

Lecture 1

*Lecturer: Haim Permuter**Scribe: Noa Zuretz*

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, \dots, x_n) for $n \geq 1$. If $n = 0$ then the vector is empty.
- x_i^j - is the vector $(x_i, x_{i+1}, \dots, x_j)$, 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 ? The following definition answer this question.

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

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

if the limit exists.

Exercise 1 Show that 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

$$H'(\mathcal{X}) \triangleq \lim_{n \rightarrow \infty} H(X_n | X^{n-1}) \quad (2)$$

if the limit exists.

$H(\mathcal{X})$ and $H'(\mathcal{X})$ 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 Find the entropy rate of the following process which has memory:

$$X^n = \begin{cases} 000\dots 0 & p = \frac{1}{2} \\ 111\dots 1 & p = \frac{1}{2}. \end{cases} \quad (3)$$

Answer:

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

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 (Entropy rate of stationary processes) For any stationary process $\{X_i\}_{i \geq 1}$ the limits in (1) and (2) exist and are equal, i.e.,

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

Proof:

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

$$\stackrel{(b)}{=} H(X_n|X_1^{n-1}), \quad (7)$$

where

(a) follows from the fact that conditioning reduces entropy

(b) follows from the stationarity properties

The sequence $H(X_n|X^{n-1})$ is decreasing and positive, hence the limit exists. Finally we conclude that $\lim \frac{1}{n}H(X^n) = \lim H(X_n|X^{n-1})$, using Cesàro mean lemma which is proved below. ■

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

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

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

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

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

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

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

III. DEFINITIONS AND PROPERTIES OF DIRECTED INFORMATION AND CAUSAL CONDITIONING

A. Causal Conditioning

Introduced and employed by Kramer [4], [3] and by Massey [1], *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

$$P(x^n||y^n) \triangleq \prod_{i=1}^n P(x_i|x^{i-1}, y^i) \quad (12)$$

In addition, we introduce the following notation:

$$P(x^n||y^{n-1}) \triangleq \prod_{i=1}^n P(x_i|x^{i-1}, y^{i-1}) \quad (13)$$

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

$$\sum_{x^n} P(x^n||y^n) = 1 \quad (14)$$

$$\sum_{x^n} P(x^n||y^{n-1}) = 1 \quad (15)$$

Proof:

$$\sum_{x^n} P(x^n||y^n) = \sum_{x_n} \sum_{x^{n-1}} P(x^{n-1}||y^{n-1}) P(x_n|x^{n-1}, y^n) \quad (16)$$

$$= \sum_{x_n} \sum_{x^{n-1}} P(x^{n-1}||y^{n-1}) P(x_n|x^{n-1}, y^n) \quad (17)$$

$$= \sum_{x^{n-1}} \left\{ P(x^{n-1}||y^{n-1}) \left(\sum_{x_n} P(x_n|x^{n-1}, y^n) \right) \right\}$$

(18)

$$= \sum_{x^{n-1}} P(x^{n-1}||y^{n-1}). \quad (19)$$

Now, by iterating equation (16) n times we obtain (14). Similarly, we obtain (15) ■

Lemma 3 (Chain Rule for Causal Conditioning.) Any joint distribution can be decompose as

$$P(x^n, y^n) = P(x^n||y^n)P(x^n||y^{n-1}), \quad (20)$$

and any product $P(x^n||y^n)P(x^n||y^{n-1})$ results in a joint distribution.

Proof:

$$P(x^n||y^{n-1})P(y^n||x^n) = P(y^n, x^n) \quad (21)$$

$$= \prod_{i=1}^n P(x_i|x^{i-1}, y^{i-1})P(y_i|y^{i-1}, x^{i-1}, x_i) \quad (22)$$

$$= \prod_{i=1}^n P(x_i, Y_i|x^{i-1}, y^{i-1}) \quad (23)$$

$$= P(x^n, y^n) \quad (24)$$

■

Lemma 4 (Equivalence of $P(x^n||y^n)$ and $\{P(x_i|x^{i-1}, y^{i-1})\}$.) 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 $P(x^{i-1}, y^{i-1}) > 0$, then according to Lemma 3, it also implies that $P(x^{i-1}||y^{i-2}) > 0$. In addition, we always have equality

$$P(x^{n-1}||y^{n-2}) = \sum_{x_n} P(x^n||y^{n-1}) \quad (25)$$

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 $P(x^i||y^{i-1})$. Since $P(x^i||y^{i-2}) > 0$, we can use the equality

$$P(x_i|x^{i-1}, y^{i-1}) = \frac{P(x_i||y^{i-1})}{P(x^{i-1}||y^{i-2})} \quad (26)$$

