§ 8. Compressing stationary ergodic sources

We have examined the compression of i.i.d. sequence $\{S_i\}$, for which

$$\frac{1}{n} l(f^*(S^n)) \to H(S) \quad \text{in prob.} \tag{8.1}$$

$$\lim_{n\to\infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases} \tag{8.2}$$

In this lecture, we shall examine similar results for ergodic processes and we first state the main theory as follows:

Theorem 8.1 (Shannon-McMillan)**.** *Let $\{S_1, S_2, \dots\}$ be a stationary and ergodic discrete process, then*

$$\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} \mathcal{H}, \quad \textit{also a.s. and in } L_1 \tag{8.3}$$

where $\mathcal{H} = \lim_{n\to\infty} \frac{1}{n} H(S^n)$ is the entropy rate.

Corollary 8.1. *For any stationary and ergodic discrete process $\{S_1, S_2, \dots\}$, (8.1) – (8.2) hold with $H(S)$ replaced by $\mathcal{H}$.*

Proof. Shannon-McMillan (we only need convergence in probability) + Theorem 6.4 + Theorem 7.1 which tie together the respective CDF of the random variable $l(f^*(S^n))$ and $\log \frac{1}{P_{S^n}(s^n)}$. □

In Lecture 7 we learned the asymptotic equipartition property (AEP) for iid sources. Here we generalize it to stationary ergodic sources thanks to Shannon-McMillan.

Corollary 8.2 (AEP for stationary ergodic sources)**.** *Let $\{S_1, S_2, \dots\}$ be a stationary and ergodic discrete process. For any $\delta > 0$, define the set*

$$T_n^\delta = \left\{ s^n : \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| \le \delta \right\}.$$

Then

1. $\mathbb{P}\left[S^n \in T_n^\delta\right] \to 1$ *as* $n \to \infty$.
2. $2^{n(\mathcal{H}-\delta)}(1 + o(1)) \le |T_n^\delta| \le 2^{(\mathcal{H}+\delta)n}(1 + o(1))$.

Note:

- Convergence in probability for stationary ergodic Markov chains [Shannon 1948]
- Convergence in L_1 for stationary ergodic processes [McMillan 1953]

- Convergence almost surely for stationary ergodic processes [Breiman 1956] (Either of the last two results implies the convergence Theorem 8.1 in probability.)
- For a Markov chain, existence of typical sequences can be understood by thinking of Markov process as sequence of independent decisions regarding which transitions to take. It is then clear that Markov process's trajectory is simply a transformation of trajectories of an i.i.d. process, hence must similarly concentrate similarly on some typical set.

8.1 Bits of ergodic theory

Let's start with a dynamic system view and introduce a few definitions:

Definition 8.1 (Measure preserving transformation)**.** $\tau : \Omega \to \Omega$ is measure preserving (more precisely, probability preserving) if

$$\forall E \in \mathcal{F}, P(E) = P(\tau^{-1}E).$$

The set E is called τ-invariant if $E = \tau^{-1}E$. The set of all τ-invariant sets forms a σ-algrebra (check!) denoted $\mathcal{F}_{inv}$.

Definition 8.2 (stationary process)**.** A process $\{S_n, n = 0, \dots\}$ is stationary if there exists a measure preserving transformation $\tau : \Omega \to \Omega$ such that:

$$S_j = S_{j-1} \circ \tau = S_0 \circ \tau^j$$

Therefore a stationary process can be described by the tuple $(\Omega, \mathcal{F}, \mathbb{P}, \tau, S_0)$ and $S_k = S_0 \circ \tau^k$.

Notes:

1. Alternatively, a random process $(S_0, S_1, S_2, \dots)$ is stationary if its joint distribution is invariant with respect to shifts in time, i.e., $P_{S_n^m} = P_{S_{n+t}^{m+t}}$, $\forall n, m, t$. Indeed, given such a process we can define a m.p.t. as follows:
$$(s_0, s_1, \dots) \xrightarrow{\tau} (s_1, s_2, \dots) \tag{8.4}$$
So τ is a shift to the right.

2. An event $E \in \mathcal{F}$ is shift-invariant if
$$(s_1, s_2, \dots) \in E \Rightarrow \forall s_0 (s_0, s_1, s_2, \dots) \in E$$
or equivalently $E = \tau^{-1}E$ (check!). Thus τ-invariant events are also called shift-invariant, when τ is interpreted as (8.4).

3. Some examples of shift-invariant events are $\{\exists n : x_i = 0 \forall i \ge n\}$, $\{\limsup x_i < 1\}$ etc. A non shift-invariant event is $A = \{x_0 = x_1 = \cdots = 0\}$, since $\tau(1, 0, 0, \dots) \in A$ but $(1, 0, \dots) \notin A$.

