Multi-User Information Theory October 26th, 2009

Lecture 2

Lecturer: Haim Permuter *Scribe: Yaniv Nissenboim*

I. MARKOV CHAINS

A. Markov Process

Definition 1 A discrete stochastic process $\{X_i\}_{i\geq 1}$ is said to be *Markov Process* if

$$P(x_{n+1}|x^n) = P(x_{n+1}|x_n), \quad \forall n \tag{1}$$

In this case, the joint probability mass function of the random variables can be written as

$$P(x_1, x_2, ..., x_n) = P(x_1)P(x_2|x_1)P(x_3|x_2)\cdots P(x_n|x_{n-1}) \tag{2}$$

Definition 2 The Markov Process is said to be *time invariant* if the conditional probability $P(x_{n+1}|x_n)$ does not depend on n, that is

$$P(x_{n+1} = i|x^n = j) = P(x_2 = i|x_1 = j) = p_{ij}, \quad \forall n \ \ and \ \ \forall i, j \in \mathcal{X}. \tag{3}$$

B. Markov Chain

Definition 3 A *Markov Chain* is finite Markov Process. If $\{X_i\}$ is a Markov Chain, X_n is called state at time *n*. A time invariant(stationary) Markov Chain is characterized by a *transition matrix*,

$$\Pi = P_{i,j}, \quad i, j \in \{1, 2, ..., m\}. \tag{4}$$

The initial state probability is $P_0(i) = \Pr(x_0 = 1)$.

Let $P_t = [\Pr(x_t = 1), P_r(x_t = 2), ..., \Pr(x_t = m)]$ be a probability vector, and Π the transition matrix of the stationary Markov Chain, thus we can write:

$$P_t = P_{t-1} \cdot \Pi \tag{5}$$

$$P_t = P_0 \cdot \Pi^t. \tag{6}$$

Equation 5 follows from the following equations:

$$P(x_t = j) = \sum_{i=1}^{m} P(x_t = j, x_{t-1} = i) = \sum_{i=1}^{m} P(x_{t-1} = i)P(x_t = j|x_{t-1} = i) = \sum_{i=1}^{m} P_{t-1}(i)P_{i,j} \tag{7}$$

Properties of Markov Chain:

1) *Irreducible* There exists a positive probability of getting to any state from any state. That is, all the states are connected.

2) *Aperiodic* The largest common factor (GCD) of all possible loops in state is 1, i.e. returns to state i can occur at irregular times.

These properties can be demonstrated by the following figure:

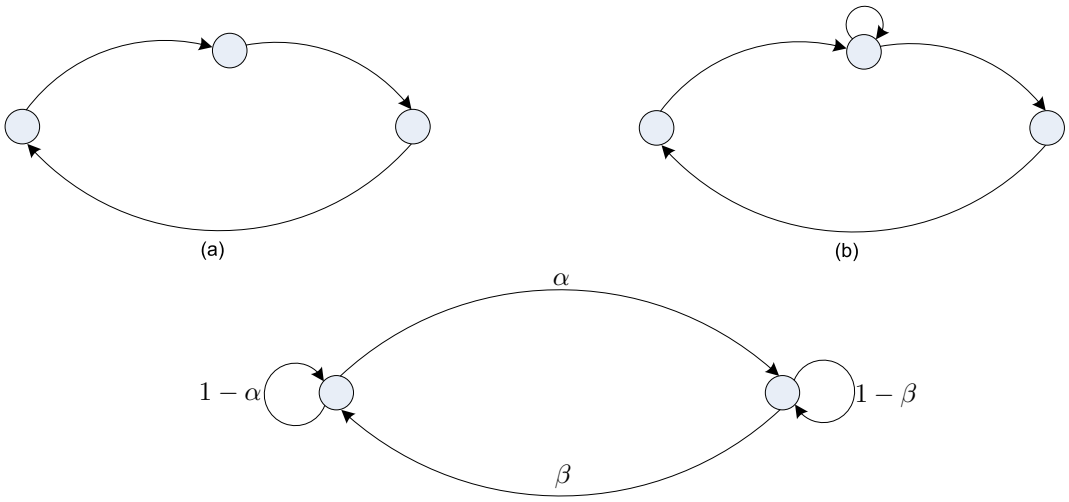

Fig. 1. a. Irreducible Non-Aperiodic, b. Irreducible Aperiodic, c. Irreducible Non-Aperiodic

Only when the Markov Chain is irreducible, the aperiodic property is defined. Also, when the Markov Chain is irreducible, then if of the states is aperiodic, then all the states are aperiodic.

Definition 4 μ is *stationary distribution* if exists μ such that $\mu\Pi = \mu$. That is for each initial state we start from, we will get μ after finite number of steps.

