p_1, \dots, p_N) : p_i \geq 0 \text{ and } \sum p_i \leq 1\}$ be the set of sub-probability vectors with N components. Define

$$H(\bar{p}) = \sum_{i=1}^N \phi(p_i); \quad (3)$$

this can be interpreted as the expected information associated to a collection of mutually exclusive events with probabilities p_1, \dots, p_N .

Exercise 5 Show that $H(\bar{p}) \leq \log N$ for all $\bar{p} \in \Delta_N$, with equality if and only if $p_i = \frac{1}{N}$ for all i .

Given $\mu \in \mathcal{M}_\sigma$, we have for each n a probability vector with $\#\mathcal{L}_n$ components; writing $\mu(w) = \mu([w])$ for convenience, the entropy (expected information) associated to this vector is

$$H_n(\mu) := \sum_{w \in \mathcal{L}_n} \phi(\mu(w)). \quad (4)$$

Lemma 4 $H_{m+n}(\mu) \leq H_m(\mu) + H_n(\mu)$

PROOF:

$$\begin{aligned} H_{m+n}(\mu) &\leq \sum_{v \in \mathcal{L}_m} \sum_{w \in \mathcal{L}_n} \mu(vw) I\left(\frac{\mu(vw)}{\mu(v)} \cdot \mu(v)\right) \\ &= \sum_{w \in \mathcal{L}_n} \sum_{v \in \mathcal{L}_m} \mu(v) \phi\left(\frac{\mu(vw)}{\mu(v)}\right) + \sum_{v \in \mathcal{L}_m} \sum_{w \in \mathcal{L}_n} \mu(vw) I(\mu(v)) \\ &\leq \sum_{w \in \mathcal{L}_n} \phi\left(\sum_{v \in \mathcal{L}_m} \mu(vw)\right) + \sum_{v \in \mathcal{L}_m} \phi(\mu(v)) \\ &= H_n(\mu) + H_m(\mu), \end{aligned}$$

where the first line is by definition, the second is since $I(pq) = I(p) + I(q)$, the third uses concavity of ϕ , and the fourth uses invariance of μ to get $\sum_{v \in \mathcal{L}_m} \mu(vw) = \mu(w)$. \square

This lemma has the following intuitive interpretation: the expected information from the first $m+n$ symbols is at most the expected information from the first m symbols plus the expected information from the next n symbols.

Now Fekete's lemma implies that the following *measure-theoretic entropy* exists:

$$h(\mu) := \lim_{n \rightarrow \infty} \frac{1}{n} H_n(\mu). \quad (5)$$

2.3 Variational principle

Using Exercise 5 we see that $H_n(\mu) \leq c_n = \log \#\mathcal{L}_n$, with equality if and only if $\mu(w) = 1/\#\mathcal{L}_n$ for all $w \in \mathcal{L}_n$. This immediately proves that

$$h(\mu) \leq h_{\text{top}}(X) \text{ for all } \mu \in \mathcal{M}_\sigma, \quad (6)$$

and suggests that in order to have equality we should look for a measure with

$$\mu(w) \approx \frac{1}{\#\mathcal{L}_n} \approx e^{-nh_{\text{top}}(X)} \text{ for all } w \in \mathcal{L}_n. \quad (7)$$

The following makes this more precise.

Definition 5 $\mu \in \mathcal{M}_\sigma$ is a Gibbs measure if there are $h, c, C > 0$ such that for all $n \in \mathbb{N}$ and $w \in \mathcal{L}_n$, we have $ce^{-nh} \leq \mu(w) \leq Ce^{-nh}$.

Now Theorem 3 is a consequence of the following two results, which we prove below.

Theorem 6 If μ is an ergodic Gibbs measure for X , then $h = h_{\text{top}}(X) = h(\mu)$ and μ is the unique MME.

Theorem 7 If X has specification, then it has an ergodic Gibbs measure.

In fact the construction of μ below always gives equality in (6), without relying on the specification property (or obtaining uniqueness), but we will not prove this.

2.4 Convex combinations

Before embarking on the proof of Theorems 6 and 7, we establish a general property of entropy that will be important.