4. Also recall that the tail σ-algebra is defined as
$$\mathcal{F}_{tail} \triangleq \bigcap_{n \ge 1} \sigma\{S_n, S_{n+1}, \dots\}.$$
It is easy to check that all shift-invariant events belong to $\mathcal{F}_{tail}$. The inclusion is strict, as for example the event
$$\{\exists n : x_i = 0, \forall \ \underline{\text{odd}} \ i \ge n\}$$
is in $\mathcal{F}_{tail}$ but not shift-invariant.

Proposition 8.1 (Poincare recurrence)**.** *Let τ be measure-preserving for $(\Omega, \mathcal{F}, \mathbb{P})$. Then for any measurable A with $\mathbb{P}[A] > 0$ we have*

$$\mathbb{P}[\bigcup_{k\geq 1} \tau^{-k} A | A] = \mathbb{P}[\tau^k(\omega) \in A - -\textit{infinitely often} | A] = 1\,.$$

Proof. Let $B = \bigcup_{k\geq 1} \tau^{-k} A$. It is sufficient to show that $\mathbb{P}[A \cap B] = \mathbb{P}[A]$ or equivalently

$$\mathbb{P}[A \cup B] = \mathbb{P}[B]\,. \tag{8.5}$$

To that end notice that $\tau^{-1} A \cup \tau^{-1} B = B$ and thus

$$\mathbb{P}[\tau^{-1}(A \cup B)] = \mathbb{P}[B]\,,$$

but the left-hand side equals $\mathbb{P}[A \cup B]$ by the measure-preservation of τ, proving (8.5). □

Note: Consider τ mapping initial state of the conservative (Hamiltonian) mechanical system to its state after passage of a given unit of time. It is known that τ preserves Lebesgue measure in phase space (Liouville's theorem). Thus Poincare recurrence leads to rather counter-intuitive conclusions. For example, opening the barrier separating two gases in a cylinder allows them to mix. Poincare recurrence says that eventually they will return back to the original separated state (with each gas occupying roughly its half of the cylinder).

Definition 8.3 (Ergodicity)**.** A transformation τ is ergodic if $\forall E \in \mathcal{F}_{inv}$ we have $\mathbb{P}[E] = 0$ or 1. A process $\{S_i\}$ is ergodic if all shift invariant events are deterministic, i.e., for any shift invariant event E, $\mathbb{P}[S_1^\infty \in E] = 0$ or 1.

Example:

- $\{S_k = k^2\}$: ergodic but not stationary
- $\{S_k = S_0\}$: stationary but not ergodic (unless S_0 is a constant). Note that the singleton set $E = \{(s, s, \ldots)\}$ is shift invariant and $\mathbb{P}[S_1^\infty \in E] = \mathbb{P}[S_0 = s] \in (0,1)$ – not deterministic.
- $\{S_k\}$ i.i.d. is stationary and ergodic (by Kolmogorov's 0-1 law, tail events have no randomness)
- (Sliding-window construction of ergodic processes)
  If $\{S_i\}$ is ergodic, then $\{X_i = f(S_i, S_{i+1}, \ldots)\}$ is also ergodic. It is called a **B-process** if S_i is i.i.d.
  Example, $S_i \sim \text{Bern}(\frac{1}{2})$ i.i.d., $X_k = \sum_{n=0}^\infty 2^{-n-1} S_{k+n} = 2X_{k-1} \mod 1$. The marginal distribution of X_i is uniform on $[0,1]$. *Note that X_k's behavior is completely deterministic:* given X_0, all the future X_k's are determined exactly. This example shows that certain deterministic maps exhibit ergodic/chaotic behavior under iterative application: although the trajectory is completely deterministic, its time-averages converge to expectations and in general "look random".
- There are also stronger conditions than ergodicity. Namely, we say that τ is mixing (or strong mixing) if

  $$\mathbb{P}[A \cap \tau^{-n} B] \to \mathbb{P}[A]\mathbb{P}[B]\,.$$

  We say that τ is weakly mixing if

  $$\sum_{k=1}^{n} \frac{1}{n} |\mathbb{P}[A \cap \tau^{-n} B] - \mathbb{P}[A]\mathbb{P}[B]| \to 0\,.$$

  Strong mixing implies weak mixing, which implies ergodicity (check!).

- $\{S_i\}$: finite irreducible Markov chain with recurrent states is ergodic (in fact strong mixing), regardless of initial distribution.
  Toy example: kernel $P(0|1) = P(1|0) = 1$ with initial dist. $P(S_0 = 0) = 0.5$. This process only has two sample paths: $\mathbb{P}[S_1^\infty = (010101\ldots)] = \mathbb{P}[S_1^\infty = (101010\ldots)] = \frac{1}{2}$. It is easy to verify this process is ergodic (in the sense defined above!). Note however, that in Markov-chain literature a chain is called ergodic if it is irreducible, aperiodic and recurrent. This example does not satisfy this definition (this clash of terminology is a frequent source of confusion).

