EL-GY 6063: Information Theory
Lecture 7

April 2, 2019

1 Markov Chains

Reference: Chapter 4 of Elements of Information Theory

Definition 1 (Stochastic Process)**.** *A stochastic process (also called a random process) is an indexed sequence of random variables.*

Example. Let X_i be the indicators of a head in consecutive throws of a coin. Then, the sequence $X_1, X_2, \cdots$ is a random process. This is called the Bernoulli process with parameter p.
Example. Assume that the position of a pointer at time $t = 1$ is $X_1 = 0$, and that at time $t > 1$ the pointer moves to the right with probability p and to the left with probability $q = 1-p$. Let $X_t, t \geq 1$ be the position of the pointer at time t (note that we are assuming that time is discrete taking values $t \in \{1, 2, 3, \cdots\}$. Then $X_1, X_2, \ldots$ is a stochastic process. This is called a simple random walk.

Note that $P_{X_3|X_2,X_1}(x_3|x_2, x_1) = P_{X_3|X_2}(x_3|x_2)$. More generally, the position at any time t only depends on the position at time $t-1$.
Example. Assume that $X_t, t \in \{1, 2, \cdots\}$ is a sequence of independent Gaussian random variables with zero mean and variance t^2 (i.e. $X_t \sim N(0, t^2)$), then $X_1, X_2, \cdots$ is a stocastic process.

Definition 2. *A stochastic process is said to be stationary if the joint distribution of any subsequence of random variables is $X_1, X_2, \cdots, X_t$ is invariant with respect to shifts in the time index; that is,*

$$\forall t, l \in \mathbb{N}, : P_{X_1,X_2,\cdots,X_t}(x_1, x_2, \cdots, x_t) = P_{X_{1+l},X_{2+l},\cdots,X_{t+l}}(x_1, x_2, \cdots, x_t).$$

In the previous three examples, the first two (Bernoulli process and the simple random walk) are stationary whereas the third example is not stationary.

In all of the above examples, we have assumed that the process is discrete-time (i.e. the index t in X_t takes discrete values). Generally, in this course we will only talk about discrete-time processes. Our first two examples are discrete valued processes since the variable X_i takes a discrete set of values. The third example is a continuous-valued random process.

We are ready to define a Markov process:

Definition 3. *A discrete-time, discrete-valued stochastic process $X_1, X_2, \cdots$ is said to be an l_{th} order Markov Chain or a Markov Process if*

$$\forall n \in \mathbb{N} : P_{X_{n+1}|X_1,X_2,\cdots,X_n}(x_{n+1}|x_1, x_2, \cdots, x_n) = P_{X_{n+1}|X_n,X_{n-1},\cdots,X_{n-l+1}}(x_{n+1}|x_1, x_2, \cdots, x_{n-l+1}).$$

Particularly, for a first order Markov Chain:

$$\forall n \in \mathbb{N} : P_{X_{n+1}|X_1,X_2,\cdots,X_n}(x_{n+1}|x_1, x_2, \cdots, x_n) = P_{X_{n+1}|X_n}(x_{n+1}|x_n).$$

In this class we mainly study first order Markov Chains. For a first order Markov Chain, the value at time t only depends on the the previous realization at time $t-1$.

The simple random walk in the second example is a first order Markov Chain.

Definition 4. *For a First order Markov chain $X_1, X_2, \cdots$, the variable X_i is called the state of the Markov chain at time i. We assume that the alphabet of X_i is fixed and is equal to $\mathcal{X} = \{1, 2, \cdots, r\}$. If $X_n = j$, we say that the Markov Chain is at state j at time n.*

Example. Assume that in the random walk $X_1 = 0$, $X_2 = 1$, $X_3 = 2$, $X_4 = 1$. In that case the Markov chain starts at state 0, goes to state 1 at time 1, then goes to state 2 and then at time 3 returns to state 1.

Generally, we assume that the probability of going from one state to the other is the same regardless of time.

