1 Introduction

1.1 Today’s Topic
- Markov chains/processes
- Entropy rate of Markov chain

1.2 Motivating Example

Example 1: Let us start by considering the following example. What are the rates of $X$ and $Y$?

2 Stochastic Process

A stochastic process can be viewed as an infinite sequence of random variables, e.g., $X_{-n}$, $X_{-n+1}$, ..., $X_0$, $X_1$, $X_2$, ..., $X_n$, ..., whose distribution may be expressed by

$$\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \sim p(x_1, \ldots, x_n).$$

There are some meaningful and restricted classes of stochastic process.
Definition 1 (Stationary Process) \((X_n)_n\) is a stationary process if
\[
\Pr[X_1 = x_1, \ldots, X_n = x_n] = \Pr[X_{n+l} = x_1, \ldots, X_{n+l} = x_n], \quad \forall n, l, x_1, \ldots, x_n.
\]

Definition 2 (Markov Process/Markov Chain) \((X_n)_n\) is a Markov chain if
\[
\Pr[X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}] = \Pr[X_n = x_n | X_{n-1} = x_{n-1}], \quad \forall n, x_1, \ldots, x_n.
\]
If \(X_i \in \Omega\) and \(\Omega\) is finite, then \(\Pr[X_n = x_n | X_{n-1} = x_{n-1}]\) is just \(|\Omega|^2\) entries for every \(n\). But, can we describe it in finite terms? No.

Definition 3 (Time Invariant Markov Chain) Markov Chain is time-invariant if
\[
\Pr[X_n = a | X_{n-1} = b] = \Pr[X_{n+l} = a | X_{n+l-1} = b], \quad \forall n, l, a, b \in \Omega.
\]
Time invariant Markov chain can be specified by distribution on \(X_0\) and probability transition matrix \(P = [P_{ij}]\), where \(P_{ij} = \Pr[X_2 = j | X_1 = i]\). Throughout the rest of lecture, time invariant Markov chain will be referred to simply as Markov chain (MC).

Example 2: Consider the following three-state MC. In this case, \(P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}\).

With \(X_0 = A\), the resulting sequence will be “ABCABCABC⋯”. Note that this is not stationary because \(\Pr[X_0 = A, X_1 = B, X_2 = C] = 1\) but \(\Pr[X_1 = A, X_2 = B, X_3 = C] = 0\). Instead, \(\Pr[X_1 = B, X_2 = C, X_3 = A] = 1\).

Fact 1 For every MC, \(\exists\) stationary distribution \(\mu\) on \(X_0\) such that \(\mu\) and \(P\) define a stationary process. In the example 2, \(\mu = \left[\frac{1}{3} \frac{1}{3} \frac{1}{3}\right]\).

Because
\[
\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]
= \Pr[X_1 = x_1] \cdot \Pr[X_2 = x_2 | X_1 = x_1] \cdots \Pr[X_n = x_n | X_{n-1} = x_{n-1}]
= \Pr[X_1 = x_1] \cdot P_{x_1 x_2} \cdots P_{x_{n-1} x_n},
\]
the overall distribution depends only on the distribution on \(X_1\), which implies that the distribution \(\mu\) on \(X_0\) is stationary if \(\Pr[X_1 = i] = \mu_i (= \Pr[X_0 = i])\).
Example 3: Let us consider the following example:

In this case, \( \mu_A = \mu_C = 0, \mu_B = 1 \) is stationary, but \( \mu_A = \mu_B = 0, \mu_C = 1 \) is also stationary. More than one stationary distribution can be problematic, and this situation happens because the MC is reducible.

Definition 4 (Reducibility of Markov Chain)

1. Markov chain given by probability transition matrix \( P \) is reducible if \( P \) can be written as

\[
\begin{bmatrix}
  P_0 & P_1 \\
  0   & P_2
\end{bmatrix},
\]

where \( P_0, P_2 \) are square matrices.

2. MC is irreducible if it is not reducible.

In terms of graph structure, the “irreducible” and “aperiodic” characteristics can be interpreted as

- irreducible - strongly connected, \( \exists \) path from each state \( i \) to state \( j \).
- aperiodic - greatest common divisor of cycle lengths is 1.

Theorem 2 (Perron-Frobenius’s Theorem) Every (aperiodic) irreducible Markov chain has a unique stationary distribution.