- (optional) $\{S_i\}$: stationary zero-mean Gaussian process with autocovariance function $R(n) = \mathbb{E}[S_0 S_n^*]$.

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{t=0}^{n} R[t] = 0 \Leftrightarrow \{S_i\} \text{ ergodic} \Leftrightarrow \{S_i\} \text{ weakly mixing}$$
$$\lim_{n\to\infty} R[n] = 0 \Leftrightarrow \{S_i\} \text{ mixing}$$

Intuitively speaking, an ergodic process can have infinite memory in general, but the memory is weak. Indeed, we see that for a stationary Gaussian process ergodicity means the correlation dies (in the Cesaro-mean sense).

The *spectral measure* is defined as the (discrete time) Fourier transform of the autocovariance sequence $\{R(n)\}$, in the sense that there exists a unique probability measure μ on $[-\frac{1}{2}, \frac{1}{2}]$ such that $R(n) = \mathbb{E}\exp(i2n\pi X)$ where $X \sim \mu$. The spectral criteria can be formulated as follows:

$$\{S_i\} \text{ ergodic} \Leftrightarrow \text{spectral measure has no atoms (CDF is continuous)}$$
$$\{S_i\} \text{ B-process} \Leftrightarrow \text{spectral measure has density}$$

Detailed exposition on stationary Gaussian processes can be found in [Doo53, Theorem 9.3.2, pp. 474, Theorem 9.7.1, pp. 494–494].[1]

8.2 Proof of Shannon-McMillan

We shall show the convergence in L_1, which implies convergence in probability automatically. In order to prove Shannon-McMillan, let's first introduce the Birkhoff-Khintchine's convergence theorem for ergodic processes, the proof of which is presented in the next subsection.

Theorem 8.2 (Birkhoff-Khintchine's Ergodic Theorem). *If $\{S_i\}$ stationary and ergodic, $\forall$ function $f \in L_1$, i.e., $\mathbb{E}|f(S_1,\ldots)| < \infty$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(S_k,\ldots) = \mathbb{E} f(S_1,\ldots). \quad \textit{a.s. and in } L_1$$

In the special case where f depends on finitely many coordinates, say, $f = f(S_1,\ldots,S_m)$, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(S_k,\ldots,S_{k+m-1}) = \mathbb{E} f(S_1,\ldots,S_m). \quad \textit{a.s. and in } L_1$$

Interpretation: time average converges to ensemble average.
Example: Consider $f = f(S_1)$

[1] Thanks Prof. Bruce Hajek for the pointer.

- $\{S_i\}$ is iid. Then Theorem 8.2 is SLLN (strong LLN).
- $\{S_i\}$ is such that $S_i = S_1$ for all i – non-ergodic. Then Theorem 8.2 fails unless S_1 is a constant.

Definition 8.4. $\{S_i : i \in \mathbb{N}\}$ is an m^{th} order Markov chain if $P_{S_{t+1}|S_1^t} = P_{S_{t+1}|S_{t-m+1}^t}$ for all $t \geq m$. It is called time homogeneous if $P_{S_{t+1}|S_{t-m+1}^t} = P_{S_{m+1}|S_1^m}$.

Remark 8.1. Showing (8.3) for an m^{th} order time homogeneous Markov chain $\{S_i\}$ is a direct application of Birkhoff-Khintchine.

$$\begin{aligned}\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} &= \frac{1}{n}\sum_{t=1}^{n}\log\frac{1}{P_{S_t|S^{t-1}}(S_t|S^{t-1})}\\ &= \frac{1}{n}\log\frac{1}{P_{S^m}(S^m)} + \frac{1}{n}\sum_{t=m+1}^{n}\log\frac{1}{P_{S_t|S_{t-m}^{t-1}}(S_l|S_{l-m}^{l-1})}\\ &= \underbrace{\frac{1}{n}\log\frac{1}{P_{S_1}(S_1^m)}}_{\to 0} + \underbrace{\frac{1}{n}\sum_{t=m+1}^{n}\log\frac{1}{P_{S_{m+1}|S_1^m}(S_t|S_{t-m}^{t-1})}}_{\to H(S_{m+1}|S_1^m)\text{ by Birkhoff-Khintchine}},\end{aligned} \tag{8.6}$$

where we applied Theorem 8.2 with $f(s_1, s_2, \ldots) = \log\frac{1}{P_{S_{m+1}|S_1^m}(s_{m+1}|s_1^m)}$.