Definition 5. *The Markov Chain is said to be time invariant if the conditional probability $P_{X_{n+1}|X_n}$ is the same for all n.*

A time-invariant Markov chain is characterized by two parameters: 1) Initial state X_1, 2) Transition matrix $\Pi = [P_{i,j}]_{i,j \in [1,r]}$, where $P_{i,j} = P_{X_{n+1}|X_n}(j|i)$.

Note that the initial state is a random variable. It has distribution P_{X_1}. The distribution P_{X_1} determines the probability that the Markov chain starts at any specific state.

Properties of the transition matrix:
1) $P_{i,j} \geq 0, \forall i, j$. This is true since probabilities are always positive.
2) $\sum_{j=1}^{r} P_{i,j} = 1, \forall i$. This means that the probability that the Markov chain will go from state i to any of the states $i \in [1, r]$ is equal to 1.

Let $w_j^{(n)} = P(X_n = j)$, that is $w_j^{(n)}$ is the probability that the Markov chain is at state j at time n. Then,

$$\begin{aligned} w_j^{(n)} = P(X_n = j) &= \sum_{k=1}^{r} P(X_n = j|X_{n-1} = k)P(X_{n-1} = k) \\ &= \sum_{k=1}^{r} P_{k,j} w_k^{(n-1)}. \end{aligned}$$

This looks like matrix multiplication. Particularly, let $\underline{w}^{(n)} = [w_1^{(n)} w_2^{(n)} \cdots w_r^{(n)}]$ be the vector of probabilities of the state of the Markov chain at time r. Then,

$$\underline{w}^{(n)} = \underline{w}^{(n-1)}\Pi.$$

So, the vector of probabilities of the states of the Markov chain at time n can be computed by multiplying the transition matrix with the vector of probabilities of the states of the Markov chain at time $n - 1$. This process can be further continued:

$$\underline{w}^{(n)} = \underline{w}^{(n-1)}\Pi = \underline{w}^{(n-2)}\Pi^2 = \cdots = \underline{w}^{(1)}\Pi^{n-1}.$$

So, the vector of probabilities can be written as a product of the transition probability matrix and the vector of probabilities of the initial state.

Define $P_{i,j}^{(2)} = P(X_{n+2} = j|X_n = i)$, then:

$$P_{i,j}^{(2)} = P(X_{n+2} = j|X_n = i) = \sum_{k} P(X_{n+1} = k|X_n = i)P(X_{n+2} = j|X_{n+1} = k) = \sum_{k=1}^{n} P_{i,k}P_{k,j}.$$

This also reminds us of matrix multiplication. Define the matrix $\Pi^{(2)} = [P_{i,j}^{(2)}]_{i,j \in [1,n]}$. Then, from the above $\Pi^{(2)} = \Pi \times \Pi = \Pi^2$.

Similarly, $\Pi^{(n)} = \Pi^n$, where Π is the matrix of n step transition probabilities.

Next, we define the stationary distribution of a Markov Chain:

Definition 6. *The distribution $\underline{w} = [w_1, w_2, \cdots, w_r]$ is called the stationary distribution of the Markov chain if $\underline{w} = \underline{w}\Pi$.*

Note that if the Markov chain has initial state X_1 distributed according to $\underline{P}$, then the distribution of all X_i's is the same.

Example. Let $\Pi = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$. Find the stationary distribution of the matrix.

Solution.

$$\underline{w} = \underline{w}\Pi \Rightarrow [w_1 w_2 w_3] = [w_1 w_2 w_3] \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \Rightarrow \begin{cases} w_1 = \frac{1}{2}w_2 + \frac{1}{2}w_3 \\ w_2 = \frac{1}{3}w_2 + \frac{1}{2}w_3 \\ w_3 = w_1 + \frac{1}{6}w_2 \\ w_1 + w_2 + w_3 = 1 \end{cases} \Rightarrow \underline{w} = [\frac{1}{3}, \frac{2}{7}, \frac{8}{21}].$$

We are often interested in Markov chains for which the equation $\underline{w} = \underline{w}\Pi$ has a unique solution.

