Multi-User Information Theory October 19th, 2009

Lecture 1

Lecturer: Haim Permuter *Scribe: Noa Zuaretz*

I. Notation

- X - random variable
- $\mathcal{X}$ - alphabet
- $|\mathcal{X}|$ - cardinality of the alphabet. Unless it is said otherwise, we assume that the alphabet is finite, i.e., $|\mathcal{X}| < \infty$.
- x - an observation or a specific value. Clearly, $x \in \mathcal{X}$.
- $P_X(x)$ - the probability that the random variable X gets the value x, i.e., $P_X(x) = \Pr\{X = x\}$.
- P_X - denotes the whole vector of probabilities. One may also use the notation $P_X(\cdot)$.
- $P(x)$ - this is a short notation for $P_X(x)$.
- x^n - is the vector $(x_1, x_2, ..., x_n)$ for $n \geq 1$. If $n = 0$ then the vector is empty.
- x_i^j - is the vector $(x_i, x_{i+1}, ..., x_j)$, for $j > i$. If $j = i$, then the vector has only one element x_i and if $j < i$, the vector is empty.

II. Entropy Rates of a Stochastic Process

A. Entropy Rate

If we have a sequence of n random variables, how does the entropy of the sequence grow with n?

Definition 1 We define the entropy rate as this rate of growth. The entropy of a stochastic process $\{X_i\}$ is defined by

$$H(\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X_n) \tag{1}$$

if the limits exists.

If X is i.i.d. then $\frac{1}{n}H(X^n) = H(X)$.

We can also define a related quantity for entropy rate by

$$H'(\bar{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X_n|X^{n-1}) \tag{2}$$

if the limits exists.

$H(\bar{X})$ and $H'(\bar{X})$ correspond to two different notions of entropy rate. The first (1) is the per symbol entropy of the n random variables, and the second (2) is the conditional entropy of the last random variable given the past.

Example 1

$$X^n = \begin{cases} 000...0 \ p = \frac{1}{2} \\ 111...1 \ p = \frac{1}{2} \end{cases} \tag{3}$$

$$\frac{1}{n} H(X^n) = \frac{1}{n} \to 0 \text{ when } n \to \infty \tag{4}$$

Definition 2 A process $\{X_i\}_{i \geq 1}$ is stationary if P(X_i^{i+n}) = P(X_0^n) $\forall i, n$

Theorem 1 For any stationary process $\{X_i\}_{i \geq 1}$ the limits in (1) and (2) exist and are equal

$$H(\bar{X}) = H'(\bar{X}) \tag{5}$$

Proof:

$$H(X_{n+1}|X^n) \overset{(a)}{\leq} H(X_{n+1}|X_2^n) \tag{6}$$

$$\overset{(b)}{=} H(X_n|X_1^{n-1}) \tag{7}$$

(a) Conditioning reduces entropy

(b) Stationarity properties ■

The sequence $H(X_n|X^{n-1})$ is decreasing and positive, hence the limit exists.

Lemma 1 (Cesaro mean) Let $\{a_n\}_{n \geq 1}$ be a sequence such that $\lim_{n \to \infty} a_n = a$ then the limit of the sequence $\{b_n\}_{n \geq 1}$ where $b_n = \frac{1}{n} \sum_{i=1}^n a_n$ is $\lim_{n \to \infty} b_n = a$.

Proof: Let $\varepsilon > 0$. Since $a_n \to a$, there exists a number $N(\varepsilon)$ such that $|a_n - a| \leq \varepsilon$ for all $n \geq N(\varepsilon)$. Furthermore,

$$|b_n - a| = |\frac{1}{n} \sum_{i=1}^n (a_i - a)| \tag{8}$$

$$\leq \frac{1}{n} \sum_{i=1}^n |(a_i - a)| \tag{9}$$

$$\leq \frac{1}{n} \sum_{i=1}^{N(\varepsilon)} |(a_i - a)| + \frac{n - N(\varepsilon)}{n} \varepsilon \tag{10}$$

$$\leq \frac{1}{n} \sum_{i=1}^{N(\varepsilon)} |(a_i - a)| + \varepsilon \tag{11}$$

for all $n \geq N(\varepsilon)$. Since the first term goes to 0 as $n \to \infty$, we can make $|b_n - a| \leq 2\varepsilon$ by taking n large enough. Hence, $b_n \to a$ as $n \to \infty$. ■

III. Directed Information and Causal Conditioning

A. Causal Conditioning

Introduced and employed by Kramer [4], [3] and by Massey [2], Causal Conditioning (notation $(\cdot||\cdot)$) is defined as the probability mass function of the sequence X^n causally conditioned on the sequence Y^n.