Lemma 8 Given $\bar{p}, \bar{q} \in \Delta_N$ and $s, t \in [0, 1]$ with $s + t = 1$, we have

$$sH(\bar{p}) + tH(\bar{q}) \leq H(s\bar{p} + t\bar{q}) \leq sH(\bar{p}) + tH(\bar{q}) + \log 2. \quad (8)$$

PROOF: The first inequality follows immediately from concavity of ϕ . For the second inequality we first observe that

$$H(s\bar{p}) = \sum_{i=1}^N sp_i I(sp_i) = \sum_{i=1}^N (sp_i I(p_i) + sp_i I(s)) = sH(\bar{p}) + \phi(s) \sum_{i=1}^N p_i. \quad (9)$$

Then we use monotonicity of I to get

$$\begin{aligned} H(s\bar{p} + t\bar{q}) &= \sum sp_i I(sp_i + tq_i) + tq_i I(sp_i + tq_i) \\ &\leq \sum sp_i I(sp_i) + \sum tq_i I(tq_i) \\ &= sH(\bar{p}) + tH(\bar{q}) + \phi(s) \sum p_i + \phi(t) \sum q_i, \end{aligned}$$

where the last equality uses (9) twice; this proves (8). \square

Applying Lemma 8 to the probability vectors associated to two measures $\nu, \mu \in \mathcal{M}_\sigma$, we see that

$$sH_n(\nu) + tH_n(\mu) \leq H_n(s\nu + t\mu) \leq sH_n(\nu) + tH_n(\mu) + \log 2$$

for all n : dividing by n and sending $n \rightarrow \infty$ gives

$$h(s\nu + t\mu) = sh(\nu) + th(\mu). \quad (10)$$

Remark 9 *It is worth mentioning the deeper fact that a version of (10) holds for infinite convex combinations (even uncountable ones given by an integral); this is due to Konrad Jacobs, see Section 9.6 of "Foundations of Ergodic Theory" by Viana and Oliveira.*

3 Gibbs implies uniqueness

To prove Theorem 6, start by observing that the lower Gibbs bound gives

$$1 = \mu(X) = \sum_{w \in \mathcal{L}_n} \mu(w) \geq ce^{-nh} \#\mathcal{L}_n \quad \Rightarrow \quad \#\mathcal{L}_n \leq c^{-1}e^{nh}, \quad (11)$$

and thus $h_{\text{top}}(X) \leq h$. Meanwhile, the upper Gibbs bound gives

$$H_n(\mu) = \sum_{w \in \mathcal{L}_n} \mu(w) I(\mu(w)) \geq \sum_{w \in \mathcal{L}_n} \mu(w) I(Ce^{-nh}) = I(Ce^{-nh}) = nh - \log C,$$

and thus $h(\mu) \geq h$, so we conclude that $h_{\text{top}}(X) = h = h(\mu)$. It remains to show that every $\nu \in \mathcal{M}_\sigma$ with $\nu \neq \mu$ has $h(\nu) < h$.

First observe that by (10), we can restrict our attention to the case when $\nu \perp \mu$. Indeed, given any $\nu \in \mathcal{M}_\sigma$, the Lebesgue decomposition theorem gives $\nu = s\nu_1 + t\nu_2$ for some $\nu_1 \perp \mu$ and $\nu_2 \ll \mu$, and (10) gives $h(\nu) = sh(\nu_1) + th(\nu_2)$. By ergodicity we must have $\nu_2 = \mu$ and thus $s > 0$, so if $h(\nu_1) < h$ then the same is true of $h(\nu)$.

Now consider $\nu \perp \mu$. Then there is a Borel set $D \subset X$ with $\nu(D) = 0$ and $\mu(D) = 1$, and this in turn gives $\mathcal{D}_n \subset \mathcal{L}_n$ such that

$$\nu(\mathcal{D}_n) \rightarrow 0 \text{ and } \mu(\mathcal{D}_n) \rightarrow 1 \text{ as } n \rightarrow \infty. \quad (12)$$