Now let's prove (8.3) for a general stationary ergodic process $\{S_i\}$ which might have infinite memory. The idea is to approximate the distribution of that ergodic process by an m-th order MC (finite memory) and make use of (8.6); then let $m \to \infty$ to make the the approximation accurate (*Markov approximation*).

Proof of Theorem 8.1 in L_1. To show that (8.3) converges in L_1, we want to show that

$$\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\right| \to 0, \quad n \to \infty.$$

To this end, fix an $m \in \mathbb{N}$. Define the following auxiliary distribution for the process:

$$\begin{aligned}Q^{(m)}(S_1^\infty) &= P_{S_1^m}(S_1^m)\prod_{t=m+1}^{\infty}P_{S_t|S_{t-m}^{t-1}}(S_t|S_{t-m}^{t-1})\\ &\overset{\text{stat.}}{=} P_{S_1^m}(S_1^m)\prod_{t=m+1}^{\infty}P_{S_{m+1}|S_1^m}(S_t|S_{t-m}^{t-1})\end{aligned}$$

Note that under $Q^{(m)}$, $\{S_i\}$ is an m^{th}-order time-homogeneous Markov chain.

By triangle inequality,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\right| \leq & \underbrace{\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \frac{1}{n}\log\frac{1}{Q_{S^n}^{(m)}(S^n)}\right|}_{\triangleq A}\\ &+ \underbrace{\mathbb{E}\left|\frac{1}{n}\log\frac{1}{Q_{S^n}^{(m)}(S^n)} - H_m\right|}_{\triangleq B} + \underbrace{|H_m - \mathcal{H}|}_{\triangleq C}\end{aligned}$$

where $H_m \triangleq H(S_{m+1}|S_1^m)$.

Now

- $C = |H_m - \mathcal{H}| \to 0$ as $m \to \infty$ by Theorem 5.4 (Recall that for stationary processes: $H(S_{m+1}|S_1^m) \to H$ from above).
- As shown in Remark 8.1, for any fixed m, $B \to 0$ in L_1 as $n \to \infty$, as a consequence of Birkhoff-Khintchine. Hence for any fixed m, $\mathbb{E}B \to 0$ as $n \to \infty$.
- For term A,

  $$\mathbb{E}[A] = \frac{1}{n}\mathbb{E}_P\left|\log \frac{dP_{S^n}}{dQ_{S^n}^{(m)}}\right| \le \frac{1}{n}D(P_{S^n}\|Q_{S^n}^{(m)}) + \frac{2\log e}{en}$$

  where

  $$\begin{aligned}\frac{1}{n}D(P_{S^n}\|Q_{S^n}^{(m)}) &= \frac{1}{n}\mathbb{E}\left[\log \frac{P_{S^n}(S^n)}{P_{S^m}(S^m)\prod_{t=m+1}^n P_{S_{m+1}|S_m^1}(S_t|S_{t-m}^{t-1})}\right] \\ &\overset{\text{stat.}}{=} \frac{1}{n}(-H(S^n) + H(S^m) + (n-m)H_m) \\ &\to H_m - \mathcal{H} \text{ as } n \to \infty\end{aligned}$$

  and the next Lemma 8.1.

Combining all three terms and sending $n \to \infty$, we obtain for any m,

$$\limsup_{n\to\infty} \mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\right| \le 2(H_m - \mathcal{H}).$$

Sending $m \to \infty$ completes the proof of L_1-convergence. □

Lemma 8.1.

$$\mathbb{E}_P\left[\left|\log\frac{dP}{dQ}\right|\right] \le D(P\|Q) + \frac{2\log e}{e}.$$

Proof. $|x\log x| - x\log x \le \frac{2\log e}{e}, \forall x > 0$, since LHS is zero if $x \ge 1$, and otherwise upper bounded by $2\sup_{0\le x\le 1} x\log\frac{1}{x} = \frac{2\log e}{e}$. □

8.3* Proof of Birkhoff-Khintchine

Proof of Theorem 8.2. $\forall$ function $\tilde{f} \in L_1$, $\forall\epsilon$, there exists a decomposition $\tilde{f} = f + h$ such that f is bounded, and $h \in L_1$, $\|h\|_1 \le \epsilon$.
Let us first focus on the bounded function f. Note that in the bounded domain $\mathcal{L}_1 \subset \mathcal{L}_2$, thus $f \in \mathcal{L}_2$. Furthermore, $\mathcal{L}_2$ is a Hilbert space with inner product $(f,g) = \mathbb{E}[f(S_1^\infty)\overline{g(S_1^\infty)}]$.

