Chapter 29

Entropy Rates and Asymptotic Equipartition

Section 29.1 introduces the entropy rate — the asymptotic entropy per time-step of a stochastic process — and shows that it is well-defined; and similarly for information, divergence, etc. rates.

Section 29.2 proves the Shannon-MacMillan-Breiman theorem, a.k.a. the asymptotic equipartition property, a.k.a. the entropy ergodic theorem: asymptotically, almost all sample paths of a stationary ergodic process have the same log-probability per time-step, namely the entropy rate. This leads to the idea of "typical" sequences, in Section 29.2.1.

Section 29.3 discusses some aspects of asymptotic likelihood, using the asymptotic equipartition property, and allied results for the divergence rate.

29.1 Information-Theoretic Rates

Definition 376 (Entropy Rate) *The* entropy rate *of a random sequence X is*

$$h(X) \equiv \lim_n H_\rho[X_1^n]n \tag{29.1}$$

when the limit exists.

Definition 377 (Limiting Conditional Entropy) *The* limiting conditional entropy *of a random sequence X is*

$$h'(X) \equiv \lim_n H_\rho[X_n|X_1^{n-1}] \tag{29.2}$$

when the limit exists.

Lemma 378 *For a stationary sequence, $H_\rho[X_n|X_1^{n-1}]$ is non-increasing in n. Moreover, its limit exists if X takes values in a discrete space.*

PROOF: Because "conditioning reduces entropy", $H_\rho[X_{n+1}|X_1^n] \leq H[X_{n+1}|X_2^n]$. By stationarity, $H_\rho[X_{n+1}|X_2^n] = H_\rho[X_n|X_1^{n-1}]$. If X takes discrete values, then conditional entropy is non-negative, and a non-increasing sequence of non-negative real numbers always has a limit. □

Remark: Discrete values are a *sufficient* condition for the existence of the limit, not a necessary one.

We now need a natural-looking, but slightly technical, result from real analysis.

Theorem 379 (Cesàro) *For any sequence of real numbers $a_n \rightarrow a$, the sequence $b_n = n^{-1}\sum_{i=1}^{n} a_n$ also converges to a.*

PROOF: For every $\epsilon > 0$, there is an $N(\epsilon)$ such that $|a_n - a| < \epsilon$ whenever $n > N(\epsilon)$. Now take b_n and break it up into two parts, one summing the terms below $N(\epsilon)$, and the other the terms above.

$$\lim_n |b_n - a| = \lim_n \left| n^{-1}\sum_{i=1}^{n} a_i - a \right| \tag{29.3}$$

$$\leq \lim_n n^{-1}\sum_{i=1}^{n} |a_i - a| \tag{29.4}$$

$$\leq \lim_n n^{-1}\left(\sum_{i=1}^{N(\epsilon)} |a_i - a| + (n - N(\epsilon))\epsilon\right) \tag{29.5}$$

$$\leq \lim_n n^{-1}\left(\sum_{i=1}^{N(\epsilon)} |a_i - a| + n\epsilon\right) \tag{29.6}$$

$$= \epsilon + \lim_n n^{-1}\sum_{i=1}^{N(\epsilon)} |a_i - a| \tag{29.7}$$

$$= \epsilon \tag{29.8}$$

Since ϵ was arbitrary, $\lim b_n = a$. □

Theorem 380 (Entropy Rate) *For a stationary sequence, if the limiting conditional entropy exists, then it is equal to the entropy rate, $h(X) = h'(X)$.*

PROOF: Start with the chain rule to break the joint entropy into a sum of conditional entropies, use Lemma 378 to identify their limit as $h^{]prime}(X)$, and then use Cesàro's theorem:

$$\begin{aligned} h(X) &= \lim_n \frac{1}{n} H_\rho[X_1^n] && (29.9)\\ &= \lim_n \frac{1}{n} \sum_{i=1}^{n} H_\rho[X_i|X_1^{i-1}] && (29.10)\\ &= h'(X) && (29.11) \end{aligned}$$

as required. □

Because $h(X) = h'(X)$ for stationary processes (when both limits exist), it is not uncommon to find what I've called the limiting conditional entropy referred to as the entropy rate.