Let $\nu|_{\mathcal{D}_n}$ denote the *normalization* of ν after restricting to words in \mathcal{D}_n , and similarly for $\nu|_{\mathcal{D}_n^c}$. Recall from Fekete's lemma and subadditivity of $H_n(\nu)$ that $\frac{1}{n}H_n(\nu) \geq h(\nu)$ for all n . Then we get

$$\begin{aligned} nh(\nu) &\leq H_n(\nu) = H_n(\nu|_{\mathcal{D}_n} + \nu(\mathcal{D}_n^c)\nu|_{\mathcal{D}_n^c}) \\ &\leq \nu(\mathcal{D}_n)H_n(\nu|_{\mathcal{D}_n}) + \nu(\mathcal{D}_n^c)H_n(\nu|_{\mathcal{D}_n^c}) + \log 2 \\ &\leq \nu(\mathcal{D}_n) \log \#\mathcal{D}_n + \nu(\mathcal{D}_n^c) \log \#\mathcal{D}_n^c + \log 2 \\ &\leq \nu(\mathcal{D}_n) \log(c^{-1}e^{nh}\mu(\mathcal{D}_n)) + \nu(\mathcal{D}_n^c) \log(c^{-1}e^{nh}\mathcal{D}_n^c) + \log 2 \\ &= nh - \log c + \log 2 + \nu(\mathcal{D}_n) \log \mu(\mathcal{D}_n) + \nu(\mathcal{D}_n^c) \log \mu(\mathcal{D}_n^c), \end{aligned}$$

where the second line uses Lemma 8, the third line uses Exercise 5, and the fourth line uses the lower Gibbs bound as in (11). We conclude that

$$n(h(\nu) - h) \leq \log(2/c) + \nu(\mathcal{D}_n) \log \mu(\mathcal{D}_n) + \nu(\mathcal{D}_n^c) \log \mu(\mathcal{D}_n^c),$$

and as $n \rightarrow \infty$ the right-hand side goes to $-\infty$ by (12), which implies that $h(\nu) < h$, completing the proof.

4 Specification implies Gibbs

Now we outline the proof of Theorem 7. This comes in three steps: (1) uniform counting bounds; (2) construction of a Gibbs measure; (3) proof of ergodicity.

4.1 Uniform counting bounds

From now on we write $h = h_{\text{top}}(X)$ for convenience. Fekete's lemma gives $\frac{1}{n} \log \#\mathcal{L}_n \geq h$ for all n , or equivalently $\#\mathcal{L}_n \geq e^{nh}$. This can also be deduced by writing

$$\#\mathcal{L}_{kn} \leq (\#\mathcal{L}_n)^k \quad \Rightarrow \quad \frac{1}{kn} \log \#\mathcal{L}_{kn} \leq \frac{1}{n} \log \#\mathcal{L}_n$$

and sending $k \rightarrow \infty$ so that the left-hand side goes to h (this is basically part of the proof of Fekete's lemma).

To get an upper bound on $\#\mathcal{L}_n$ we need the specification property, which gives a 1-1 map $\mathcal{L}_m \times \mathcal{L}_n \rightarrow \mathcal{L}_{m+n+\tau}$, so that $\#\mathcal{L}_{m+n+\tau} \geq \#\mathcal{L}_m \#\mathcal{L}_n$. Then one can either apply Fekete's lemma to $b_n := -\log \#\mathcal{L}_{n+\tau}$, or observe that

$$\#\mathcal{L}_{k(n+\tau)} \geq (\#\mathcal{L}_n)^k \quad \Rightarrow \quad \frac{1}{k(n+\tau)} \log \#\mathcal{L}_{k(n+\tau)} \geq \frac{1}{n+\tau} \log \#\mathcal{L}_n.$$

Sending $k \rightarrow \infty$ the left-hand side goes to h , and so combined with the lower bound above we get the uniform counting bounds

$$e^{nh} \leq \#\mathcal{L}_n \leq e^{\tau h} e^{nh}. \quad (13)$$

4.2 A Gibbs measure

There is a standard procedure for constructing an MME: let $\nu_n \in \mathcal{M}$ be any sequence of (not necessarily invariant) Borel probability measures such that $\nu_n(w) = 1/\#\mathcal{L}_n$ for all $w \in \mathcal{L}_n$, and then put

$$\mu_n = \frac{1}{n} \sum_{k=0}^{n-1} \sigma_*^k \nu_n = \frac{1}{n} \sum_{k=0}^{n-1} \nu_n \circ \sigma^{-k}.$$

Since \mathcal{M} is weak* compact, there is a weak* convergent subsequence μ_{n_j} .

Exercise 6 Show that $\mu := \lim_{j \rightarrow \infty} \mu_{n_j}$ is σ -invariant ($\mu \circ \sigma^{-1} = \mu$).