Definition 3

$$P(\text{  }|Y^n) \triangleq \prod_{i=1}^{n} P(X_i|X^{i-1}, Y^i) \tag{12}$$

In addition, we introduce the following notation:

$$P(X^n||Y^{n-1}) \triangleq \prod_{i=1}^{n} P(X_i|X^{i-1}, Y^{i-1}) \tag{13}$$

The definition given in (13) can be considered to be a particular case of the definition given in (12) where X_0 is set to a dummy zero. This concept was captured by a notation of Massey in [2] via a concatenation at the beginning of the sequence X^{i-1} with a dummy zero.

Since, we define the causal conditioning $P(X^n||Y^n)$ as a product of $P(X_i|X^{i-1}, Y^i)$, then whenever we use $P(X^n||Y^n)$, we implicitly assume that there exists a set of that satisfies the equality in (12). We can call $P(X^n||Y^n)$ and $P(X^n||Y^{n-1})$ a causal conditional distribution since they are nonnegative for all X^n, Y^n and since they sum to one.

Lemma 2

$$\sum_{X^n} P(X^n||Y^n) = 1 \tag{14}$$

$$\sum_{X^n} P(X^n||Y^{n-1}) = 1 \tag{15}$$

Proof:

$$\begin{aligned}
\sum_{X^n} P(X^n||Y^n) &= \sum_{X^n} \sum_{X^{n-1}} P(X^{n-1}||Y^{n-1}) P(X_n|X^{n-1}, Y^n) && (16)\\
&= \sum_{X^n} \sum_{X^{n-1}} P(X^{n-1}||Y^{n-1}) P(X_n|X^{n-1}, Y^n) && (17)\\
&= \sum_{X^{n-1}} \{P(X^{n-1}||Y^{n-1})(\sum_{X^n} P(X_n|X^{n-1}, Y^n))\} && (18)\\
&= \sum_{X^{n-1}} P(X^{n-1}||Y^{n-1}) && (19)
\end{aligned}$$

■

Lemma 3 Chain Rule for Causal Conditioning. Using the chain rule, we can easily verify that $P(X^n, Y^n) = P(X^n||Y^n)P(X^n||Y^{n-1})$.

Proof:

$$\begin{aligned} P(X^n||Y^{n-1})P(Y^n||X^n) &= P(Y^n, X^n) && (20)\\ &= \prod_{i=1}^{n} P(X_i|X^{i-1}, Y^{i-1})P(Y_i|Y^{i-1}, X^{i-1}, X_i) && (21)\\ &= \prod_{i=1}^{n} P(X_i, Y_i|X^{i-1}, Y^{i-1}) && (22)\\ &= P(X^n, Y^n) && (23) \end{aligned}$$

■

Lemma 4 The causal conditioning distribution $P(X^n||Y^n)$ uniquely determines the value of $P(X_i|X^{i-1}, Y^{i-1})$ for all $i \leq N$ and all the arguments (X^{i-1}, Y^{i-1}), for which $P(X^{i-1}, Y^{i-1}) > 0$.

Proof: First we note that if $P(X^{i-1}, Y^{-1}) > 0$, then according to Lemma 3, it also implies that $P(X^{i-1}||Y^{i-2}) > 0$. In addition, we always have equality

$$P(X^{i-1}||Y^{i-2}) = \sum_{X^n} P(X^n||Y^{n-1}) \quad (24)$$

hence, $P(X^n||Y^{n-2})$ is uniquely determined from $P(X^n||Y^{n-1})$. Furthermore, by induction it can be shown that the sequence $P(X^i||Y^{i-1})_{i=1}^{n}$ is uniquely derived from $P(X^i||Y^{n-1})$. Since $P(X^i||Y^{i-2}) > 0$, we can use the equality

$$P(X_i|X^{i-1}, Y^{i-1}) = \frac{P(X_i||Y^{i-1})}{P(X^{i-1}||Y^{i-2})} \quad (25)$$