Lemma 381 *For a stationary sequence $h(X) \leq H[X_1]$, with equality iff the sequence is IID.*

PROOF: Conditioning reduces entropy, unless the variables are independent, so $H[X_n|X_1^{n-1}] < H[X_n]$, unless $X_n \perp\!\!\!\perp X_1^{n-1}$. For this to be true of all n, which is what's needed for $h(X) = H[X_1]$, all the values of the sequence must be independent of each other; since the sequence is stationary, this would imply that it's IID. □

Example 382 (Markov Sequences) *If X is a stationary Markov sequence, then $h(X) = H_\rho[X_2|X_1]$, because, by the chain rule, $H_\rho[X_1^n] = H_\rho[X_1] + \sum_{t=2}^{n} H_\rho[X_t|X_1^{t-1}]$. By the Markov property, however, $H_\rho[X_t|X_1^{t-1}] = H_\rho[X_t|X_{t-1}]$, which by stationarity is $H_\rho[X_2|X_1]$. Thus, $H_\rho[X_1^n] = H_\rho[X_1] + (n-1)H_\rho[X_2|X_1]$. Dividing by n and taking the limit, we get $H_\rho[X_1^n] = H_\rho[X_2|X_1]$.*

Example 383 (Higher-Order Markov Sequences) *If X is a k^{th} order Markov sequence, then the same reasoning as before shows that $h(X) = H_\rho[X_{k+1}|X_1^k]$ when X is stationary.*

Definition 384 (Divergence Rate) *The* divergence rate *or* relative entropy rate *of the infinite-dimensional distribution Q from the infinite-dimensional distribution P, $d(P\|Q)$, is*

$$d(P\|Q) = \lim_n \mathbf{E}_P \left[\log \left(\left. \frac{dP}{dQ} \right|_{\sigma(X_{-n}^0)} \right) \right] \qquad (29.12)$$

if all the finite-dimensional distributions of Q dominate all the finite-dimensional distributions of P. If P and Q have densities, respectively p and q, with respect to a common reference measure, then

$$d(P\|Q) = \lim_n \mathbf{E}_P \left[\log \frac{p(X_0|X_{-n}^{-1})}{q(X_0|X_{-n}^{-1})} \right] \qquad (29.13)$$

29.2 The Shannon-McMillan-Breiman Theorem or Asymptotic Equipartition Property

This is a central result in information theory, acting as a kind of ergodic theorem for the entropy. That is, we want to say that, for almost all ω,

$$-\frac{1}{n}\log \mathbb{P}\left(X_1^n(\omega)\right) \rightarrow \lim_n \frac{1}{n}\mathbf{E}\left[-\log \mathbb{P}\left(X_1^n\right)\right] = h(X)$$

At first, it looks like we should be able to make a nice time-averaging argument. We can always factor the joint probability,

$$\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) = \frac{1}{n}\sum_{t=1}^{n}\log \mathbb{P}\left(X_t|X_1^{t-1}\right)$$

with the understanding that $\mathbb{P}\left(X_1|X_1^0\right) = \mathbb{P}\left(X_1\right)$. This looks rather like the sort of Cesàro average that we became familiar with in ergodic theory. The problem is, there we were averaging $f(T^t\omega)$ for a *fixed* function f. This is not the case here, because we are conditioning on long and longer stretches of the past. There's no problem if the sequence is Markovian, because then the remote past is irrelevant, by the Markov property, and we can just condition on a fixed-length stretch of the past, so we're averaging a fixed function shifted in time. (This is why Shannon's original argument was for Markov chains.) The result nonetheless more broadly, but requires more subtlety than might otherwise be thought. Breiman's original proof of the general case was fairly involved[1], requiring both martingale theory, and a sort of dominated convergence theorem for ergodic time averages. (You can find a simplified version of his argument in Kallenberg, at the end of chapter 11.) We will go over the "sandwiching" argument of Algoet and Cover (1988), which is, to my mind, more transparent.

