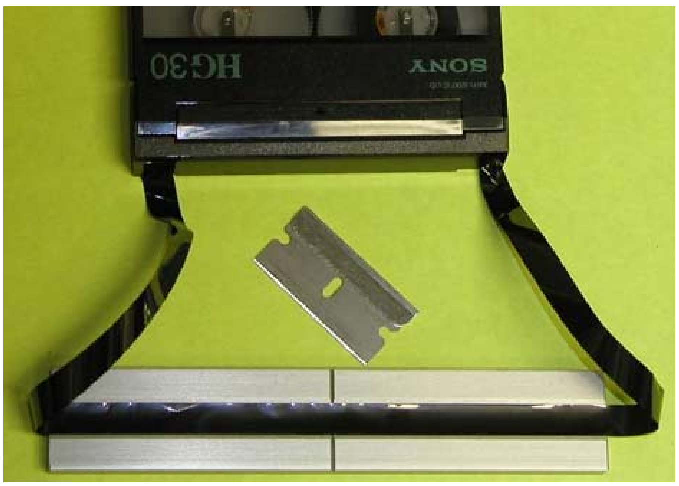
COUPLING AND SPLICING

The purpose of these lectures is to (re-)introduce an important concept from discrete probability to symbolic dynamics, namely *coupling*. We will apply ideas of coupling to create new invariant measures from old, by a process that I will call *splicing*. We will illustrate this with a number of examples from thermodynamic formalism.



A very important previous application of the idea of coupling was due to Don Ornstein and his collaborators in their study of isomorphisms between Bernoulli shifts (and many other related objects).

1. Couplings

Definition. *Let μ be a probability measure on X and μ' be a probability measure on X'. A* coupling *is any measure λ on $X \times X'$ such that $\lambda(A \times X') = \mu(A)$ and $\lambda(X \times B) = \mu'(B)$ for any subsets A of X and B of X'.*

So what?

Theorem 1. *Let (X_n) be an aperiodic irreducible Markov Chain on a finite state space, S. Let π be the stationary distribution. Then $\mathbb{P}(X_n = j | X_0 = i_0)$ converges exponentially to π_j as $n \to \infty$.*

Proof. Define a transition matrix on $S \times S$ by

$$\bar{P}_{(i,i'),(j,j')} = \begin{cases} P_{i,j} & \text{if } i = i' \text{ and } j = j' \\ 0 & \text{if } i = i' \text{ and } j \neq j' \\ P_{i,j} P_{i',j'} & \text{otherwise.} \end{cases}$$

□

For fixed i and i', $\bar{P}_{(i,i'),(j,j')}$ is a coupling of the measures $(P_{i,j})_{j \in S}$ and $(P_{i',j})_{j \in S}$. We now consider a new Markov chain on $S \times S$ with initial distribution $\delta_{i_0} \otimes \pi$ and transition matrix $\bar{P}$. Let the state at time n be (X_n, X'_n). The new Markov chain has the property that X_n and X'_n evolve independently following the Markov chain transition probabilities until they coincide. From that point, they 'stick together' and continue evolving under the Markov chain transition matrix so that $X_n = X'_n$ from that time onward. If one looks at the trajectory $(X_n, X'_n)_{n \geq 0}$, and studies just the first coordinates, $(X_n)_{n \geq 0}$, one sees exactly the original Markov chain, started from i. If one looks at $(X'_n)_{n \geq 0}$, again one sees the original Markov chain, started from a randomly chosen point of S with distribution π. Since the distribution is stationary, we see that $\mathbb{P}(X'_n = j) = \pi_j$ for all n.

By the irreducibility and aperiodicity, we observe that there exist $N \in \mathbb{N}$, $k \in S$ and $\delta > 0$ such that $\mathbb{P}(X_{n+N} = k | X_n = i) > \delta$ for each $i \in S$.

The probability of coalescing in the first N steps is at least δ^2 (a lower bound for the probability that X_N and X'_N are both at k). Given that coalescence failed in the first lN steps, the probability of failing to coalesce in the subsequent N steps is (by the Markov property), at most $1 - \delta^2$. Hence we deduce the inequality

$$\mathbb{P}(X_{lN} \neq X'_{lN}) \leq (1 - \delta^2)^l.$$

Now we put this together:

$$\begin{aligned} |\mathbb{P}(X_{lN} = j) - \pi_j| &= |\mathbb{P}(X_{lN} = j) - \mathbb{P}(X'_{lN} = j)| \\ &\leq \mathbb{P}(X_{lN} \neq X'_{lN}) \\ &\leq (1 - \delta^2)^l. \end{aligned}$$

A dynamical version of this is *joining* of measures. Let μ and ν be two T-invariant measures on a sequence space $A^{\mathbb{Z}}$. A joining of μ and ν is a $T \times T$-invariant measure $\bar{\mu}$ such that $\bar{\mu}(B \times X) = \mu(B)$ and $\bar{\mu}(X \times B) = \nu(B)$.

Just as the set of invariant measures for a continuous transformation of a compact metric space forms a weak*-compact set, so the collection of all joinings of two invariant measures μ and ν forms a compact subset of invariant measures on $X \times X$, written $J(\mu, \nu)$.

Suppose that $T\colon X \to X$ and $S\colon Y \to Y$ are dynamical systems, and that S is a factor of T: there exists a map π such that $\pi \circ T = S \circ \pi$. Then any T-invariant measure, μ *pushes forward* to an S-invariant measure $\mu \circ \pi^{-1}$. The push-forward of a measure should be interpreted as taking a typical point from the measure μ, and applying π to give a typical point of $\mu \circ \pi^{-1}$.

2. Entropy

We make extensive use of entropy, and the intuition that it measures the "amount of information" yielded by a measurement. If one determines which element of a partition $\mathcal{P} = \{B_1, \dots, B_n\}$ an unknown point ω lies in, if one defines the amount of information obtained to be $-\log \mu(B_i)$ if $\omega \in B_i$, this turns out to be consistent with a number of properties that one expects (such as the amount of information learned by recording the outcomes of tossing two coins is twice the amount of information if one coin is tossed). The entropy is the expected amount of information gained, namely $H(\mathcal{P}) = -\sum_B \mu(B) \log \mu(B)$. If one has some prior measurements, the conditional probability of being in B is $\mu(B|\mathcal{A})(\omega)$. Integrating this, one obtains $H(\mathcal{P}|\mathcal{A})$. This is interpreted as the amount of additional information obtained by measuring $\mathcal{P}$ given that the information in $\mathcal{A}$ is already known. One has the beautiful formula $H(\mathcal{P}_1 \vee \mathcal{P}_2) = H(\mathcal{P}_2) + H(\mathcal{P}_1|\mathcal{P}_2)$, which agrees with intuition about measurement of information.