Lemma 1. *The Markov chain has a steady state solution if and only if there exists a positive integer N such that π^N has at least one column all of whose elements are positive.*

Example. Let $\Pi = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$. Then, $\Pi^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{7}{36} & \frac{10}{18} \\ \frac{1}{4} & \frac{1}{6} & \frac{7}{12} \end{bmatrix}$. So, the Markov chain has a unique stationary solution.

Definition 7. *Assume that for any (i,j) the limit $\lim_{n\to\infty} P_{i,j}^{(n)}$ exists. Then we know that $\lim_{n\to\infty} P_{i,j}^{(n)}$ is the same for all $i \in [1,r]$ and fixed j. In this case the vector $\underline{w} = [w_j]_{j\in[1,n]}$, where $w_j = \lim_{n\to\infty} P_{i,j}^{(n)}$ is called the steady state distribution of the Markov chain.*

Properties of the steady state solution:
1) $\sum_{i=1}^{r} w_i = 1$.
2) $\underline{w}$ is a stationary distribution of the Markov chain.
3) $\underline{w}$ is the unique stationary distribution for the Markov chain.
proof of 1): We know that $\sum_{j=1}^{r} P_{i,j}^{(n)} = 1, \forall i \in [1,n]$. So, $\lim_{n\to\infty} \sum_{j=1}^{r} P_{i,j}^{(n)} = 1 \Rightarrow \sum_{j=1}^{r} \lim_{n\to\infty} P_{i,j}^{(n)} = 1$. However, we know from property 1) that $w_j = \lim_{n\to\infty} P_{i,j}^{(n)}, \forall i$, so $\sum_{j=1}^{r} w_j = 1$.
Proof of 2) left as exercise.
Proof of 3): Let $\underline{z}$ be a steady state distribution. Then,

$$\underline{z}\Pi = \underline{z} \Rightarrow \underline{z}\Pi^n = \underline{z} \Rightarrow z_j = \sum_{i=1}^{r} z_i P_{i,j}^{(n)}, \forall i.$$

Take the limit of the last statement as $n \to \infty$:

$$z_j = \sum_{i=1}^{r} z_i w_j = w_j.$$

So $\underline{z} = \underline{w}$.

Lemma 2. *If the matrix Π^N does not have at least one column all of whose elements are positive for all values $N \le 2^{r^2}$, then it does not have at least one column all of whose elements are positive for all values of N.*

Example. Let $\Pi = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Then $\Pi^{2n} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\Pi^{2n+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The Markov chain does not have a steady state distribution.

2 Entropy of a Stochastic Process

Definition 8. *The entropy of a stochastic process $X_i, i \in \mathbb{N}$ is defined by:*

$$H_\infty(X) = lim_{n\to\infty} \frac{1}{n} H(X^n),$$

when the limit exists. We also define $H_n(X) = \frac{1}{n}H(X^n)$.

Another quantity of interest for stochastic processes is the entropy rate which is defined below:

Definition 9. *For a stochastic process $X_i, i \in \mathbb{N}$, the entropy rate is defined as:*

$$H'_\infty(X) = \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}),$$

when the limit exists. We also define $H'_n(X) = H(X_n|X_1, X_2, \cdots, X_{n-1})$.

Roughly, the entropy rates gives the expected number of bits of information necessary to store X_n if all of the previous symbols X^{n-1} are already known.

Example. Assume that X^n is an i.i.d sequence. Then $H_\infty(X) = H'_\infty(X) = H(X)$.
Example. Assume that the sequence $X_i, i \in \mathbb{N}$ is a sequence of variables which are all equal to each other. That is $X_1 = X_2 = \cdots$. Then, $H_\infty = H'_\infty(X) = 0$.