For the measure preserving transformation τ that generates the stationary process $\{S_i\}$, define the operator $T(f) = f\circ\tau$. Since τ is measure preserving, we know that $\|Tf\|_2^2 = \|f\|_2^2$, thus T is a unitary and bounded operator.

Define the operator

$$A_n(f) = \frac{1}{n}\sum_{k=1}^n f\circ\tau^k$$

Intuitively:

$$A_n = \frac{1}{n}\sum_{k=1}^n T^k = \frac{1}{n}(I - T^n)(I-T)^{-1}$$

Then, if $f \perp \ker(I-T)$ we should have $A_n f \to 0$, since only components in the kernel can blow up. This intuition is formalized in the proof below.

Let's further decompose f into two parts $f = f_1 + f_2$, where $f_1 \in \ker(I-T)$ and $f_2 \in \ker(I-T)^\perp$. Observations:

- if $g \in \ker(I-T)$, g must be a constant function. This is due to the ergodicity. Consider indicator function $\mathbf{1}_A$, if $\mathbf{1}_A = \mathbf{1}_A \circ \tau = \mathbf{1}_{\tau^{-1}A}$, then $\mathbb{P}[A] = 0$ or 1. For a general case, suppose $g = Tg$ and g is not constant, then at least some set $\{g \in (a,b)\}$ will be shift-invariant and have non-trivial measure, violating ergodicity.
- $\ker(I-T) = \ker(I-T^*)$. This is due to the fact that T is unitary:

  $$g = Tg \Rightarrow \|g\|^2 = (Tg,g) = (g,T^*g) \Rightarrow (T^*g,g) = \|g\|\|T^*g\| \Rightarrow T^*g = g$$

  where in the last step we used the fact that Cauchy-Schwarz $(f,g) \le \|f\| \cdot \|g\|$ only holds with equality for $g = cf$ for some constant c.
- $\ker(I-T)^\perp = \ker(I-T^*)^\perp = [\mathrm{Im}(I-T)]$, where $[\mathrm{Im}(I-T)]$ is an $\mathcal{L}_2$ closure.
- $g \in \ker(I-T)^\perp \iff \mathbb{E}[g] = 0$. Indeed, only zero-mean functions are orthogonal to constants.

With these observations, we know that $f_1 = m$ is a const. Also, $f_2 \in [\mathrm{Im}(I-T)]$ so we further approximate it by $f_2 = f_0 + h_1$, where $f_0 \in \mathrm{Im}(I-T)$, namely $f_0 = g - g\circ\tau$ for some function $g \in \mathcal{L}_2$, and $\|h_1\|_1 \le \|h_1\|_2 < \epsilon$. Therefore we have

$$\begin{aligned} A_n f_1 &= f_1 = \mathbb{E}[f] \\ A_n f_0 &= \frac{1}{n}(g - g\circ\tau^n) \to 0 \text{ a.s. and } L_1 \\ (\text{since } \mathbb{E}[\sum_{n\ge1}(\frac{g\circ\tau^n}{n})^2] &= \mathbb{E}[g^2]\sum\frac{1}{n^2} < \infty \quad \Longrightarrow \quad \frac{1}{n}g\circ\tau^n \to 0 \text{ a.s.}) \end{aligned}$$

The proof completes by showing

$$\mathbb{P}[\limsup_n A_n(h+h_1) \ge \delta] \le \frac{2\epsilon}{\delta}. \tag{8.7}$$

Indeed, then by taking $\epsilon \to 0$ we will have shown

$$\mathbb{P}[\limsup_n A_n(f) \ge \mathbb{E}[f] + \delta] = 0$$

as required. □

Proof of (8.7) makes use of the Maximal Ergodic Lemma stated as follows:

Theorem 8.3 (Maximal Ergodic Lemma). *Let $(\mathbb{P},\tau)$ be a probability measure and a measure-preserving transformation. Then for any $f \in L_1(\mathbb{P})$ we have*

$$\mathbb{P}\left[\sup_{n\ge1} A_n f > a\right] \le \frac{\mathbb{E}[f\mathbf{1}_{\sup_{n\ge1} A_n f > a}]}{a} \le \frac{\|f\|_1}{a}$$

where $A_n f = \frac{1}{n}\sum_{k=0}^{n-1} f\circ\tau^k$.

Note: This is a so-called "weak L_1" estimate for a sublinear operator $\sup_n A_n(\cdot)$. In fact, this theorem is exactly equivalent to the following result:

Lemma 8.2 (Estimate for the maximum of averages)**.** *Let $\{Z_n, n = 1, \dots\}$ be a stationary process with $\mathbb{E}[|Z|] < \infty$ then*

$$\mathbb{P}\left[\sup_{n\geq 1} \frac{|Z_1 + \ldots + Z_n|}{n} > a\right] \leq \frac{\mathbb{E}[|Z|]}{a} \qquad \forall a > 0$$

Proof. The argument for this Lemma has originally been quite involved, until a dramatically simple proof (below) was found by A. Garcia.