Well known equalities state that for symbolic dynamical systems, $h(\mu) = \lim_{n\to\infty}(1/n)H(\bigvee_{j=0}^{n-1} T^{-j}\mathcal{P}) = H(\mathcal{P}|\bigvee_{j=1}^{\infty} T^{-j}\mathcal{P})$. This says, in the information interpretation, that the entropy of the measure is the limit of the expected amount of information per step measured in n steps, which is the same as the expected amount of information measured in a single step when the entire past is known.

It is straightforward to show that $h(\mu \circ \pi^{-1}) \leq h(\mu)$ for a factor map π.

3. $\bar{d}$ distance

Let μ and ν be two ergodic invariant measures on $A^{\mathbb{Z}}$. The $\bar{d}$-distance from μ to ν is defined by

$$\bar{d}(\mu,\nu) = \inf_{\bar{\mu}\in J(\mu,\nu)} \bar{\mu}\{(x,y) \colon x_0 \neq y_0\}.$$

By compactness, this infimum is attained at some optimal joining. By taking ergodic decompositions, the infimum is attained at an ergodic optimal joining, $\bar{\mu}$. Notice that by the ergodic theorem, for $\bar{\mu}$-a.e. pair $(x,y) \in A^{\mathbb{Z}} \times A^{\mathbb{Z}}$, $(1/N)\#\{j < N \colon x_j \neq y_j\} \to \bar{d}(\mu,\nu)$.

4. Continuous entropy reduction

Let X be a shift space with a safe symbol, '0'. That is: if $x \in X$ then $\tilde{x} \in X$ where $\tilde{x}$ is obtained by replacing any subset of coordinates with 0's.

We use coupling and splicing to show the following:

Theorem 2 (Konieczny, Kupsa, Kwietniak)**.** *Let X be a shift with a safe symbol. Let μ be an ergodic invariant measure on X. Then there is a family μ_t of invariant measures on X such that $\bar{d}(\mu_t,\mu_s) \leq |t-s|$ with $\mu_1 = \mu$, $\mu_0 = \delta_0$ and $t \mapsto h(\mu_t)$ is a non-decreasing function.*

Proof. Let $\bar{X} = X \times [0,1)$ and let $\bar{T} = T \times R_\alpha$, where R_α is a rotation through the irrational angle α. Let $\bar{\mu} = \mu \times \lambda$, where λ is Lebesgue measure and define a family of maps $\Phi_t \colon \bar{X} \to \bar{X}$ by $\Phi_t(x,y) = (z,y)$ where

$$z_n = \begin{cases} x_n & \text{if } R_\alpha^n(y) < t; \\ 0 & \text{otherwise,} \end{cases}$$

so that $\Phi_1(x,y) = (x,y)$ and $\Phi_0(x,y) = (\mathbf{0},y)$. Define $\bar{\mu}_t = \bar{\mu} \circ \Phi_t^{-1}$ and $\mu_t = \bar{\mu}_t \circ \pi_1^{-1}$, where π_1 is the projection onto the first coordinate. That is, we see μ_t as the measure on the sequences obtained by starting with a μ-typical element of X and then replacing a fraction $1-t$ of the symbols with 0's. Notice that if $s < t$, then $\Phi_s \circ \Phi_t = \Phi_s$, so that $\bar{\mu}_s$ is a factor of $\bar{\mu}_t$.

Let $\mathcal{P}_1$ be the partition of $\bar{X}$ given by $\{[a] \times [0,1) \colon a \in A\}$ and $\mathcal{P}_2$ be the partition $\{X \times [0,\alpha), X \times [\alpha,1)\}$. Then $h(\mu_t) = h(\bar{\mu}_t,\mathcal{P}_1) \leq h(\bar{\mu}_t,\mathcal{P}_1 \vee \mathcal{P}_2) = h(\bar{\mu}_t) \leq h(\bar{\mu}_t,\mathcal{P}_1) + h(\bar{\mu}_t,\mathcal{P}_2) = h(\bar{\mu}_t,\mathcal{P}_1) = h(\mu_t)$, so that $h(\mu_t) = h(\bar{\mu}_t)$ for each t.

Since $h(\bar{\mu}_s) \le h(\bar{\mu}_t)$ for $s < t$, it follows that $h(\mu_s) \le h(\mu_t)$ for $s < t$ as required.

Finally, define $\Psi_{s,t} \colon X \times [0,1) \to X \times X$ by

$$\Psi_{s,t}(x,y)_n = (\pi_1(\Phi_s(x,y)), \pi_1(\Phi_t(x,y))).$$

Now we see immediately from the definitions that $\bar{\nu}\bar{\mu} \circ \Psi_{s,t}$ is a joining of μ_s and μ_t.

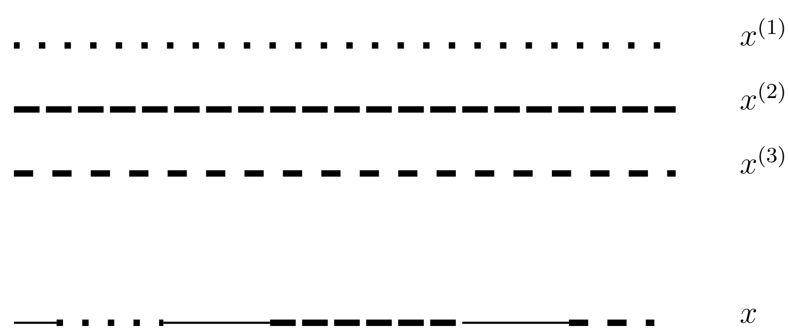
$$\begin{aligned}\bar{\nu}(\{(u,v) \colon u_0 \ne v_0\}) &= \bar{\mu}(\Psi_{s,t}^{-1}\{(u,v) \colon u_0 \ne v_0\}) \\ &\le \bar{\mu}(\{(x,y) \colon s \le y < t\} = t - s.\end{aligned}$$

□

5. Specification

Recall that if X is a subshift and ϕ is a continuous function, then the *pressure* of ϕ is $\sup_{\mu \in \mathcal{M}_{\text{inv}}}(h(\mu) + \int \phi \, d\mu)$. In fact, a compactness argument shows that there is at least one invariant measure for which the supremum is attained: $\mu \mapsto h(\mu)$ is upper semi-continuous with respect to the weak*-topology; and $\mu \mapsto \int \phi \, d\mu$ is continuous, so that the sum is upper semi-continuous. Upper semi-continuous functions on compact sets attain their suprema by the standard argument. An invariant measure for which the supremum is attained is called an *equilibrium state*.