Lemma 3. *For a stationary process X^n the function $H'_n(X)$ is non-increasing.*

For a stationary process, $P_{X_1,X_2,\cdots,X_{n-1}} = P_{X_2,X_3,\cdots,X_n} \forall n$. So, $H'_n(X) = H(X_n|X_1, X_2, \cdots, X_{n-1}) \leq H(X_n|X_2, X_3, \cdots, X_{n-1}) = H(X_{n-1}|X_1, X_2, \cdots, X_{n-2}) = H'_{n-1}(X)$.

Theorem 1. *Assume that the stochastic process $X_i, i \in \mathbb{N}$ is a stationary process. Then, the entropy rate of the process exists.*

proof. Remember from calculus that for any sequence of number $a_n, n \in \mathbb{N}$, if the sequence if non-increasing and bounded from below, then $\lim_{n\to\infty} a_n$ exists.

Lemma 4. *For a stationary process X^n we have $H_n(X) \geq H'_n(X)$.*

proof. $H_n(X) = \frac{1}{n}H(X^n) = \frac{1}{n}\sum_{i=1}^n H(X_i|X_1, X_2, \cdots, X_{i-1}) \geq \frac{1}{n}\sum_{i=1}^n H(X_n|X_1, X_2, \cdots, X_{n-1}) = H(X_n|X_1, X_2, \cdots, X_{n-1}) = H'_n(X)$.

Lemma 5. *For a stationary process X^n the function $H_n(X)$ is non-increasing.*

proof. $H_n(X) = \frac{1}{n}H(X^n) = \frac{1}{n}(H(X^{n-1}) + H(X_n|X_1, X_2, \cdots, X_{n-1})) = \frac{1}{n}(H_{n-1}(X) + H'_n(X)) \leq \frac{1}{n}(H_{n-1}(X) + H_n(X))$. Simplifying the two sides gives $H_n(X) \leq H_{n-1}(X)$.

Theorem 2. *Assume that the stochastic process $X_i, i \in \mathbb{N}$ is a stationary process. Then, the entropy of the process exists.*

Theorem 3. *Assume that the stochastic process $X_i, i \in \mathbb{N}$ is a stationary process. Then, the entropy of the process is equal to the entropy rate of the process.*

proof. We use without proof the following result which is called the Cesaro mean result: If $a_n \to a$ as $n \to \infty$, then $\frac{1}{n}\sum_{i=1}^n a_i \to a$ as $n \to \infty$. Intuitively, the Cesaro mean result holds because if $a_n \to a$, then most of the terms in the summation $\sum_{i=1}^n a_i$ are close to a for large enough n. So, the average is also close to a. For a formal proof see Theorem 4.2.3. in the textbook.

Note that $\frac{1}{n}H(X^n) = \frac{1}{n}\sum_{i=1}^n H(X_i|X_1, X_2, \cdots, X_{i-1})$. Let $a_i = H(X_i|X_1, X_2, \cdots, X_{i-1})$. From the previous theorem we know that $\lim_{i\to\infty} a_i = H'_\infty(X)$ exists. On the other hand, from the Cesaro mean theorem and the chain rule of entropy we get that $H_\infty(X) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n a_i = H'_\infty(X)$.

Theorem 4. *Let $X_i, i \in \mathbb{N}$ be a first order Markov Chain. Then, $H_\infty(X) = H'_\infty(X) = H(X_2|X_1)$.*

proof. For exercise, we calculate the entropy and entropy rate separately.

$$\begin{aligned}
H_\infty(X) &= \lim_{n\to\infty} \frac{1}{n} H(X^n) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} H(X_i | X_1, X_2, \cdots, X_{i-1}) \\
&= \lim_{n\to\infty} \frac{1}{n} (H(X_1) + H(X_2|X_1) + H(X_3|X_2) + \cdots + H(X_n|X_{n-1})) \\
&= \lim_{n\to\infty} \frac{1}{n} H(X_1) + \frac{n-1}{n} H(X_2|X_1) = H(X_2|X_1).
\end{aligned}$$

Also,

$$H'_\infty(X) = \lim_{n\to\infty} H(X_n|X_1, X_2, \cdots, X_{n-1}) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1).$$