Define

$$S_n = \sum_{k=1}^{n} Z_k \tag{8.8}$$

$$L_n = \max\{0, Z_1, \dots, Z_1 + \dots + Z_n\} \tag{8.9}$$

$$M_n = \max\{0, Z_2, Z_2 + Z_3, \dots, Z_2 + \dots + Z_n\} \tag{8.10}$$

$$Z^* = \sup_{n\geq 1} \frac{S_n}{n} \tag{8.11}$$

It is sufficient to show that

$$\mathbb{E}[Z_1 1_{\{Z^*>0\}}] \geq 0\,. \tag{8.12}$$

Indeed, applying (8.12) to $\tilde{Z}_1 = Z_1 - a$ and noticing that $\tilde{Z}^* = Z^* - a$ we obtain

$$\mathbb{E}[Z_1 1_{\{Z^*>a\}}] \geq a\mathbb{P}[Z^* > a]\,,$$

from which Lemma follows by upper-bounding the left-hand side with $\mathbb{E}[|Z_1|]$.

In order to show (8.12) we first notice that $\{L_n > 0\} \nearrow \{Z^* > 0\}$. Next we notice that

$$Z_1 + M_n = \max\{S_1, \dots, S_n\}$$

and furthermore

$$Z_1 + M_n = L_n \qquad \text{on } \{L_n > 0\}$$

Thus, we have

$$Z_1 1_{\{L_n>0\}} = L_n - M_n 1_{\{L_n>0\}}$$

where we do not need indicator in the first term since $L_n = 0$ on $\{L_n > 0\}^c$. Taking expectation we get

$$\mathbb{E}[Z_1 1_{\{L_n>0\}}] = \mathbb{E}[L_n] - \mathbb{E}[M_n 1_{\{L_n>0\}}] \tag{8.13}$$

$$\geq \mathbb{E}[L_n] - \mathbb{E}[M_n] \tag{8.14}$$

$$= \mathbb{E}[L_n] - \mathbb{E}[L_{n-1}] = \mathbb{E}[L_n - L_{n-1}] \geq 0\,, \tag{8.15}$$

where we used $M_n \geq 0$, the fact that M_n has the same distribution as L_{n-1}, and $L_n \geq L_{n-1}$, respectively. Taking limit as $n \to \infty$ in (8.15) we obtain (8.12). □

8.4* Sinai's generator theorem

It turns out there is a way to associate to every probability-preserving transformation τ a number, called Kolmogorov-Sinai entropy. This number is invariant to isomorphisms of p.p.t.'s (appropriately defined).

Definition 8.5. Fix a probability-preserving transformation τ acting on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Kolmogorov-Sinai entropy of τ is defined as

$$\mathcal{H}(\tau) \triangleq \sup_{X_0} \lim_{n\to\infty} \frac{1}{n} H(X_0, X_0 \circ \tau, \dots, X_0 \circ \tau^{n-1}),$$

where supremum is taken over all random variables $X_0 : \Omega \to \mathcal{X}$ with finite range $\mathcal{X}$ and measurable with respect to $\mathcal{F}$.

Note that every random variable X_0 generates a stationary process adapted to τ, that is

$$X_k \triangleq X_0 \circ \tau^k .$$

In this way, Kolmogorov-Sinai entropy of τ equals the maximal entropy rate among all stationary processes adapted to τ. This quantity may be extremely hard to evaluate, however. One help comes in the form of the famous criterion of Y. Sinai. We need to elaborate on some more concepts before:

- σ-algebra $\mathcal{G} \subset \mathcal{F}$ is $\mathbb{P}$-dense in $\mathcal{F}$, or sometimes we also say $\mathcal{G} = \mathcal{F} \mod \mathbb{P}$ or even $\mathcal{G} = \mathcal{F} \mod 0$, if for every $E \in \mathcal{F}$ there exists $E' \in \mathcal{G}$ s.t.

$$\mathbb{P}[E \Delta E'] = 0 .$$

- Partition $\mathcal{A} = \{A_i, i = 1, 2, \dots\}$ measurable with respect to $\mathcal{F}$ is called generating if

$$\bigvee_{n=0}^{\infty} \sigma\{\tau^{-n}\mathcal{A}\} = \mathcal{F} \mod \mathbb{P} .$$

- Random variable $Y : \Omega \to \mathcal{Y}$ with a *countable* alphabet $\mathcal{Y}$ is called a generator of $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$ if

$$\sigma\{Y, Y \circ \tau, \dots, Y \circ \tau^n, \dots\} = \mathcal{F} \mod \mathbb{P}$$