A subshift X is said to satisfy *specification* if there exists an ℓ (the *specification distance*) such that for all $x, y \in X$, there exists a point $z \in X$ such that $z_n = x_n$ for all $n < -\ell$ and $z_n = y_n$ for all $n \ge 0$. Applying the specification condition inductively, it is not hard to see that if $x^{(1)}, \ldots, x^{(k)}$ are k points in X and $I^{(1)}, \ldots, I^{(k)}$ are k subintervals, there is a point $x \in X$ such that $x_n = x_n^{(i)}$ for all $n \in I^{(i)}$. This is illustrated in the figure.



Applying induction and compactness, one may also do this for countably many points. However, there is a serious drawback from the thermodynamic formalism point of view. One would like to be able to obtain an invariant measure by splicing together pieces. The pieces that are used for filling may in general depend on the order in which the filling is done. This will break invariance of any measure created in this way.

5.1. **The specification interpolation method.** A solution (the solution?) is to perform a hierarchical filling as specified by an external process: use an i.i.d. process (for example) to assign each gap a positive integer: assign the integer n with probability 2^{-n} for example. Then iteratively fill in all level n gaps using the lexicographically smallest filler. In the end, all gaps are filled, and invariance is maintained!

Let's give a more precise description of this in a special case. Let X be a subshift with specification length ℓ and let μ be an ergodic invariant measure on X. Let W be a legal word of length n in X and let ν be a mixing invariant measure on $\{0,1\}^{\mathbb{Z}}$, in which 1's are separated by at least $2\ell+n+1$ (mixing so that the product $\mu\times\nu$ is ergodic). Finally, let λ be an i.i.d. measure on $\mathbb{N}^{\mathbb{Z}}$ where the symbol n appears with frequency 2^{-n}. The master space is then $\bar{X}=X\times\{0,1\}^{\mathbb{Z}}\times\mathbb{N}^{\mathbb{Z}}$, equipped with the measure $\bar{\mu}=\mu\times\nu\times\lambda$. We define a sequence of maps $\Phi_n\colon\bar{X}\to\bar{A}^{\mathbb{Z}}$, where $\bar{A}=A\cup\{\star\}$.

First, define

$$\Phi_0(x,y,z)_n=\begin{cases}W_i & \text{if } y_{n-i}=1 \text{ for some } 0\le i<|W|;\\ \star & \text{if } y_{n+i}=1 \text{ for some } 0<i\le\ell;\\ \star & \text{if } y_{n-i}=1 \text{ for some } |W|\le i<|W|+\ell\\ x_n & \text{otherwise.}\end{cases}$$

That is, $\Phi_0(x,y,z)$ is a sequence in $\bar{A}^{\mathbb{Z}}$, in which at each 1 in the y sequence $|W|$ symbols of x have been replaced by the word W; and ℓ $\star$'s have been added on either side. We call the blocks of $\star$'s *star blocks*. The gap between 1's ensures that the star blocks are non-overlapping and non-contiguous. A key property of this map is that $\Phi_0(\sigma(x),\sigma(y),\sigma(z))=\sigma(\Phi_0(x,y,z))$. We call this property *shift-commuting*.

We then define the maps $\Phi_n(x,y,z)$. For each star block, its *order* is the z label of the first $\star$ in the block. The map Φ_n will then fill in all of the nth order star blocks. More precisely, in $\Phi_{n-1}(x,y,z)$, there are (with probability 1) infinitely many star blocks, of all orders.

Let $a_0^{(n)}(x,y,z)$ be the smallest positive integer such that (1) a star block of order n starts at $a_0^{(n)}(x,y,z)$; and (2) there is a star block of higher order starting at some $0 \le j < a_0^{(n)}(x,y,z)$. If $a_i^{(n)}(x,y,z)$ is defined, then let $b_i^{(n)}(x,y,z)$ be the position of the first star block of order higher than n that starts after $a_i^{(n)}(x,y,z)$ and let $a_{i+1}^{(n)}(x,y,z)$ be the first star block of order n that starts after $b_i^{(n)}(x,y,z)$. The same inductive process is also run along the negative integers, so that all star blocks occurring in $[a_i^{(n)}(x,y,z), b_i^{(n)}(x,y,z))$ are of order n; these blocks cover all star blocks of order n; and there is at least one star block of higher order between any pair.

By the specification property, there is at least one way to replace the star blocks in $[\Phi_{n-1}(x,y,z)]_{a_i^{(n)}(x,y,z)}^{b_i^{(n)}(x,y,z)-1}$ in such a way that the new block belongs to the language of X. We choose the lexicographically smallest such completion in each star block simultaneously, and denote the point obtained by $\Phi_n(x,y,z)$. The map Φ_n is defined on a set of measure 1 (where there are infinitely many star blocks of all orders in both halves of the coordinates) and is shift-commuting (but not continuous).

Notice that $\Phi_n(x,y,z)$ converges to a point $\Phi(x,y,z)$ with no $\star$'s. Also, every block of $\Phi(x,y,z)$ appears as a block of $\Phi_n(x,y,z)$ for sufficiently large n, but every $\star$-free block of $\Phi_n(x,y,z)$ belongs to the language of X, so that every block of $\Phi(x,y,z)$ belongs to the language of X. Hence, since X is a subshift, $\Phi(x,y,z) \in X$. Since the Φ_n are shift-commuting, so is the limit, Φ. In particular, $\Phi(x,y,z)$ is a point of X agreeing with x, except that at every n such that y_n=1, a segment of x is replaced by a W, and the l surrounding symbols are replaced in order to ensure that the resulting point lies in X.

Since Φ is a shift-commuting measurable map from $X \times \{0,1\}^{\mathbb{Z}} \times \mathbb{N}^{\mathbb{Z}}$ to X, and $\mu \times \nu \times \lambda$ is a shift invariant measure, the push-forward, $(\mu \times \nu \times \lambda) \circ \Phi^{-1}$ is a shift-invariant measure on X.

6. A theorem of Bowen

In this section, we describe a coupling and splicing proof of a theorem of Bowen.

Theorem 3. *(Bowen) Let X be a subshift with specification, and let ϕ be a function with the property (the* Bowen condition*) that there exists an M such that for all $k > 0$, if $x, y \in X$ satisfy $x_0^{k-1} = y_0^{k-1}$, then*

$|S_k\phi(x) - S_k\phi(y)| \le M$, where $S_k\phi(x) = \sum_{i=0}^{k-1} \phi(T^i x)$. Then there is a unique equilibrium state.

The proof goes via a number of steps. The first of these is to show that any equilibrium state has positive entropy.

Lemma 4. *Let X be a subshift with specification on an alphabet A with $|A| \ge 2$. and let ϕ satisfy Bowen's condition. Then any equilibrium state has positive entropy.*

Proof. Suppose for a contradiction that μ is an equilibrium state, and the μ has 0 entropy. Let ℓ be the specification distance.