Theorem 1 (Sufficient condition for existance of a stationary distribution) If a finite-state Markov Chain is aperiodic and irreducible there exists an unique stationary distribution. That is $P_{X_n} = P_{X_{n+1}}$.

Example 1 ([1], Chapter 4.1) Consider a two-state Markov chain that is irreducible and aperiodic with a probability transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix},$$

as shown in the Figure 1, (c).

Let the stationary distribution be represented by the vector $\mu = [\mu_1 \quad \mu_2]$, such that each component of the vector is the stationary probability of states 1 and 2. We can find the stationary probability by solving $\mu P = \mu$. From the fact that $\mu_1 + \mu_2 = 1$ we get:

$$\mu_1 = \frac{\beta}{\alpha + \beta}, \tag{8}$$

$$\mu_1 = \frac{\alpha}{\alpha + \beta}. \tag{9}$$

Recall that the *entropy rate* of a stationary process is,

$$H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X^{n-1}). \tag{10}$$

Now we compute $H(X_n|X^{n-1})$ for stationary Markov process and by so we will get the entropy rate of stationary Markov process,

$$\begin{aligned} H(X_n|X^{n-1}) &\overset{(a)}{=} H(X_n|X_{n-1}) \\ &\overset{(b)}{=} H(X_1|X_0) \\ &= \sum_{i=1}^{m} P(i) H(X_1|X_0 = i) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{m} P(i) P_{ij} \log(P_{ij}) \end{aligned} \tag{11}$$

where

(a) follows from Markovity.

(b) follows from stationary process.

The entropy rate of example 1 is,

$$H(X) = \frac{\beta}{\alpha + \beta} H_b(\alpha) + \frac{\alpha}{\alpha + \beta} H_b(\beta), \tag{12}$$

where $H_b(p) = -p \log p - (1-p) \log(1-p)$.

C. HMM - Hidden Markov Model [1], Chapter 4.5

Let us consider a Markov process $X_1, X_2, ..., X_n$, and define a new process $Y_1, Y_2, ..., Y_n$, where each Y_i is drawn according to $P(y_i|x_i)$, conditionally independent of all the other $X_j, j \neq i$; that is,

$$P(x^n, y^n) = P(x_1) \Pi_{i=1}^{n-1} P(x_{i+1}|x_i) \Pi_{i=1}^{n} P(y_i|x_i). \tag{13}$$

Such a process, called a *Hidden Markov model (HMM)*, is used extensively in speech recognition, handwriting recognition, and so on. The same argument as that used above for functions of a Markov chain carry over to hidden Markov models, and we can lower bound the entropy rate of a hidden Markov

model by conditioning it on the underlying Markov state.

In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output dependent on the state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by a HMM gives some information about the sequence of states.

Another way to represent the HMM is by $Y_i = \phi(X_i)$, whereas Y_i is deterministic function of the state, X_i.

This two HMM representations, i.e., $P(y_i|x_i)$ and $Y_i = \phi(X_i)$, are equivalent. If we consider X_i as the state of the Markov chain, we can define $P(y_i|x_i)$ by

$$P(y_i|x_i) = \left\{ \begin{array}{ll} \phi(x_i), & w.p.\ \ 1 \\ others, & w.p.\ \ 0 \end{array} \right\},$$

by doing so we represent $y_i = \phi(x_i)$ by $P(y_i|x_i)$. Now, in order to represent $P(y_i|x_i)$ by $y_i = \phi(x_i)$ we can construct new states $\tilde{x_i} = (x_i, y_i)$, which is Markovian and clearly $y_i = \phi(\tilde{x}_i)$. We need to check the Markov property of $\tilde{x}_i$, i.e. $\tilde{x}_i - \tilde{x}_{i-1} - \tilde{x}^{i-1}$,

$$\begin{aligned} P(\tilde{x_i}|\tilde{x^{i-1}}) &= P(x_i, y_i|x^{i-1}, y^{i-1}) \\ &= P(x_i|x^{i-1}, y^{i-1})P(y_i|x^i, y^{i-1}) \\ &= P(x_i|x_{i-1})P(y_i|x_i) \\ &= P(\tilde{x_i}|\tilde{x_{i-1}}). \end{aligned} \tag{14}$$

Note that the underlying states of the Markov chain cannot be observed. They are said to be hidden.

II. Gambling

In this section we will show that there is strong duality between the growth rate of investment in a horse race and the entropy rate of the horse race.