The idea of the sandwich argument is to show that, for large n, $-n^{-1}\log \mathbb{P}\left(X_1^n\right)$ must lie between an upper bound, h_k, obtained by approximating the sequence by a Markov process of order k, and a lower bound, which will be shown to be h. Once we establish that $h_k \downarrow h$, we will be done.

Definition 385 (Markov Approximation) *For each k, define the order k Markov approximation to X by*

$$\mu_k(X_1^n) = \mathbb{P}\left(X_1^k\right)\prod_{t=k+1}^{n}\mathbb{P}\left(X_t|X_{t-k}^{t-1}\right) \qquad (29.14)$$

μ_k is the distribution of a stationary Markov process of order k, where the distribution of X_1^{k+1} matches that of the original process.

[1] Notoriously, the proof in his original paper was actually invalid, forcing him to publish a correction.

Lemma 386 *For each k, the entropy rate of the order k Markov approximation is is equal to $H[X_{k+1}|X_1^k]$.*

PROOF: Under the approximation (but not under the original distribution of X), $H[X_1^n] = H[X_1^k] + (n-k)H[X_{k+1}|X_1^k]$, by the Markov property and stationarity (as in Examples 382 and 383). Dividing by n and taking the limit as $n \to \infty$ gives the result. □

Lemma 387 *If X is a stationary two-sided sequence, then $Y_t = f(X_{-\infty}^t)$ defines a stationary sequence, for any measurable f. If X is also ergodic, then Y is ergodic too.*

PROOF: Because X is stationary, it can be represented as a measure-preserving shift on sequence space. Because it is measure-preserving, $\theta X_{-\infty}^t \overset{d}{=} X_{-\infty}^t$, so $Y(t) \overset{d}{=} Y(t+1)$, and similarly for all finite-length blocks of Y. Thus, all of the finite-dimensional distributions of Y are shift-invariant, and these determine the infinite-dimensional distribution, so Y itself must be stationary.

To see that Y must be ergodic if X is ergodic, recall that a random sequence is ergodic iff its corresponding shift dynamical system is ergodic. A dynamical system is ergodic iff all invariant functions are a.e. constant (Theorem 304). Because the Y sequence is obtained by applying a measurable function to the X sequence, a shift-invariant function of the Y sequence is a shift-invariant function of the X sequence. Since the latter are all constant a.e., the former are too, and Y is ergodic. □

Lemma 388 *If X is stationary and ergodic, then, for every k,*

$$\mathbb{P}\left(\lim_n -\frac{1}{n}\log \mu_k(X_1^n(\omega)) = h_k\right) = 1 \tag{29.15}$$

i.e., $-\frac{1}{n}\log \mu_k(X_1^n(\omega))$ converges a.s. to h_k.

PROOF: Start by factoring the approximating Markov measure in the way suggested by its definition:

$$-\frac{1}{n}\log \mu_k(X_1^n) = -\frac{1}{n}\log \mathbb{P}\left(X_1^k\right) - \frac{1}{n}\sum_{t=k+1}^{n} \log \mathbb{P}\left(X_t|X_{t-k}^{t-1}\right) \tag{29.16}$$

As n grows, $\frac{1}{n}\log \mathbb{P}\left(X_1^k\right) \to 0$, for every fixed k. On the other hand, $-\log \mathbb{P}\left(X_t|X_{t-k}^{t-1}\right)$ is a measurable function of the past of the process, and since X is stationary and ergodic, it, too, is stationary and ergodic (Lemma 387). So

$$\begin{aligned}
-\frac{1}{n}\log \mu_k(X_1^n) &\to -\frac{1}{n}\sum_{t=k+1}^{n} \log \mathbb{P}\left(X_t|X_{t-k}^{t-1}\right) && (29.17)\\
&\overset{a.s.}{\to} \mathbf{E}\left[-\log \mathbb{P}\left(X_{k+1}|X_1^k\right)\right] && (29.18)\\
&= h_k && (29.19)
\end{aligned}$$

by Theorem 312. □

Definition 389 *The* infinite-order approximation *to the entropy rate of a discrete-valued stationary process X is*

$$h_\infty(X) \equiv \mathbf{E}\left[-\log \mathbb{P}\left(X_0|X_{-\infty}^{-1}\right)\right] \tag{29.20}$$

Lemma 390 *If X is stationary and ergodic, then*

$$\lim_n -\frac{1}{n}\log \mathbb{P}\left(X_1^n|X_{-\infty}^0\right) = h_\infty \tag{29.21}$$

almost surely.