■

B. Causal Conditioning Entropy

Definition 4 The entropy of the sequence X^n causally conditioned on the sequence Y^n is

$$\begin{aligned} H(X^n||Y^n) &\triangleq E[-\log P(X^n||Y^n)] && (26)\\ &= -\sum_{X^n, Y^n} P(X^i, Y^i) \log P(X^i||Y^i) && (27) \end{aligned}$$

C. Directed Information

The directed information flowing from a sequence X^n to a sequence Y^n was introduced by Massey [1] and can be written as

Definition 5

$$I(X^n \to Y^n) = H(Y^n) - H(Y^n||X^n) \quad (28)$$

$$= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i) \quad (29)$$

$$= \sum_{i=1}^{n} I(Y_i, X^i|Y^{i-1}) \quad (30)$$

which hints, in a rough analogy to mutual information, a possible interpretation of directed information $I(X^n \to Y^n)$ as the amount of information causally available side information Y^n can provide about X^n.

D. Memoryless Channel

A discrete channel with finite input alphabet x and finite output alphabet y is the specification of the conditional probability $P(y_n|x^n, y^{n-1})$ for all $n \geq 1$, all $X^n \epsilon x^n$ and all $Y^n \epsilon y^n$. The discrete channel is memoryless if this conditional probability satisfies

Definition 6

$$P(Y_n|X^n, Y^{n-1}) = P(Y^n|X^n) \quad (31)$$

Lemma 5 For the memoryless channel $\max_{P_{X^n||Y^{n-1}}} \frac{1}{n} I(X^n \to Y^n) = \max_{P_X} I(X;Y)$, the capacity is defined as the supremum over all achievable rates [7].

$$C = \max\{I(X^n \to Y^n)\} \quad (32)$$

$$= \max\{H(Y^n) - H(Y^n||X^n)\} \quad (33)$$

$$= \max\{\sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}, X^i)\} \quad (34)$$

$$\overset{(a)}{\leq} max\{\sum_{i=1}^{n} H(Y_i) - H(Y_i|X^i)\} \quad (35)$$

$$= \max\{\sum_{i=1}^{n} I(Y_i, X_i)\} \quad (36)$$

(a) Conditioning reduces entropy

Note: max $\{\frac{1}{n} I(X^n \to Y^n)\}$ = max $\{I(X;Y)\}$

E. Conservation law

Massey [2] showed the following Conservation Law

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n) \quad (37)$$

Proof:

$$I(X^n \to Y^n) = H(Y^n) - H(Y^n||X^n) \quad (38)$$

$$= E\left(log \frac{P(Y^n||X^n)}{P(Y^n)}\right) \quad (39)$$

$$I(Y^{n-1} \to X^n) \stackrel{(a)}{=} I(\emptyset Y^{n-1} \to X^n) \tag{40}$$

$$= H(X^n) - H(X^n||Y^{n-1}) \tag{41}$$

$$= E\left(\log \frac{P(X^n||Y^{n-1})}{P(X^n)}\right) \tag{42}$$

(a) Øis added in order to achieve same length on both sides $length\emptyset Y^{n-1} = lengthX^n$.

$$I(X^n \to Y^n) + I(Y^{n-1} \to X^n) = E\left(\log \frac{P(Y^n||X^n)}{P(Y^n)} \frac{P(X^n||Y^{n-1})}{P(X^n)}\right) \tag{43}$$

$$= E\left(\log \frac{P(X^n, Y^n)}{P(Y^n)P(X^n)}\right) \tag{44}$$

$$= I(X^n \to Y^n) \tag{45}$$

■

REFERENCES

[1] J. Massey, *Causality, feedback and directed information,* in Proc. Int. Symp. Information Theory and Its Applications (ISITA-90), Waikiki, HI, Nov. 1990, pp. 303305.

[2] J. Massey, *Conservation of mutual and directed information,* in Proc. Int. Symp. Information Theory (ISIT-05), Adelaide, Australia, Sep. 2005, pp. 157158.

[3] G. Kramer, *Capacity results for the discrete memoryless network,* IEEE Trans. Inf. Theory, vol. 49, no. 1, pp. 421, Jan. 2003.

[4] G. Kramer, *Directed information for channels with feedback,* Ph.D. Dissertation, Swiss Federal Institute of Technology (ETH) Zurich, 1998.

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

[6] H. Permuter, T. Weissman and A. Goldsmith, *Finite state channels with time-invariant deterministic feedback,* IEEE Trans. Info. Theory, Feb 2009.

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