A. The horse race

In order to describe the horse race let us define the following:

- m - Number of horses.
- X_i - Random variable that tells us which horse wins at time i. $X_i = \{1, 2, ..., m\}$.
- P_X - The pmf of the winning horse.
- $o(X)$ - The amount of money we get, for each dollar we put, if horse X wins.
- $b(X)$ - Betting strategy on horse X. $b(X) \geq 0 \ \ \forall X$ and $\sum_x b(X) = 1$.

- S_n - Money, after n rounds of gambling.

The following are elaborations on S_n for different n:

$*S_0 = 1$ - Money the gambler has in time 0.

$*S_1 = b(X_1) \cdot o(X_1) \cdot S_0$ - Money the gambler has in time 1.

$*S_2 = b(X_2) \cdot o(X_2) \cdot S_1$ - Money the gambler has in time 2.

$*S_n = b(X_n) \cdot o(X_n) \cdot S_{n-1} = \Pi_{i=1}^{n} b(X_i) \cdot o(X_i)$ - Money the gambler has in time n.

In the horse race we assume that:

- The gambler distributes all of his wealth across the horses.
- The winning probability is time invariant.
- The wealth at the end of the race is a random variable.
- The gambler wishes to maximize the value of this random variable.

The objective goal will be to find

$$b = \arg\max_{b(\cdot)} \mathbb{E}[\log S_n] \tag{15}$$

$$\begin{aligned}
\max_{b(x)} \mathbb{E}[\log S_n] &= \max_{b(x)} \mathbb{E}[\log \Pi_{i=1}^{n} b(x_i) + \log \Pi_{i=1}^{n} O(x_i)] \\
&= \max_{b(x)} \sum_{i=1}^{n} (\mathbb{E}[\log b(x_i)] + \mathbb{E}[\log O(x_i)]) \\
&\overset{(a)}{=} \max_{b(x)} \sum_{i=1}^{n} (\mathbb{E}[\log b(x_i)]) \\
&\overset{(b)}{=} n \cdot \arg\max_{b(x)} \mathbb{E}[\log b(x_i)] \\
&= n \cdot \max_{b(x)} \sum_{x} P(x) \log b(x) \\
&= n \cdot \max_{b(x)} \sum_{x} (P(x) \log \frac{b(x)}{P(x)} + P(x) \log P(x)) \\
&\overset{(c)}{=} n \cdot \max_{b(x)} [-D(P(x)||b(x)) - H(X)] \\
&\overset{(d)}{\leq} -n \cdot H(X)
\end{aligned} \tag{16}$$

Thus,

$$b(x) = P(x) \tag{17}$$

where

(a) follows from the fact that $O(x_i)$ does not depend on $b(x_i)$.

(b) follows from the fact that we are maximizing over the same argument.

(c) follows from definition of divergence and entropy.

(d) follows from the fact that D is non-negative.

Equation 17 follows from the fact that (d) in Eq. 16 reach equality when $b(x) = P(x)$ due to the properties of divergence.

B. Gambling with causal side information [2]

Assume there are m racing horses where X_i denotes the horse that wins at time i, i.e., $X_i \in \mathcal{X} := [1, 2, ..., m]$. At time i, the gambler knows some side information which we denote as Y_i. We assume that the gambler invests all his capital in the horse race as a function of the information that he knows at time i, i.e., the previous horse race outcomes X^{i-1} and side information Y^i up to time i. Let $b(x_i|x^{i-1}, y^i)$ be the portion of wealth that the gambler bets on horse x_i given $X^{i-1} = x^{i-1}$ and $Y^i = y^i$. Obviously, the betting scheme should satisfy $b(x_i|x^{i-1}, y^i) \geq 0$ and $\sum_{x_i} b(x_i|x^{i-1}, y^i) = 1$ for any history x^{i-1}, y^i. Let $o(x_i|x^{i-1})$ denote the odds of a horse x_i given the previous outcomes x^{i-1}, which is the amount of capital that the gambler gets for each unit capital that the gambler invested in the horse. We denote by $S(x^n||y^n)$ the gambler's wealth after n races where the race outcomes were x^n and the side information that was causally available was y^n. We assume that the gambler wishes to maximize his wealth which is a random variable.

Here is a summary of the notation:

- X_i is the outcome of the horse race at time i.
- Y_i is the the side information at time i.
- $o(X_i|X^{i-1})$ is the payoffs at time i for horse X_i given that in the previous race the horses X^{i-1} won.
- $b(X_i|Y^i, X^{i-1})$ betting strategy - the fractions of the gambler's wealth invested in horse X_i at time i given that the outcome of the previous races are X^{i-1} and the side information available time i is Y^i.
- $S(X^n||Y^n)$ the gambler's wealth after n races when the outcomes of the races are X^n and the side information Y^n is causally available.