Proof: Via Theorem 312 again, as in Lemma 388. □

Lemma 391 *For a stationary, ergodic, finite-valued random sequence, $h_k(X) \downarrow h_\infty(X)$.*

Proof: By the martingale convergence theorem, for every $x_0 \in \Xi$,

$$\mathbb{P}\left(X_0 = x_0|X_n^{-1}\right) \overset{a.s.}{\rightarrow} \mathbb{P}\left(X_0 = x_0|X_\infty^{-1}\right) \tag{29.22}$$

Since Ξ is finite, the probability of any point in Ξ is between 0 and 1 inclusive, and $p \log p$ is bounded and continuous. So we can apply bounded convergence to get that

$$\begin{aligned} h_k &= \mathbf{E}\left[-\sum_{x_0}\mathbb{P}\left(X_0 = x_0|X_{-k}^{-1}\right)\log\mathbb{P}\left(X_0 = x_0|X_{-k}^{-1}\right)\right] && (29.23)\\ &\rightarrow \mathbf{E}\left[-\sum_{x_0}\mathbb{P}\left(X_0 = x_0|X_{-\infty}^{-1}\right)\log\mathbb{P}\left(X_0 = x_0|X_{-\infty}^{-1}\right)\right] && (29.24)\\ &= h_\infty && (29.25)\end{aligned}$$

Lemma 392 *$h_\infty(X)$ is the entropy rate of X, i.e. $h_\infty(X) = h(X)$.*

Proof: Clear from Theorem 380 and the definition of conditional entropy. □

We are almost ready for the proof, but need one technical lemma first.

Lemma 393 *If $R_n \geq 0$, $\mathbf{E}[R_n] \leq 1$ for all n, then*

$$\limsup_n \frac{1}{n}\log R_n \leq 0 \tag{29.26}$$

almost surely.

Proof: Pick any $\epsilon > 0$.

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n}\log R_n \geq \epsilon\right) &= \mathbb{P}\left(R_n \geq e^{n\epsilon}\right) && (29.27)\\ &\leq \frac{\mathbf{E}[R_n]}{e^{n\epsilon}} && (29.28)\\ &\leq e^{-n\epsilon} && (29.29)\end{aligned}$$

by Markov's inequality. Since $\sum_n e^{-n\epsilon} \leq \infty$, by the Borel-Cantelli lemma, $\limsup_n n^{-1}\log R_n \leq \epsilon$. Since ϵ was arbitrary, this concludes the proof. □

Theorem 394 (Asymptotic Equipartition Property) *For a stationary, ergodic, finite-valued random sequence X,*

$$-\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) \rightarrow h(X) \ a.s. \tag{29.30}$$

Proof: For every k, $\mu_k(X_1^n)/\mathbb{P}\left(X_1^n\right) \geq 0$, and $\mathbf{E}\left[\mu_k(X_1^n)/\mathbb{P}\left(X_1^n\right)\right] \leq 1$. Hence, by Lemma 393,

$$\limsup_n \frac{1}{n}\log\frac{\mu_k(X_1^n)}{\mathbb{P}\left(X_1^n\right)} \leq 0 \tag{29.31}$$

a.s. Manipulating the logarithm,

$$\limsup_n \frac{1}{n}\log \mu_k(X_1^n) \leq -\limsup_n -\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) \tag{29.32}$$

From Lemma 388, $\limsup_n \frac{1}{n}\log\mu_k(X_1^n) = \lim_n \frac{1}{n}\log\mu_k(X_1^n) = -h_k(X)$, a.s. Hence, for each k,

$$h_k(X) \geq \limsup_n -\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) \tag{29.33}$$

almost surely.