Let a, b be two distinct symbols in A. For a small parameter $\epsilon > 0$, consider the i.i.d. measure on symbols 0, a and b, where a and b appear with frequency $\epsilon/2$. Let Φ be the map from $\{0, a, b\}^{\mathbb{Z}}$ to itself that replaces any a or b within $2\ell + 1$ of any other a or b with a 0. We let the push-forward of this i.i.d. measure under Φ be ν_ϵ, so that ν_ϵ has zeros with frequency $1 - \epsilon + O(\epsilon^2)$ and a's and b's with frequency $\frac{\epsilon}{2} - O(\epsilon^2)$.

We then apply the method specification interpolation described above to obtain a new shift invariant measure μ' from $\mu \times \nu_\epsilon \times \lambda$, where the push-forward map Ψ ensures that

$$\Psi(x, y, z)_n = \begin{cases} x_n & \text{if } y_{n-\ell} = \ldots = y_{n+\ell} = 0; \\ y_n & \text{if } y_n = a \text{ or } b. \end{cases}$$

We now estimate $\int \phi \, d\mu'$ and $h(\mu')$ to show that for sufficiently small $\epsilon > 0$, $h(\mu') + \int \phi \, d\mu' > h(\mu) + \int \phi \, d\mu$.

First, we bound the entropy from below. Let $\mathcal{P}$ be the time zero partition and let $L = 2\ell + 1$. We consider $H(\bigvee_{j=0}^{N-1} \sigma^{-jL}\mathcal{P})$. A simple counting argument shows that

$$\begin{aligned} (1/LN) H_{\mu'} \left(\bigvee_{j=0}^{LN-1} \sigma^{-j}\mathcal{P} \right) &\ge (1/LN) H_{\mu'} \left(\bigvee_{j=0}^{N-1} \sigma^{-jL}\mathcal{P} \right) \\ &\gtrsim (1/L)\mathcal{H}(\tfrac{\epsilon}{2}, \tfrac{\epsilon}{2}, 1 - \epsilon) \\ &= \Omega(-\epsilon \log \epsilon), \end{aligned}$$

where $A = \Omega(B)$ means $A \ge cB$ for all small enough ϵ; and $\mathcal{H}(p_1, \ldots, p_n) = \sum_i -p_i \log p_i$. In particular $h(\mu') \ge -c\epsilon \log \epsilon$.

Now we compare $\int \phi \, d\mu'$ to $\int \phi \, d\mu$. For this, we use Bowen's condition, and we also use the map Ψ defined above to build a coupling of μ and

μ'. Namely we define the map $\bar{\Psi}(x,y,z) = (x, \Psi(x,y,z))$ and let $\bar{\mu}$ be the push-forward of $\mu \times \nu_\epsilon \times \lambda$ under $\bar{\Psi}$. The two points x and $\Psi(x,y,z)$ differ only on blocks of size $2\ell+1$ surrounding places where y is non-zero. Hence for $\bar{\mu}$-a.e. pair (x,x'), the point x is a generic point for μ; the point x' is a generic point for μ' and they disagree on blocks of length $2\ell+1$ that occur with frequency ϵ.

We are then able to use the Bowen condition, to show that the difference in the orbit sums of ϕ along x and x' disagree by at most M on the common orbit segments and by at most $2(2\ell+1)\|\phi\|$ on the interpolation segments. Hence for $\bar{\mu}$-almost every pair (x,x'), $|S_N\phi(x) - S_N\phi(x')| \lesssim (\epsilon N)(M + (2\ell+1)\|\phi\|)$.

In particular, since for $\bar{\mu}$-a.e. (x,x'), x is generic for μ and x' is generic for μ', we see on dividing by N (recalling that M, ℓ and $\|\phi\|$ are constant) that

$$\left| \int \phi\, d\mu - \int \phi\, d\mu' \right| = O(\epsilon)$$

Combining the two estimates, we see that for small ϵ, $h(\mu') + \int \phi\, d\mu' > h(\mu) + \int \phi\, d\mu$, giving the required contradiction. □

We now invoke a theorem of Ornstein and Weiss.

Theorem 5 (Ornstein and Weiss). *Let μ be an ergodic measure on a subshift with $h(\mu) = h > 0$. Then for μ-a.e. $x \in X$,*

$$\lim_{n\to\infty} \tfrac{1}{n} \log \min\{j > 0 \colon x_0^{n-1} = x_j^{j+(n-1)}\} = h.$$

As a corollary, we obtain

Corollary 6. *Let μ be an ergodic measure of positive entropy. Then for all $\epsilon > 0$, there exists an $n_0 > 0$ such that for all $n > n_0$,*

$$\mu\{x \colon x_0^{j-1} = x_{n-j}^{n-1} \text{ for some } \tfrac{n}{3} \le j < n\} < \epsilon.$$

Proposition 7. *Let μ be an ergodic measure of positive entropy. Then for all $\epsilon > 0$, there exists an $n_0 > 0$ such that for all $n \ge n_0$ and all $x \in X$,*

$$\mu\left(\bigcup_{j=0}^{n-1} \sigma^{-j}([x_0 \dots x_{n/3}]) \right) < \epsilon.$$

Now let no_n (for non-overlapping) be the collection of n-words W for which $W_0^{j-1} \ne W_{n-j}^{n-1}$ for all $\frac{n}{3} \le j < n$.

We define a relation $\xrightarrow{\mu}$ on no_n by $W \xrightarrow{\mu} W'$ if

$$\frac{\mu\left([W] \cap \sigma^{-(n+\ell)} \bigcup_{j=n/3}^{n} [W'_{n-j} \dots W'_{n-1}]\right)}{\mu([W])} < \tfrac{1}{4}; \text{ and}$$
$$\frac{\mu\left([W] \cap \bigcup_{j=n/3}^{n-1} \sigma^{\ell+j} [W'_0 \dots W'_{j-1}]\right)}{\mu([W])} < \tfrac{1}{4}.$$

that is, the probability of seeing a long suffix of W' starting at coordinate $n+\ell$ conditional on being in $[W]$ is small; and also that the probability of seeing a long prefix of W' ending at coordinate $-\ell$ conditional on being in $[W]$ is small.

Similarly, write $W \xleftrightarrow{\mu} W'$ if $W \xrightarrow{\mu} W'$ and $W' \xrightarrow{\mu} W$.

Lemma 8. *There exists a constant M, such that for all sufficiently large n, if W and W' belong to no_n and $W \xleftrightarrow{\mu} W'$, then*

$$\frac{1}{M}\frac{e^{\Phi(W)}}{e^{\Phi(W')}} \le \frac{\mu([W])}{\mu([W'])} \le M\frac{e^{\Phi(W)}}{e^{\Phi(W')}},$$

where $\Phi(W)$ denotes $S_n\phi(x)$ for an arbitrarily chosen $x \in [W]$ (this quantity does not vary by more than an additive constant if x is changed to another element of $[W]$).