Theorem 8.4 (Sinai's generator theorem)**.** *Let Y be the generator of a p.p.t. $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$. Let $H(\mathbb{Y})$ be the entropy rate of the process $\mathbb{Y} = \{Y_k = Y \circ \tau^k, k = 0, \dots\}$. If $H(\mathbb{Y})$ is finite, then $\mathcal{H}(\tau) = H(\mathbb{Y})$.*

Proof. Notice that since $H(\mathbb{Y})$ is finite, we must have $H(Y_0^n) < \infty$ and thus $H(Y) < \infty$. First, we argue that $\mathcal{H}(\tau) \ge H(\mathbb{Y})$. If Y has finite alphabet, then it is simply from the definition. Otherwise let Y be $\mathbb{Z}_+$-valued. Define a truncated version $\tilde{Y}_m = \min(Y, m)$, then since $\tilde{Y}_m \to Y$ as $m \to \infty$ we have from lower semicontinuity of mutual information, cf. (3.9), that

$$\lim_{m\to\infty} I(Y; \tilde{Y}_m) \ge H(Y) ,$$

and consequently for arbitrarily small ϵ and sufficiently large m

$$H(Y|\tilde{Y}) \le \epsilon ,$$

Then, consider the chain

$$\begin{aligned}H(Y_0^n) &= H(\tilde{Y}_0^n, Y_0^n) = H(\tilde{Y}_0^n) + H(Y_0^n|\tilde{Y}_0^n)\\ &= H(\tilde{Y}_0^n) + \sum_{i=0}^n H(Y_i|\tilde{Y}_0^n, Y_0^{i-1})\\ &\le H(\tilde{Y}_0^n) + \sum_{i=0}^n H(Y_i|\tilde{Y}_i)\\ &= H(\tilde{Y}_0^n) + nH(Y|\tilde{Y}) \le H(\tilde{Y}_0^n) + n\epsilon\end{aligned}$$

Thus, entropy rate of $\tilde{\mathbb{Y}}$ (which has finite-alphabet) can be made arbitrarily close to the entropy rate of $\mathbb{Y}$, concluding that $\mathcal{H}(\tau) \ge \mathcal{H}(\mathbb{Y})$.

The main part is showing that for any stationary process $\mathbb{X}$ adapted to τ the entropy rate is upper bounded by $H(\mathbb{Y})$. To that end, consider $X : \Omega \to \mathcal{X}$ with finite $\mathcal{X}$ and define as usual the process $\mathbb{X} = \{X \circ \tau^k, k = 0, 1, \ldots\}$. By generating property of $\mathbb{Y}$ we have that X (perhaps after modification on a set of measure zero) is a function of Y_0^∞. So are all X_k. Thus

$$H(X_0) = I(X_0; Y_0^\infty) = \lim_{n\to\infty} I(X_0; Y_0^n)\,,$$

where we used the continuity-in-σ-algebra property of mutual information, cf. (3.10). Rewriting the latter limit differently, we have

$$\lim_{n\to\infty} H(X_0|Y_0^n) = 0\,.$$

Fix $\epsilon > 0$ and choose m so that $H(X_0|Y_0^m) \le \epsilon$. Then consider the following chain:

$$\begin{aligned}H(X_0^n) &\le H(X_0^n, Y_0^n) = H(Y_0^n) + H(X_0^n|Y_0^n)\\ &\le H(Y_0^n) + \sum_{i=0}^n H(X_i|Y_i^n)\\ &= H(Y_0^n) + \sum_{i=0}^n H(X_0|Y_0^{n-i})\\ &\le H(Y_0^n) + m\log|\mathcal{X}| + (n-m)\epsilon\,,\end{aligned}$$

where we used stationarity of (X_k, Y_k) and the fact that $H(X_0|Y_0^{n-i}) < \epsilon$ for $i \le n-m$. After dividing by n and passing to the limit our argument implies

$$H(\mathbb{X}) \le H(\mathbb{Y}) + \epsilon\,.$$

Taking here $\epsilon \to 0$ completes the proof.