A similar manipulation of $\mathbb{P}\left(X_1^n\right)/\mathbb{P}\left(X_1^n|X_{-\infty}^0\right)$ gives

$$h_\infty(X) \leq \liminf_n -\frac{1}{n}\log \mathbb{P}\left(X_1^n\right) \tag{29.34}$$

a.s.

As $h_k \downarrow h_\infty$, it follows that the liminf and the limsup of the normalized log likelihood must be equal almost surely, and so equal to h_∞, which is to say to $h(X)$. □

Why is this called the AEP? Because, to within an $o(n)$ term, all sequences of length n have the same log-likelihood (to within factors of $o(n)$, if they have positive probability at all. In this sense, the likelihood is "equally partitioned" over those sequences.

29.2.1 Typical Sequences

Let's turn the result of the AEP around. For large n, the probability of a given sequence is either approximately 2^{-nh} or approximately zero[2]. To get the total probability to sum up to one, there need to be about 2^{nh} sequences with positive probability. If the size of the alphabet is s, then the fraction of sequences which are actually exhibited is $2^{n(h-\log s)}$, an increasingly small fraction (as $h \leq \log s$). Roughly speaking, these are the *typical* sequences, any one of which, via ergodicity, can act as a representative of the complete process.

[2] Of course that assumes using base-2 logarithms in the definition of entropy.

29.3 Asymptotic Likelihood

29.3.1 Asymptotic Equipartition for Divergence

Using methods analogous to those we employed on the AEP for entropy, it is possible to prove the following.

Theorem 395 *Let P be an asymptotically mean-stationary distribution, with stationary mean $\overline{P}$, with ergodic component function ϕ. Let M be a homogeneous finite-order Markov process, whose finite-dimensional distributions dominate those of P and $\overline{P}$; denote the densities with respect to M by p and $\overline{p}$, respectively. If $\lim_n n^{-1} \log \overline{p}(X_1^n)$ is an invariant function $\overline{P}$-a.e., then*

$$-\frac{1}{n} \log p(X_1^n(\omega)) \overset{a.s.}{\to} d(\overline{P}_{\phi(\omega)} \| M) \tag{29.35}$$

where $\overline{P}_{\phi(\omega)}$ is the stationary, ergodic distribution of the ergodic component.

Proof: See Algoet and Cover (1988, theorem 4), Gray (1990, corollary 8.4.1).

Remark. The usual AEP is in fact a consequence of this result, with the appropriate reference measure. (Which?)

29.3.2 Likelihood Results

It is left as an exercise for you to obtain the following result, from the AEP for relative entropy, Lemma 367 and the chain rules.

Theorem 396 *Let P be a stationary and ergodic data-generating process, whose entropy rate, with respect to some reference measure ρ, is h. Further let M be a finite-order Markov process which dominates P, whose density, with respect to the reference measure, is m. Then*

$$-\frac{1}{n} \log m(X_1^n) \to h + d(P\|M) \tag{29.36}$$

P-almost surely.

29.4 Exercises

Exercise 29.1 *Markov approximations are maximum-entropy approximations. (You may assume that the process X takes values in a finite set.)*

a Prove that μ_k, as defined in Definition 385, gets the distribution of sequences of length $k+1$ correct, i.e., for any set $A \in \mathcal{X}^{k+1}$, $\nu(A) = \mathbb{P}\left(X_1^{k+1} \in A\right)$.

b Prove that $\mu_{k'}$, for any any $k' > k$, also gets the distribution of length $k+1$ sequences right.

c *In a slight abuse of notation, let $H[\nu(X_1^n)]$ stand for the entropy of a sequence of length n when distributed according to ν. Show that $H[\mu_k(X_1^n)] \geq H[\mu_{k'}(X_1^n)]$ if $k' > k$. (Note that the $n \leq k$ case is easy!)*

d *Is it true that that if ν is any other measure which gets the distribution of sequences of length $k+1$ right, then $H[\mu_k(X_1^n)] \geq H[\nu(X_1^n)]$? If yes, prove it; if not, find a counter-example.*

Exercise 29.2 *Prove Theorem 396.*