Proof Sketch. Suppose $W \xleftrightarrow{\mu} W'$. Then we create a new measure by randomly converting some W to W' and filling in using the specification interpolation method.

We define pm to be the set of *potential marks*,

$$\mathsf{pm} = [W] \setminus \left(\sigma^{-(n+l)} \bigcup_{j=n/3}^{n} [W'_{n-j} \dots W'_{n-1}] \cup \bigcup_{j=n/3}^{n-1} \sigma^{l+j} [W'_0 \dots W'_{j-1}]\right).$$

That is, these are places where converting a W to a W' will not inadvertently create additional copies of W', even when the specification interpolation is done. By assumption, we have $\mu(\mathsf{pm}) > \frac{1}{2}\mu([W])$.

To construct the new measure, let ν be the Bernoulli measure on $\{0,1\}^{\mathbb{Z}}$ where $\nu([1]) = \epsilon$ (and ϵ is a parameter to be determined later). Now define a map $\Psi_1 \colon X \times \{0,1\}^{\mathbb{Z}} \to X \times \{0,1\}^{\mathbb{Z}}$ by

$$\Psi_1(x,u)_k = \begin{cases} (x_k, 1) & \text{if } u_k = 1,\ \sigma^k(x) \in \mathsf{pm} \text{ and } u_j = 0 \text{ if } |j-k| < n+\ell \\ (x_k, 0) & \text{otherwise} \end{cases}$$

This map just removes marks that are too close together. The push-forward of $\mu \times \nu$ under Φ_1 is called μ_{mark}.

One can show that for small ϵ, $h(\mu_{\text{mark}}|\mu) > \mu(\mathsf{pm})(-\epsilon \log \epsilon)$. (To see this, notice that the right side under-estimates the entropy by $-\mu(\mathsf{pm})(1-\epsilon)\log(1-\epsilon) = \Omega(\epsilon)$, but over-estimates the entropy by ignoring collisions between marks, leading to an over-estimate of size $O(\epsilon^2|\log \epsilon|)$. This implies that for small ϵ

$$h(\mu_{\text{mark}}) \geq h(\mu) - \epsilon\mu(\mathsf{pm}) \log \epsilon. \tag{1}$$

We then use the specification interpolation method to produce a measure $\mu_{\text{mark,switch}}$ on $X \times \{0,1\}^{\mathbb{Z}}$ in which the marked W's are replaced by W''s and the interpolation is done to obtain an invariant measure. The marks are preserved. Finally, we project onto the first coordinate to obtain a measure μ_{switch} on X, where the marks are erased.

Simple estimates show

$$h(\mu_{\text{mark}}|\mu_{\text{mark,switch}}) \leq 2\ell\epsilon\mu(\mathsf{pm}) \log |\mathcal{A}|, \tag{2}$$

(as to reconstruct the point before the replacement was made, it is only necessary to recover the ℓ symbols on each side of the replacement).

We now estimate $h(\mu_{\text{mark,switch}}|\mu_{\text{switch}})$. This amounts to taking a point from the measure μ_{switch} (in which some W's in the original sequence have been replaced by W''s) and 'guessing' where the replacements took place. We consider the case in which W and W' cannot overlap (so no W' are destroyed in the replacement process). In case overlaps are possible, the estimates below still hold. The frequency of W' in the new sequence is $\mu(W') + \epsilon\mu(\mathsf{pm}) + O(\epsilon^2)$. Write $a = \mu(W')$ and $\delta = \mu_{\text{switch}}(W') - \mu(W') = \epsilon\mu(\mathsf{pm}) + O(\epsilon^2)$. For each occurrence of W', by concavity of entropy, the expected amount of information is at most $H(a/(a+\delta), \delta/(a+\delta))$, where $H(p, 1-p) = -p\log p - (1-p)\log(1-p)$.

This yields

$$\begin{aligned} h(\mu_{\text{mark,switch}}|\mu_{\text{switch}}) &\leq (a+\delta)\log(a+\delta) - a\log a - \delta\log\delta \\ &\leq \delta(1+\log a) - \delta\log\delta + C\delta \\ &= \epsilon\mu(\mathsf{pm})(\log\mu(W') - \log(\epsilon\mu(\mathsf{pm})) + C\epsilon\mu(\mathsf{pm}) \\ &= \epsilon\mu(\mathsf{pm})(-\log\epsilon - \log(\mu([W])/\mu([W'])) + C\epsilon\mu(\mathsf{pm}). \end{aligned}$$

Combining with (2), we obtain

$$h(\mu_{\text{mark}}|\mu_{\text{switch}}) \leq \epsilon\mu(\mathsf{pm})(-\log\epsilon - \log(\mu([W])/\mu([W'])) + C\epsilon\mu(\mathsf{pm}).$$

or

$$h(\mu_{\text{switch}}) \geq h(\mu_{\text{mark}}) - \epsilon\mu(\mathsf{pm})(-\log\epsilon - \log(\mu([W])/\mu([W']))) - C\epsilon\mu(\mathsf{pm}).$$

Combining with (1), we see

$$h(\mu_{\text{switch}}) \geq h(\mu) + \epsilon\mu(\mathsf{pm})\log(\mu([W])/\mu([W'])) - C\epsilon\mu(\mathsf{pm})$$

for a constant C that does not depend on n, W or W'.

As in the proof of Lemma 4, we see that for sufficiently small ϵ,

$$\int \phi\, d\mu_{\text{switch}} \geq \int \phi\, d\mu + \epsilon\mu(\mathsf{pm})(\Phi(W') - \Phi(W)) - C\epsilon\mu(\mathsf{pm}).$$

Hence

$$P(\mu_{\text{switch}}) \geq P(\mu) + \epsilon\mu(\mathsf{pm})\left(\log\left(\frac{\mu([W])e^{-\Phi(W)}}{\mu([W'])e^{-\Phi(W')}}\right) - C\right),$$

where $P(\mu) = h(\mu) + \int \phi\, d\mu$. Since the constant C is independent of n, W and W', if $\frac{\mu([W])}{\mu([W'])} > e^C e^{\Phi(W)-\Phi(W')}$, then $P(\mu_{\text{switch}}) > P(\mu)$ for small ϵ, giving the required contradiction. □

Lemma 9. *Let $\overset{\mu}{\longleftrightarrow}$ be the relation on no_n defined above. For any U and W in no_n, there is a $V \in \mathsf{no}_n$ such that $U \overset{\mu}{\longleftrightarrow} V \overset{\mu}{\longleftrightarrow} W$.*

Proof sketch. Suppose that $U, W \in \mathsf{no}_n$ have the property that there is no $V \in \mathsf{no}_n$ such that $U \overset{\mu}{\longleftrightarrow} V \overset{\mu}{\longleftrightarrow} W$. Then for each V, one of $U \overset{\mu}{\not\longrightarrow} V$, $V \overset{\mu}{\not\longrightarrow} W$, $W \overset{\mu}{\not\longrightarrow} V$ or $V \overset{\mu}{\not\longrightarrow} U$ holds.

