

ECE 7680

Lecture 5 – Entropy Rates

Objective: The concept of entropy rate allows us to talk about entropy for sequences of random variables that are not independent.

Up to this point we have made the assumption that the random variables that we have been dealing with have been **independent** and identically distributed. Of course, in the real world, independence is not commonly encountered: the letters emerging from a stream of text are not independent.

In this lecture we will introduce the means of treating sequences of *dependent* random variables.

Definition 1 The **entropy rate** of a stochastic process $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

provided that the limit exists. □

There is also a related quantity for entropy:

$$H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, X_{n-2}, \dots, X_1).$$

These are two different concepts of entropy: the first is the (average) per-symbol entropy of all n random variables. The second is the conditional entropy, conditioned upon all prior random variables. However, (and somewhat surprisingly), for stationary sequences these are the same:

Theorem 1 *For a stationary stochastic process, the two defined entropy rates exist and are equal:*

$$H(\mathcal{X}) = H'(\mathcal{X}).$$

Before proving this, we will prove another necessary result: $\lim_{n \rightarrow \infty} H(X_n | X_{n-1}, X_{n-2}, \dots, X_1)$ exists:

Theorem 2 *For a stationary random process, $H(X_n | X_{n-1}, X_{n-2}, \dots, X_1)$ is decreasing in n and has a limit $H'(\mathcal{X})$.*

Proof

$$\begin{aligned} H(X_{n+1} | X_1, X_2, \dots, X_n) &\leq H(X_{n+1} | X_n, \dots, X_2) \\ &= H(X_n | X_{n-1}, \dots, X_1). \end{aligned}$$

(Conditioning reduces entropy; stationarity). Therefore $H(X_n | X_{n-1}, X_{n-2}, \dots, X_1)$ is a decreasing sequence of non-negative numbers, and must have a limit. □

And now a result from analysis:

Theorem 3 (*Cesaro mean*) *If $a_n \rightarrow a$ and $b_n = \frac{1}{n} \sum_{i=1}^n a_i$ then $b_n \rightarrow a$.*

The idea is that since most of the numbers in a_n are eventually close to a , then b_n , which is the average of the first n terms must also be close to a : as n gets large, the first terms become increasingly less important.

The proof is a lot of the ϵ -ish sort of stuff that analysis thrive on and most of us simply tolerate at best:

Proof Since $a_n \rightarrow a$, then for any $\epsilon > 0$ there is a number $N(\epsilon)$ such that for $n \geq N(\epsilon)$, $|a_n - a| \leq \epsilon$ (definition of convergence). Hence

$$\begin{aligned} |b_n - a| &= \left| \frac{1}{n} \sum_{i=1}^n a_i - a \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n |a_i - a| \quad \text{triangle inequality} \\ &\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{n - N(\epsilon)}{n} \epsilon \\ &\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon \end{aligned}$$

The first term goes to 0 as $n \rightarrow \infty$ (the summation is bounded, but n can grow). So the difference $|b_n - a|$ can be made as small as desired. \square

Now we can prove the equality $H(X) = H'(X)$.

Proof By the chain rule,

$$\frac{H(X_1, \dots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1).$$

But we know that the conditional entropies have a limit. By the Cesaro mean, we know that for a sequence with a limit, the running average of the sequence has a limit, which is equal to the limit of the conditional entropies. Hence

$$H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{H(X_1, \dots, X_n)}{n} = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1) = H'(\mathcal{X}).$$

\square

The generalization of the AEP theorem of the last chapter is true (but we won't prove it here): for a sequence of identically distributed (but not necessarily independent r.v.s),

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \rightarrow H(\mathcal{X}).$$

with probability 1 (strong convergence!). Based on this generalization, it is possible to define a notion of typical sequences, and determine the number of typical sequences (approximately $2^{nH(\mathcal{X})}$), each with probability about $2^{-nH(\mathcal{X})}$. A representation therefore exists which requires approximately $nH(\mathcal{X})$ bits.

There is a lot more material in the chapter about Markov processes and Hidden Markov models. However, in the interest of moving toward our goal, I will not talk about it in class.