The preceding exercise is basically the proof of the Krylov-Bogolyubov theorem, and does not require any properties of the measures ν_n beyond the fact that they are Borel probability measures.

One can prove that μ is an MME whether or not X has specification, but we will use specification to directly prove the stronger Gibbs property.

Given any $w \in \mathcal{L}_m$ and any $\tau \leq k \leq n - \tau$, we bound $\nu_n(\sigma^{-k}[w])$ by estimating how many words in $\#\mathcal{L}_n$ are of the form uvw for some $u \in \mathcal{L}_k$ and $v \in \mathcal{L}_{n-m-k}$. Arguments similar to those in the uniform counting bounds show that

$$\#\mathcal{L}_{k-\tau}\#\mathcal{L}_{n-k-m-\tau} \leq (\# \text{ of such words}) \leq \#\mathcal{L}_k\#\mathcal{L}_{n-k-m},$$

where the first inequality requires specification. Dividing by $\#\mathcal{L}_n$ and using the uniform counting bounds (13) gives

$$\nu_n(\sigma^{-k}[w]) \leq \frac{\#\mathcal{L}_k\#\mathcal{L}_{n-k-m}}{\#\mathcal{L}_n} \leq \frac{e^{\tau h} e^{kh} e^{\tau h} e^{(n-k-m)h}}{e^{nh}} = e^{2\tau h} e^{-mh};$$

using a similar estimate from below we get

$$e^{-3\tau h} e^{-mh} \leq \nu_n(\sigma^{-k}[w]) \leq e^{2\tau h} e^{-mh}. \quad (14)$$

Averaging over all k and sending $n \rightarrow \infty$ along the subsequence n_j gives

$$e^{-3\tau h} e^{-mh} \leq \mu(w) \leq e^{2\tau h} e^{-mh}, \quad (15)$$

so μ is a Gibbs measure.

4.3 Ergodicity

To prove that μ is ergodic, start by fixing $v, w \in \mathcal{L}$, with lengths $|v|$ and $|w|$, respectively. Given $j \gg |v|$ and $k \ll n$, follow the same procedure as above to estimate the number of words in \mathcal{L}_n with the form $xvywz$, where $x \in \mathcal{L}_k$, $y \in \mathcal{L}_{j-|v|}$, and $z \in \mathcal{L}_{n-k-j-|w|}$, and obtain the bounds

$$\frac{\#\mathcal{L}_{k-\tau}\#\mathcal{L}_{j-|v|-\tau}\#\mathcal{L}_{n-k-j-|w|-\tau}}{\#\mathcal{L}_n} \leq \sigma_*^k \nu_n([v] \cap \sigma^{-j}[w]) \leq \frac{\#\mathcal{L}_k\#\mathcal{L}_{j-|v|}\#\mathcal{L}_{n-k-j-|w|}}{\#\mathcal{L}_n}.$$

Averaging over k , sending $n \rightarrow \infty$, and using the uniform counting bounds (13) gives

$$e^{-4\tau h} e^{-|v|h} e^{-|w|h} \leq \mu([v] \cap \sigma^{-j}[w]) \leq e^{3\tau h} e^{-|v|h} e^{-|w|h}.$$

Using the Gibbs bounds (15) gives

$$e^{-8\tau h} \mu(v)\mu(w) \leq \mu([v] \cap \sigma^{-j}[w]) \leq e^{9\tau h} \mu(v)\mu(w). \quad (16)$$

Exercise 7 Given Borel sets $V, W \subset X$, approximate V and W with cylinders and use (16) to get

$$e^{-8\tau h} \mu(V)\mu(W) \leq \underline{\lim}_{j \rightarrow \infty} \mu(V \cap \sigma^{-j}W) \leq \overline{\lim}_{j \rightarrow \infty} \mu(V \cap \sigma^{-j}W) \leq e^{9\tau h} \mu(V)\mu(W).$$

Now if $E \subset X$ is invariant, then taking $V = E$ and $W = E^c$ in Exercise 7, the lower bound gives $e^{-8\tau h} \mu(E)(1 - \mu(E)) = 0$, so $\mu(E)$ is either 0 or 1. This proves ergodicity and completes the proof of Theorem 7.

Remark 10 In fact, the upper bound in Exercise 7 can be used to show that μ is mixing; see Proposition 20.3.6 in Katok and Hasselblatt.