We show that the measure of the union of the cylinder sets $[V]$ for which one of these conditions is satisfied is small.

To show that the union of the set of $[V]$ such that $U \overset{\mu}{\not\longrightarrow} V$ is small, notice that the conditional probability of seeing a fragment of such a V to the right (or left) of U is large, so that there cannot be more than 8 V's that don't contain a common fragment of length $n/3$. Choose a fragment of length $n/3$, and removing all V's that contain it. Then repeat. We see that we will obtain at most 8 such $n/3$ fragments. Now appealing to Proposition 7, we see that $\mu(\bigcup_{U \overset{\mu}{\not\rightarrow} V}[V]) < 8\epsilon$. Similarly $\mu(\bigcup_{W \overset{\mu}{\not\rightarrow} V}[V]) < 8\epsilon$.

To estimate $\mu(\bigcup_{V \overset{\mu}{\not\rightarrow} U}[V])$, notice that this entails a large entropy drop (the conditional probability of seeing a fragment of the fixed word U given that one is in V is at least $\frac{1}{4}$), so that the entropy conditioned on being in V is small. By the Shannon-Macmillan-Breiman theorem, this happens on a small subset of the space.

A similar estimate applies to $\mu(\bigcup_{V \overset{\mu}{\not\to} W}[V])$. Since the measure of the combined bad cylinder sets is smaller (for large n) than the measure of no_n, we see there exist $V \in \mathsf{no}_n$ such that $U \overset{\mu}{\longleftrightarrow} V \overset{\mu}{\longleftrightarrow} W$. □

Combining the last two lemmas, we see there exists an M such that for all large n and all $U, W \in \mathsf{no}_n$,

$$\frac{1}{M}\frac{e^{\Phi(U)}}{e^{\Phi(W)}} \le \frac{\mu([U])}{\mu([W])} \le M\frac{e^{\Phi(U)}}{e^{\Phi(W)}},$$

If ν were a distinct ergodic equilibrium state, then it would satisfy the same inequality for all large n. Also both μ and ν would satisfy $\mu(\bigcup_{V \in \mathsf{no}_n}[V]) \to 1$ as $n \to \infty$. It is then straightforward to show that $\mu([V]/\nu([V])$ is uniformly bounded for cylinder sets in a large part of the space. It then follows that μ and ν are equivalent measures, and so are equal.

7. Average Sample Complexity

Average Sample Complexity is a notion introduced by Karl Petersen and collaborators, related to the notion of *intricacy*, introduced by Edelman, Sporns and Tononi in neurophysiology as a measure of the self-dependence of a sequence of measurements.

Let μ be a shift-measure on a shift space X and let the coordinate partition be $\mathcal{P}$. If A is a finite subset of $\mathbb{N}$, let $\mathcal{P}_A$ denote $\bigvee_{j \in A} \sigma^{-j}\mathcal{P}$.

Then the *average sample complexity* (ASC) of μ is defined by

$$\mathrm{ASC}(\mu) = \lim_{N\to\infty} \frac{1}{N}\frac{1}{2^N} \sum_{S \subset \{0,\ldots,N-1\}} H_\mu(\mathcal{P}_S),$$

where $H_\mu(\cdot)$ denotes the entropy of a partition with respect to μ as usual. Here, the 2^N is just normalizing over the number of sets being summed over, while the N is the standard normalization appearing in the entropy.

It is not hard to see from the inequality $H(\mathcal{P} \vee \mathcal{Q}) \le H(\mathcal{P}) + H(\mathcal{Q})$ and T-invariance of μ that

$$\frac{1}{2^{N+M}} \sum_{S \subset [N+M-1]} H_\mu(\mathcal{P}_S) \le \frac{1}{2^N} \sum_{S \subset [N]} H_\mu(\mathcal{P}_S) + \frac{1}{2^M} \sum_{S \subset [M]} H_\mu(\mathcal{P}_S),$$

where $[N]$ denotes $\{0, 1\ldots, N-1\}$. Hence the limit in the definition of average sample complexity exists.

There is also a topological notion of average sample complexity, $\mathrm{ASC}_{\mathrm{top}}$ given by

$$\mathrm{ASC}_{\mathrm{top}}(X) = \lim_{N\to\infty} \frac{1}{N}\frac{1}{2^N} \sum_{S\subset[N]} \log \mathcal{N}(S),$$

where $\mathcal{N}(S)$ denotes the number of elements of $\{x|_S \colon x \in X\}$.

Unlike for regular entropy, there is no variational principle, and even for the golden mean shift of finite type, the topological average sample complexity is strictly larger than the maximum measure-theoretic average sample complexity.

We briefly develop a way of representing the measure-theoretic average sample complexity that is reminiscent of the formula $h(\mu) = H(\mathcal{P}|\bigvee_{n=1}^{\infty} T^{-n}\mathcal{P})$. Let $\mathbb{P}$ denote the measure on subsets of $\mathbb{Z}^{0-}$ in which each element is independently present or absent with probability $\frac{1}{2}$. Given a subset S of $\mathbb{Z}^{0-}$, write $S \in T_0$ if $0 \in S$ and write S^- for $S \setminus \{0\}$. As usual, we write $\mathcal{P}$ for the partition of X into cylinder sets defined by the symbol in the 0th position, and we write $\mathcal{P}_S$ for the σ-algebra $\bigvee_{j\in S} \sigma^{-j}(\mathcal{P})$.

Then

$$\mathrm{ASC}(\mu) = \int_{T_0} H_\mu(\mathcal{P}_0|\mathcal{P}_{S^-})\, d\mathbb{P}(S).$$

The proof follows from the facts that (1) $H(\mathcal{P} \vee \mathcal{Q}) = H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q})$; (2) $H(\mathcal{P}|\mathcal{Q})$ decreases as $\mathcal{Q}$ increases; and (3) the monotone convergence theorem.

