Electrical Engineering 229A Lecture 8 Notes

Daniel Raban

September 21, 2021

1 Entropy Rate, Markov Processes, and Data Compression for Sequences

1.1 Entropy rate

Last time, we introduced the entropy rate of a stationary stochastic process. If $\mathscr{X}$ is a finite or countably infinite set, a **stationary stochastic process** is a sequence of random variables $(X_k)_{k=-\infty}^{\infty}$ with the property that

$$\mathbb{P}(X_k = x_0, X_{k+1} = x_1, \ldots, X_{k+t} = x_t)$$

does not depend on k (for all $t \geq 0$, $x_0, \ldots, x_t$). The **entropy rate** of the process is the limit

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

where the limit exists because $H(X_1) \geq H(X_2 \mid X_1) \geq \cdots \geq H(X_n \mid X_1, \ldots, X_{n-1})$, and in fact,

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n\to\infty} H(X_n \mid X_1, \ldots, X_{n-1})$$

because of the chain rule, $H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1^{n-1})$. We can think of the entropy rate as the asymptotic amount of information we learn from the next random variable in the sequence.

Observe that

$$p(x_1, \ldots, x_t) = p(x_1)p(x_2 \mid x_1) \cdots p(x_t \mid x_1, \ldots, x_{t-1}).$$

For large t and $k < t$,

$$p(x_1)p(x_2 \mid x_1)p(x_k \mid x_1, \ldots, x_{k-1}) \prod_{j=1}^{t-k} p(x_{k+j} \mid x_{k+j-t+1}, \ldots, x_{k+j-1})$$

might be a decent approximation from a modeling point of view. This defines a $(k-1)$-order stationary Markov process.

A **first order Markov process**[1] is defined by the transition probabilities $p(x_2 \mid x_1)$ for $x_1, x_2 \in \mathscr{X}$ and an initial distribution $(p(x), x \in \mathscr{X})$. A **stationary Markov process** is defined by the transition probabilities and a probability distribution $(\pi(x), x \in \mathscr{X})$ such that $\sum_{x \in \mathscr{X}} \pi(x) p(y \mid x) = \pi(y)$ for all $y \in \mathscr{X}$. This would mean that

$$p(x_1, \ldots, x_t) = \pi(x_1) p(x_2 \mid x_1) p(x_3 \mid x_2) \cdots p(x_t \mid x_{t-1}).$$

For a k-th order Markov process, we need $p(x_{k+1} \mid x_1, \ldots, x_k)$ with $x_i \in \mathscr{X}$ $i = 1, \ldots, k+1$ and an initial distribution $(p(x_1, \ldots, x_k), x_1^k \in \mathscr{X}^k)$. For stationarity, we need a distribution $(\pi(x_1, \ldots, x_k), x_1^k \in \mathscr{X}^k)$ such that

$$\sum_{x_1, \ldots, x_k} \pi(x_1, \ldots, x_k) p(x_{k+1} \mid x_1, \ldots, x_k) = \pi(x_2, \ldots, x_{k+1}).$$

The entropy rate for a stationary Markov process is $H(X_2 \mid X_1)$, while the entropy rate for a k-th order stationary Markov process is

$$H(X_{k+1} \mid X_1, \ldots, X_k) = \sum_{x_1, \ldots, x_k} \pi(x_1, \ldots, x_k) H((p(x_{k+1} \mid x_1, \ldots, x_k), x_{k+1} \in \mathscr{X})).$$

For $k = 1$, this is

$$H(X_2 \mid X_1) = -\sum_{x,y} \pi(x) p(y \mid x) \log p(y \mid x)$$

1.2 Time reversal and reversible Markov processes

An important class of examples is reversible stationary Markov processes.

Definition 1.1. A stationary Markov process is **reversible** if

$$\pi(x) p(y \mid x) = \pi(y) p(x \mid y) \qquad \forall x, y \in \mathscr{X}.$$

For a general stationary Markov chain depending on the transition probability matrix $[p(y \mid x)]_{x,y \in \mathscr{X}}$ and stationay distribution $(\pi(x), x \in \mathscr{X})$, one can define

$$\widetilde{p}(y \mid x) := \frac{\pi(y) p(x \mid y)}{\pi(x)}$$

(assuming $\pi(x) > 0$ for all $x \in \mathscr{X}$). Then

$$\sum_{y \in \mathscr{X}} \widetilde{p}(y \mid x) = \frac{\sum_{y \in \mathscr{X}} \pi(y) p(x \mid y)}{\pi(x)} = \frac{\pi(x)}{\pi(x)} = 1$$