For stationary distribution, the probability distribution on \( X_1 \) should be the same as \( \mu \), the probability distribution of \( X_0 \). \( \Rightarrow \) \( \Pr[X_1 = j] = \sum_{i=1}^{N} \mu_i P_{ij} = \mu_j \), where \( N = |\Omega| \) and \( \Omega = \{1, 2, \cdots, N\} \). If we use vector-matrix notation,

\[
[ \mu ] \begin{bmatrix}
  P \\
\end{bmatrix} = [ \mu ],
\]

and \( \mu \) corresponds to an eigenvector. For the example 1,

\[
P = \begin{bmatrix}
  0.9 & 0.1 & 0 \\
  0 & 2/3 & 1/3 \\
  2/3 & 1/3 & 0
\end{bmatrix}.
\]

Theorem 2 implies that there exists a unique eigenvector with all entries non-negative. We can compute \( \mu = [\mu_1 \mu_2 \mu_3] \) using (1) and \( \mu_1 + \mu_2 + \mu_3 = 1 \). \( \Rightarrow \mu = [\frac{20}{32} \frac{9}{32} \frac{3}{32}] \).
3 Entropy Rate of Stochastic Process

There are two reasonable notions for measuring the uncertainty of \( X = \langle X_n \rangle_n \).

- Entropy rate:
  \[
  H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n)
  \]
  if the limit exists.

- Entropy’ rate:
  \[
  H'(X) = \lim_{n \to \infty} H(X_n | X_1, \ldots, X_{n-1})
  \]
  if the limit exists.

**Theorem 3** Entropy rate of a stationary stochastic process exists and equals entropy’ rate.

\[
H(X) = H'(X).
\]

**Proof Idea** The following inequality can be used for the proof of the existence of \( H'(X) \).

\[
H(X_n | X_1, \ldots, X_{n-1}) \leq H(X_n | X_2, \ldots, X_{n-1}) = H(X_{n-1} | X_1, \ldots, X_{n-1}).
\]

For complete proof, refer to pp.64-65 of Cover.

**Theorem 4** If irreducible MC has probability transition matrix \( P \) and stationary distribution \( \mu \),

\[
H(X) = H'(X) = - \sum_{i,j} \mu_i P_{ij} \log P_{ij}.
\]

Proof

\[
H'(X) = \lim_{n \to \infty} H(X_n | X_1, \ldots, X_{n-1})
\]

\[
= \lim_{n \to \infty} H(X_n | X_{n-1})
\]

\[
= H(X_2 | X_1)
\]

\[
= \sum_i \Pr[X_1 = i] \cdot H(X_2 | X_1 = i)
\]

\[
= - \sum_i \mu_i \sum_j P_{ij} \log P_{ij}.
\]

Using (2), \( H(X) \) of the example 1 can be computed:

\[
H(X) = \frac{5}{8} H(0.9) + \frac{3}{8} H\left(\frac{2}{3}\right).
\]

**AEP for Markov Chain:**

\[
- \frac{1}{n} \log p(X_1, \ldots, X_n) \to H(X).
\]

This doesn’t follow from our law of large numbers because random variables may be dependent on each other.

**Hidden Markov Model:** Now, let us consider the rate of \( \langle Y_n \rangle_n \) in the example 1. \( H'(\mathcal{Y}) = \lim_{n \to \infty} H(Y_n | Y_1, \ldots, Y_{n-1}) \), and is bounded by

\[
H(Y_n | Y_1, \ldots, Y_{n-1}, X_1) \leq H'(\mathcal{Y}) = \lim_{n \to \infty} H(Y_n | Y_1, \ldots, Y_{n-1}) \leq H(Y_n | Y_1, \ldots, Y_{n-1}) \ \forall n.
\]

5-4
(Try to prove the inequality at the left-hand side!) If we denote the interval between the upper and the lower bounds by $\epsilon_n$,

$$
\epsilon_n = H(Y_n|Y_1, \ldots, Y_{n-1}) - H(Y_n|Y_1, \ldots, Y_{n-1}, X_1) = I(X_1; Y_n|Y_1, \ldots, Y_{n-1}),
$$

and

$$
\sum_{n=1}^{M} \epsilon_n = \sum_{n=1}^{M} I(X_1; Y_n|Y_1, \ldots, Y_{n-1}) \leq H(X_1).
$$