■

B. Causal Conditioning Entropy

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

$$H(X^n||Y^n) \triangleq E[-\log P(X^n||Y^n)] \quad (27)$$

$$= - \sum_{x^n, y^n} P(x^n, y^n) \log P(x^n||y^n) \quad (28)$$

C. Directed Information

Definition 5 The directed information flowing from a sequence X^n to a sequence Y^n was introduced by Massey [1]

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

Note that

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

$$= \sum_{i=1}^n H(Y_i | Y^{i-1}) - H(Y_i | Y^{i-1}, X^i) \quad (31)$$

$$= H(Y^n) - H(Y^n | X^n), \quad (32)$$

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

Example 2 (Memoryless Channel.) A discrete channel with finite input alphabet \mathcal{X} and finite output alphabet \mathcal{Y} is the specification of the conditional probability $P(y_n | x^n, y^{n-1})$ for all $n \geq 1$, all $x^n \in \mathcal{X}^n$ and all $y^n \in \mathcal{Y}^n$.

Recall the definition of a memoryless channel.

Definition 6 The discrete channel is *memoryless* if

$$P(y_n | x^n, y^{n-1}) = P(y_n | x_n) \quad \forall n, x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n, \quad (33)$$

or equivalently

$$P(y^n | x^n) = \prod_{i=1}^n P(y_i | x_i) \quad \forall n, x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n, \quad (34)$$

Show that for the memoryless channel $\max_{P_{X^n | Y^{n-1}}} \frac{1}{n} I(X^n \rightarrow Y^n) = \max_{P_X} I(X; Y)$. (Later in the course we will see that for channels with memories and feedback $\max_{P_{X^n | Y^{n-1}}} \frac{1}{n} I(X^n \rightarrow Y^n)$ characterize the capacity.

Answer

$$\max_{P_{X^n | Y^{n-1}}} I(X^n \rightarrow Y^n) = \max_{P_{X^n | Y^{n-1}}} H(Y^n) - H(Y^n | X^n) \quad (35)$$

$$= \max_{P_{X^n | Y^{n-1}}} \sum_{i=1}^n H(Y_i | Y^{i-1}) - H(Y_i | Y^{i-1}, X^i) \quad (36)$$

$$\stackrel{(a)}{\leq} \max_{P_{X^n | Y^{n-1}}} \sum_{i=1}^n H(Y_i) - H(Y_i | X^i) \quad (37)$$

$$= \max_{P_{X_i}} \sum_{i=1}^n I(Y_i, X_i), \quad (38)$$

$$= n \max_{P_X} I(Y, X), \quad (39)$$

where,

(a) follows from the memoryless property and because conditioning reduces entropy

Now, note that if we choose $P_{X_n|X^{n-1}, Y^{n-1}} = P_X$ then inequality (a) becomes equality, hence $\max_{P_{X^n|Y^{n-1}}} \frac{1}{n} I(X^n \rightarrow Y^n) = n \max_{P_X} I(X; Y)$

D. Conservation law

Lemma 5 (Conservation law of directed information[1]) The following conservation law holds for any random vectors X^n, Y^n

$$I(X^n; Y^n) = I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n) \quad (40)$$

Proof:

$$I(X^n \rightarrow Y^n) = H(Y^n) - H(Y^n || X^n) \quad (41)$$

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

$$I(Y^{n-1} \rightarrow X^n) \stackrel{(a)}{=} I(\emptyset Y^{n-1} \rightarrow X^n) \quad (43)$$

$$= H(X^n) - H(X^n || Y^{n-1}) \quad (44)$$

$$= E \left(\log \frac{P(X^n || Y^{n-1})}{P(X^n)} \right) \quad (45)$$

(a) the sign \emptyset stands for Null, and is added in order to achieve same length on both sides length, i.e., $\{\emptyset Y^{n-1}\} = \text{length}\{X^n\}$, as required from the definition of directed information.

$$I(X^n \rightarrow Y^n) + I(Y^{n-1} \rightarrow X^n) = E \left(\log \frac{P(Y^n || X^n)}{P(Y^n)} \frac{P(X^n || Y^{n-1})}{P(X^n)} \right) \quad (46)$$

$$= E \left(\log \frac{P(X^n, Y^n)}{P(Y^n)P(X^n)} \right) \quad (47)$$

$$= I(X^n; Y^n) \quad (48)$$

■

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.