Without loss of generality, we assume that, initially, the gambler's capital is 1; therefore $S_0 = 1$. We assume that at any time n the gambler invests all his capital and therefore we have

$$S(X^n||Y^n) = b(X_n|X^{n-1}, Y^n)o(X_n|X^{n-1})S(X^{n-1}||Y^{n-1}). \tag{18}$$

This also implies that

$$S(X^n||Y^n) = \prod_{i=1}^{n} b(X_i|X^{i-1}, Y^i)o(X_i|X^{i-1}). \tag{19}$$

The objective goal will be to find

$$b = \arg \max_{b(X_i|X^{i-1},Y^i)} \mathbb{E}[\log S(X^n||Y^n)] \tag{20}$$

$$\begin{aligned}
\max_{b(X_i|X^{i-1},Y^i)} \mathbb{E}[\log S(X^n||Y^n)] &\overset{(a)}{=} \max_{b(x^n||y^n)} \mathbb{E}[\log b(x^n||y^n) + \log o(X_i|X^{i-1})] \qquad (21)\\
&\overset{(b)}{=} \max_{b(x^n||y^n)} \mathbb{E}[\log b(x^n||y^n))\\
&= \max_{b(x^n||y^n)} \sum_{x^n,y^n} P(x^n, y^n) \log b(x^n||y^n)\\
&= \max_{b(x^n||y^n)} \sum_{x^n,y^n} P(x^n, y^n) \log[b(x^n||y^n) \frac{P(x^n||y^n)}{P(x^n||y^n)}]\\
&= \max_{b(x^n||y^n)} \sum_{x^n,y^n} P(x^n, y^n) \log P(x^n||y^n) + \sum_{x^n,y^n} P(x^n, y^n) \log \frac{b(x^n||y^n)}{P(x^n||y^n)}\\
&\overset{(c)}{=} \max_{b(x^n||y^n)} -H(X^n||Y^n) + \sum_{x^n,y^n} P(x^n, y^n) \log \frac{b(x^n||y^n)}{P(x^n||y^n)}\\
&\overset{(d)}{\leq} -H(X^n||Y^n) \qquad (22)
\end{aligned}$$

Thus,

$$b(x^n||y^n) = P(x^n||y^n) \tag{23}$$

where

(a) follows from the fact that $b(x_i|x^{i-1}, y^i)$ uniquely determines $b(x^n||y^n)$.

(b) follows from the face that $o(X_i|X^{i-1})$ does not depend on $b(x_i|x^{i-1}, y^i)$.

(c) follows from the fact that $\sum_{x^n,y^n} P(x^n, y^n) \log P(x^n||y^n) == -H(X^n||Y^n)$.

(d) follows from:

$$\begin{aligned}
\sum_{x^n,y^n} P(x^n, y^n) \log \frac{b(x^n||y^n)}{P(x^n||y^n)} &\leq \log[\sum_{x^n,y^n} \frac{P(x^n, y^n)b(x^n||y^n)}{P(x^n||y^n)}]\\
&= \log[\sum_{x^n,y^n} \frac{P(x^n||y^n)P(y^n||x^{n-1})b(x^n||y^n)}{P(x^n||y^n)}]\\
&= \log[\sum_{x^n,y^n} P(y^n||x^{n-1})b(x^n||y^n)]\\
&= \log 1 = 0
\end{aligned}$$

Note that since $\{P(x_i|x^{i-1},y^i)\}_{i=1}^n$ uniquely determines $P(x^n||y^n)$, and since $\{b(x_i|x^{i-1},y^i)\}_{i=1}^n$ uniquely determines $b(x^n||y^n)$, then ($b(x^n||y^n)=P(x^n||y^n)$) is equivalent to

$$b(x_i|x^{i-1},y^i)=P(x_i|x^{i-1},y^{i-1}). \tag{24}$$

and so in order to maximize the gambler wealth the betting strategy will be,

$$b(x_i|x^{i-1},y^i)=P(x_i|x^{i-1},y^{i-1}),\quad \forall i\in[1,...,n],x^i\in\mathcal{X}^i,y^i\in\mathcal{Y}^i \tag{25}$$

Theorem 2 If we wish the evaluate the value of side information in gambling we need to compute the following,

$$\begin{aligned}\mathbb{E}[\log S(X^n||Y^n)]-\mathbb{E}[\log S(X^n)] &= -H(X^n||Y^n)+H(X^n) \\ &= I(Y^n\to X^n)\end{aligned} \tag{26}$$

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New-York: Wiley, 2006.

[2] H. Permuter, Y. H. Kim and T. Weissman, *On Directed Information and Gambling,* ISIT 2008, Toronto, Canada.