Theorem 10. *Let X be a shift of finite type with a safe symbol. A measure of maximal ASC must be fully supported.*

Lemma 11. *Let $(p_i)_{i=1}^n$ and $(q_i)_{i=1}^n$ be two probability vectors satisfying $q_i \geq (1-\delta)p_i$. Then $H(q) \geq H(p) - \delta(H(p)+1)$.*

Proof. Let (p_i) and (q_i) be as in the statement of the lemma. As they each sum to 1, we have

$$\sum_{q_i>p_i} (q_i - p_i) = \sum_{p_i>q_i} (p_i - q_i) \leq \sum_{p_i>q_i} \delta p_i \leq \delta. \tag{3}$$

Letting $\phi(x) = -x \log x$, we see that $\phi(x)/x$ is a decreasing function. So if $(1-\delta)p_i \le q_i \le p_i$, then $\phi(q_i) \ge (1-\delta)\phi(p_i)$. Now

$$\begin{aligned} H(q) = \sum_{i=1}^{n} \phi(q_i) &= \sum_{q_i<p_i} \phi(q_i) + \sum_{q_i \ge p_i} \phi(q_i) \\ &\ge (1-\delta) \sum_{q_i<p_i} \phi(p_i) + \sum_{q_i \ge p_i} (\phi(p_i) - (q_i - p_i)) \\ &\ge (1-\delta)H(p) - \delta, \end{aligned}$$

where, for the second term, we used (3) and the fact that $\phi'(x) \ge -1$ for all $x \in [0,1]$. ☐

Proof of Theorem 10. Let μ be a measure of maximal average symbolic complexity. Suppose for a contradiction that μ is not fully supported. That is: there is a word W such that $\mu([W]) = 0$. By extending W if necessary, we may assume that W begins and ends with a safe symbol.

As usual, we build another measure on X by techniques of coupling and splicing.

First, let ν be the Bernoulli measure on $\{0,1\}^{\mathbb{Z}}$ in which 1's appear with frequency ϵ (ϵ is a parameter to be adjusted later).

Let ℓ be the length of the word W. We first modify ν by removing any two 1's in the second component that are separated by less than ℓ. This gives a measure ν_2 on $\{0,1\}^{\mathbb{Z}}$. Now let $\bar{\mu} = \mu \times \nu_2$.

We then define a map $\Phi\colon X \times \{0,1\}^{\mathbb{Z}} \to X$ by

$$\Phi(x,y)_n = \begin{cases} W_i & \text{if } y_{n-i} = 1 \text{ for some } 0 \le i < \ell; \\ x_i & \text{otherwise.} \end{cases}$$

Notice that since ν_2-a.e. y has no two 1's separated by less than ℓ, there is no ambiguity in the definition of Φ. Also, since W begins and ends with a safe symbol, $\Phi(x,y) \in X$ for $\bar{\mu}$-a.e. (x,y). Let $\mu' = \bar{\mu} \circ \Phi^{-1}$, the push-forward of $\bar{\mu}$ by Φ. As usual, we visualize typical points of μ' as points of μ in which occasional copies of W have been spliced in.

We now aim to show that for sufficiently small ϵ, we have $\mathrm{ASC}(\mu') > \mathrm{ASC}(\mu)$, which will contradict the maximality of the average symbolic complexity of μ, and complete the proof.

As before, we notice that the ASC's of both μ and μ' may be expressed in terms of entropies on the coupled space $X \times \{0,1\}^{\mathbb{Z}}$ with the measure $\bar{\mu}$.

Let $\pi\colon X\times\{0,1\}^{\mathbb{Z}}$ be projection onto the first coordinate and we define $\bar{\mathcal{P}}_S=\pi^{-1}\mathcal{P}_S$. Similarly, define $\bar{\mathcal{Q}}_S=\Phi^{-1}\mathcal{P}_S$. We let $\mathcal{R}$ be the partition of $X\times\{0,1\}^{\mathbb{Z}}$ given by $\{A_0,\ldots,A_{n-1},(A_0\cup\ldots\cup A_{n-1})^c\}$, where $A_i=\bar{\sigma}^i\pi_2^{-1}[1]$. That is, $(x,y)\in A_i$ indicates that $\Phi(x,y)_0=W_i$.

We then have

$$\begin{aligned}\mathrm{ASC}(\mu)&=\int_{T_0}H_{\bar{\mu}}(\bar{\mathcal{P}}_0|\bar{P}_{S^-})\,d\mathbb{P}(S);\text{ and}\\ \mathrm{ASC}(\mu')&=\int_{T_0}H_{\bar{\mu}}(\bar{\mathcal{Q}}_0|\bar{\mathcal{Q}}_{S^-})\,d\mathbb{P}(S),\end{aligned}$$

thereby suggesting that the coupling will allow us to make pointwise comparisons.

For now, write (X_n) for the $\bar{\mathcal{P}}$ process; and (Z_n) for the $\bar{\mathcal{Q}}$ process. Fix an $S\in T_0$ and write $A=S\cap[-(n-1),-1]$, and $B=S\cap(-\infty,-n]$.

For a point $\omega\in\bar{X}$ such that $\sigma^l\omega\notin A_0\cup\cdots\cup A_{n-2}$ (that is such that x_{l-1} and x_l are not part of a common W replacement), we have

$$\begin{aligned}&\mathbb{P}(Z_0=i|Z_A=z_A;Z_B,X_B,R_B)\\ &=\frac{\mathbb{P}(Z_0=i,Z_A=z_A|Z_B,X_B,R_B)}{\mathbb{P}(Z_A=z_A|Z_B,X_B,R_B)}\\ &=\frac{\sum_{j,x_A}\mathbb{P}(Z_0=i,Z_A=z_A|X_0=j,X_A=x_A;X_B)\mathbb{P}(X_0=j,X_A=x_A|X_B)}{\sum_{x_A}\mathbb{P}(Z_A=z_A|X_A=x_A;R_B,X_B)\mathbb{P}(X_A=x_A|X_B)}.\end{aligned}$$

We bound this quantity from below. The numerator is bounded below by

$$\begin{aligned}&\mathbb{P}(Z_0=i,Z_A=z_A|X_0=i,X_A=z_A;X_B)\mathbb{P}(X_0=i,X_A=z_A|X_B)\\ &\geq(1-\ell\epsilon)\mathbb{P}(X_0=i,X_A=z_A|X_B)\\ &=(1-\ell\epsilon)\mathbb{P}(X_0=i|X_A=z_A;X_B)\mathbb{P}(X_A=z_A|X_B).\end{aligned}$$

The denominator is bounded above by $\mathbb{P}(X_A=z_A|X_B)+\ell\epsilon\sum_{x_A\neq z_A}\mathbb{P}(X_A=x_A|X_B)\leq\mathbb{P}(X_A=z_A|X_B)+\ell\epsilon$.