Alternative proof: Suppose X_0 is taking values on a finite alphabet $\mathcal{X}$ and $X_0 = f(Y_0^\infty)$. Then (this is a measure-theoretic fact) for every $\epsilon > 0$ there exists $m = m(\epsilon)$ and a function $f_\epsilon : \mathcal{Y}^{m+1} \to \mathcal{X}$ s.t.

$$\mathbb{P}[f(Y_0^\infty) \ne f_\epsilon(Y_0^m)] \le \epsilon\,.$$

(This is just another way to say that $\bigcup_n \sigma\{Y_0^n\}$ is $\mathbb{P}$-dense in $\sigma(Y_0^\infty)$.) Define a stationary process $\tilde{\mathbb{X}}$ as

$$\tilde{X}_j \triangleq f_\epsilon(Y_j^{m+j})\,.$$

Notice that since $\tilde{X}_0^n$ is a function of Y_0^{n+m} we have

$$H(\tilde{X}_0^n) \le H(Y_0^{n+m})\,.$$

Dividing by m and passing to the limit we obtain that for entropy rates

$$H(\tilde{\mathbb{X}}) \le H(\mathbb{Y}) .$$

Finally, to relate $\tilde{\mathbb{X}}$ to $\mathbb{X}$ notice that by construction

$$\mathbb{P}[\tilde{X}_j \ne X_j] \le \epsilon .$$

Since both processes take values on a fixed finite alphabet, from Corollary 5.2 we infer that

$$|H(\mathbb{X}) - H(\tilde{\mathbb{X}})| \le \epsilon \log |\mathcal{X}| + h(\epsilon) .$$

Altogether, we have shown that

$$H(\mathbb{X}) \le H(\mathbb{Y}) + \epsilon \log |\mathcal{X}| + h(\epsilon) .$$

Taking $\epsilon \to 0$ we conclude the proof. □

Examples:

- Let $\Omega = [0,1]$, $\mathcal{F}$–Borel σ-algebra, $\mathbb{P} = \text{Leb}$ and

$$\tau(\omega) = 2\omega \mod 1 = \begin{cases} 2\omega, & \omega < 1/2 \\ 2\omega - 1, & \omega \ge 1/2 \end{cases}$$

  It is easy to show that $Y(\omega) = 1\{\omega < 1/2\}$ is a generator and that $\mathbb{Y}$ is an i.i.d. Bernoulli(1/2) process. Thus, we get that Kolmogorov-Sinai entropy is $\mathcal{H}(\tau) = \log 2$.

- Let Ω be the unit circle $\mathbb{S}^1$, $\mathcal{F}$ – Borel σ-algebra, $\mathbb{P}$ be the normalized length and

$$\tau(\omega) = \omega + \gamma$$

  i.e. τ is a rotation by the angle γ. (When $\frac{\gamma}{2\pi}$ is irrational, this is known to be an ergodic p.p.t.). Here $Y = 1\{|\omega| < 2\pi\epsilon\}$ is a generator for arbitrarily small ϵ and hence

$$\mathcal{H}(\tau) \le H(\mathbb{X}) \le H(Y_0) = h(\epsilon) \to 0 \qquad \text{as } \epsilon \to 0 .$$

  This is an example of a zero-entropy p.p.t.

Remark 8.2. Two p.p.t.'s $(\Omega_1, \tau_1, \mathbb{P}_1)$ and $(\Omega_0, \tau_0, \mathbb{P}_0)$ are called isomorphic if there exists $f_i : \Omega_i \to \Omega_{1-i}$ defined $\mathbb{P}_i$-almost everywhere and such that 1) $\tau_{1-i} \circ f_i = f_{1-i} \circ \tau_i$; 2) $f_i \circ f_{1-i}$ is identity on Ω_i (a.e.); 3) $\mathbb{P}_i[f_{1-i}^{-1} E] = \mathbb{P}_{1-i}[E]$. It is easy to see that Kolmogorov-Sinai entropies of isomorphic p.p.t.s are equal. This observation was made by Kolmogorov in 1958. It was revoluationary, since it allowed to show that p.p.t.s corresponding shifts of iid Bern(1/2) and iid Bern(1/3) procceses are not isomorphic. Before, the only invariants known were those obtained from studying the spectrum of a unitary operator

$$U_\tau : L_2(\Omega, \mathbb{P}) \to L_2(\Omega, \mathbb{P}) \tag{8.16}$$
$$\phi(x) \mapsto \phi(\tau(x)) . \tag{8.17}$$

However, the spectrum of τ corresponding to any non-constant i.i.d. process consists of the entire unit circle, and thus is unable to distinguish Bern(1/2) from Bern(1/3).[2]

[2] To see the statement about the spectrum, let X_i be iid with zero mean and unit variance. Then consider $\phi(x_1^\infty)$ defined as $\frac{1}{\sqrt{m}} \sum_{k=1}^m e^{i\omega k} x_k$. This ϕ has unit energy and as $m \to \infty$ we have $\|U_\tau \phi - e^{i\omega}\phi\|_{L_2} \to 0$. Hence every $e^{i\omega}$ belongs to the spectrum of U_τ.

MIT OpenCourseWare
https://ocw.mit.edu

6.441 Information Theory
Spring 2016

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.