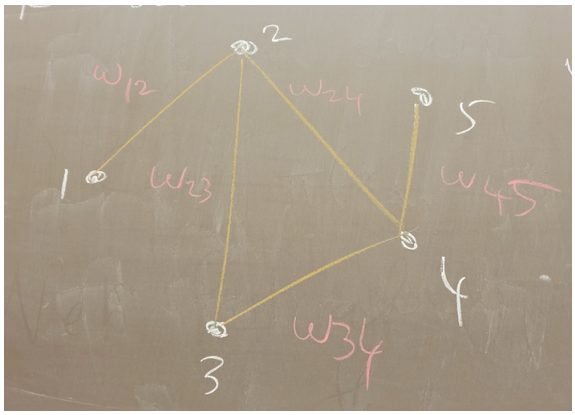
[1] For finite or countable state spaces, these are often referred to as "Markov chains."

and

$$\sum_{x\in\mathscr{X}} \pi(x)\widetilde{p}(y \mid x) = \sum_{x\in\mathscr{X}} \pi(y)p(x \mid y) = \pi(x),$$

so $[\widetilde{p}(y \mid x)]_{x,y\in\mathscr{X}}$ defines a transition probability matrix with stationary distribution $(\pi(x), x \in \mathscr{X})$. This is called the **time reversal** of the original process. A Markov process is time reversible if and only if its time reversal has the same joint distributions as as the original process.

Example 1.1. Stationary random walks on weighted graphs give rise to examples.



At any time t, X_t belongs to the set of vertices, and

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = \frac{w_{i,j}}{\sum_{k\in V} w_{i,k}}.$$

The stationary distribution will be

$$\pi(i) = \frac{\sum_{j\in V} w_{i,j}}{2\sum_{i,j} w_{i,j}},$$

and this process is reversible.

This is of huge importance in algorithms, and it has connections to resistive network theory.[2]

1.3 Overview of data compression for sequences

The next 2-3 lectures will be about various schemes for lossless data compression. The goal is to represent observed data efficiently (using as few bits /symbols as possible). We have already seen, for example, that if $X_1, X_2, \ldots$ are iid with marginal distribution

[2]This is covered in a book by Doyle an Snell called *Random Walks and Electrical Networks*.

$(p(x), x \in \mathscr{X})$, there exists an encoding map $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ and a decompression map $d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathscr{X}^n$ (for each $n \geq 1$) such that $d_n \circ e_n$ is the identity map and

$$\frac{1}{n}\,\mathbb{E}[\text{length}(e_n(X_1, \ldots, X_n))] \leq H + \varepsilon.$$

Morover, we have also seen that for any $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ and $d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathscr{X}^n$ with $d_n \circ e_n = \text{id}$, for any $\varepsilon > 0$,

$$\frac{1}{n}\,\mathbb{E}[\text{length}(e_n(X_1, \ldots, X_n))] \geq H - \varepsilon.$$

There is a book called *Handbook of Data Compression* by Salamon which discusses this.[3]

We would like a version of this for stationary processes. We'll see this as we go along, but here are some big picture facts related to this.

We cannot get an analog of the Strong Law of Large Numbers for stationary processes without assuming an additional condition called **ergodicity** which excludes examples like $p(\ldots, X_0 = 1, \ldots, X_t = 1) = \mathbb{P}(X_0 = 0, \ldots, X_t = 0) = 1/2$ for all $t \geq 0$.

For a stationary ergodic process, we have (under some conditions) the pointwise ergodic theorem:

Theorem 1.1 (Pointwise ergodic theorem, Birkhoff)**.** *Let $(X_k)_{k=-\infty}^{\infty}$ be a stationary, ergodic process with random variables taking values in $\mathscr{X}$. Given $f : \mathscr{X} \to \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \mathbb{E}[f(X_1)]$$

almost surely.

But even this is not enough for us to replace the Strong Law of Large Numbers applied to information densities. We need a further statement:

Theorem 1.2 (Shannon-McMillan-Breiman)**.** *If $(X_k)_{k=-\infty}^{\infty}$ is a stationary, ergodic process,*

$$\lim_{n\to\infty} -\frac{1}{n} \log p(X_1, \ldots, X_n) = \textit{entropy rate of process}$$

almost surely.

From a practical point of viewpoint, $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ needs to be construction from "smaller pieces." For example, start with $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ and define $e_n(x_1, \ldots, x_n) = e(x_1)e(x_2)\cdots e(x_n)$. This function e needs to be 1 to 1 for invertibility. But even if $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ is 1 to 1, e_n might not be.

[3] The book is on the order of 1000 pages long.

Example 1.2. Let $\mathscr{X} = \{1, 2, 3\}$ with $e(1) = 0$, $e(2) = 00$, and $e(3) = 1$. Then

$$e_3(12) = 000, \qquad e_3(2, 1) = 000.$$

Definition 1.2. $e : \mathscr{X} \to \{0, 1\}^* \setminus \{\varnothing\}$ is called **uniquely decodable** if each e_n is one to one.

One way to get this property is to make e **instantaneous** (or **prefix-free**) if no $e(x)$ is a prefix of $e(y)$ for $x \neq y$.

Example 1.3. If $\mathscr{X} = \{1, 2, 3, 4\}$, we can take

$$e(1) = 1, \qquad e(2) = 01, \qquad e(3) = 001, \qquad e(4) = 000.$$