Combining the estimates, we see that for each i,

$$\begin{aligned}&\mathbb{P}(Z_0=i|Z_A=z_A;Z_B,X_B,R_B)\\ &\geq(1-\ell\epsilon)\mathbb{P}(X_0=i|X_A=z_A;X_B)\frac{\mathbb{P}(X_A=z_A|X_B)}{\mathbb{P}(X_A=z_A|X_B)+\ell\epsilon}.\end{aligned}$$

Define $\delta(z_A)$ by:

$$\delta(z_A) = \frac{\ell\epsilon(1 + \mathbb{P}(X_A = z_A|X_B))}{\mathbb{P}(X_A = z_A|X_B) + \ell\epsilon}$$

and let $\mathcal{B}_B$ be the σ-algebra generated by Z_B, X_B and R_B.

Then the above can be rewritten

$$\mathbb{P}(Z_0 = i|Z_A = z_A; \mathcal{B}_B) \geq (1 - \delta(z_A))\mathbb{P}(X_0 = i|X_A = z_A; X_B).$$

Notice that

$$\delta(z_A) < \frac{2\ell\epsilon}{\mathbb{P}(X_A = z_A|X_B)}. \tag{4}$$

Now by Lemma 11,

$$\begin{aligned}
&H(Z_0|Z_A, Z_B) \geq H(Z_0|Z_A; \mathcal{B}_B) \\
=&\sum_{z_A} H(Z_0|Z_A = z_A; \mathcal{B}_B)\mathbb{P}(Z_A = z_A|\mathcal{B}_B) \\
\geq&\sum_{z_A} \Big(H(X_0|X_A = z_A; X_B) - \delta(z_A)\big(1 + H(X_0|X_A = z_A; x_B)\big)\Big) \\
&\quad \mathbb{P}(Z_A = z_A|\mathcal{B}_B) + E \\
\geq&(1 - \ell\epsilon)\sum_{z_A} \Big(H(X_0|X_A = z_A; X_B) - \delta(z_A)\big(1 + H(X_0|X_A = z_A; x_B)\big)\Big) \\
&\quad \mathbb{P}(X_A = z_A|\mathcal{B}_B) + E \\
=&(1 - \ell\epsilon) \int \Big(H(X_0|X_A, X_B) - \delta(X_A)\big(1 + H(X_0|X_A, X_B)\big)\Big)\, d\mathbb{P} + E \\
\geq&(1 - \ell\epsilon) \int_{\{\mathbb{P}(X_A|X_B) > 2\ell\epsilon\}} \Big(H(X_0|X_A, X_B) - \delta(X_A)\big(1 + H(X_0|X_A, X_B)\big)\Big)\, d\mathbb{P} + E \\
\geq&(1 - \ell\epsilon) \int_{\{\mathbb{P}(X_A|X_B) > 2\ell\epsilon\}} \Big(H(X_0|X_A, X_B) - \delta(X_A)\big(1 + \log N\big)\Big)\, d\mathbb{P} + E,
\end{aligned}$$

where the E that appears in the second inequality, and is carried throughout the proof is the contribution to the entropy of Z occurring from configurations that have 0 probability in the X sequence.

In the above calculation, the estimate in the third line relies on the assumption that $\sigma^l\omega \notin A_0 \cup \cdots \cup A_{n-2}$, as otherwise the hypotheses of Lemma 11 are not satisfied. A correction needs to be made for these σ's. However, this set is $\mathcal{B}_B$-measurable, so the entropy contribution from these σ's is of $O(\epsilon)$.

Using (4), we see

$$\int_{\{\mathbb{P}(X_A|X_B)>2\ell\epsilon\}} \delta(X_A)\,d\mathbb{P} < 2\ell\epsilon N^n$$

Let $S = \{\omega\colon \mathbb{P}(X_A|X_B) \le 2\ell\epsilon\}$. We have

$$\mathbf{1}_S = \sum_{U\in\mathcal{A}^A} \mathbf{1}_{[U]_A}(\omega)\mathbf{1}_{\{\mathbb{P}([U]_A|X_B)\le 2\ell\epsilon\}}.$$

Notice that the second term is X_B-measurable. Using properties of conditional expectations, we see that

$$\mathbb{P}(S|X_B) = \sum_{U\in\mathcal{A}^A} \mathbb{P}([U]_A|X_B)\mathbf{1}_{\{\mathbb{P}([U]_A|X_B)\le 2\ell\epsilon\}},$$

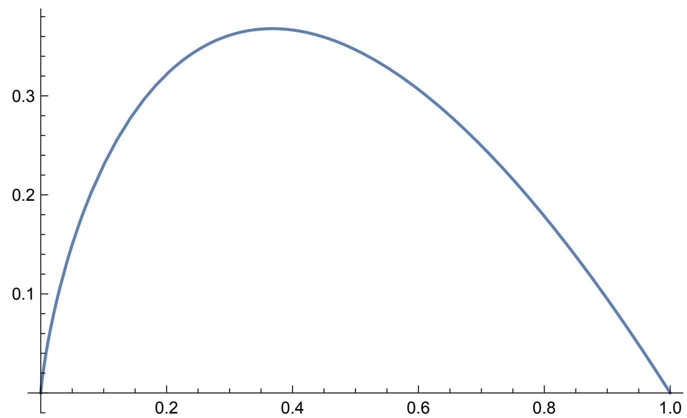
which is bounded above pointwise by $2\ell\epsilon N^n$. Now taking an expectation and using the tower law, we see that

$$\mathbb{P}\Big(\{\mathbb{P}(X_A|X_B) \le 2\ell\epsilon\}\Big) \le 2\ell\epsilon N^n.$$

Therefore

$$\begin{aligned}
\mathrm{ASC}_{A\cup B}(Z) \ge (1-\ell\epsilon)&\int H(X_0|X_A,X_B)\,d\mathbb{P} + E \\
&- \int_{\{\mathbb{P}(X_A|X_B)\le 2\ell\epsilon\}} H(X_0|X_A,X_B)\,d\mathbb{P} - 2\ell\epsilon N^n(1+\log N) + E \\
&\ge (1-\ell\epsilon)\,\mathrm{ASC}_{A\cup B}(X) - \int_{\{\mathbb{P}(X_A|X_B)\le 2\ell\epsilon\}} \log N\,d\mathbb{P} + E - O(\epsilon) \\
&= \mathrm{ASC}_{A\cup B}(X) + E - O(\epsilon).
\end{aligned}$$

In particular, by integrating, we see that $\mathrm{ASC}(Z) \ge \mathrm{ASC}(X) + E - O(\epsilon)$. However, the quantity E is of order $-\epsilon\log\epsilon$. Including those terms, we see that for sufficiently small ϵ, $\mathrm{ASC}(Z) > \mathrm{ASC}(X)$. Hence we see that a measure of maximum average symbolic complexity is fully supported, as claimed.


0.3
0.2
0.1
0.2
0.4
0.6
0.8
1.0

□