ITCT Lecture 3: The Asymptotic Equipartition Property and Entropy Rates

Prof. Ja-Ling Wu

Department of Computer Science
and Information Engineering
National Taiwan University

1

- Law of large Numbers:

  for independent, identically distributed (i.i.d.) random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to its expected value EX for large values of n.

- The Asymptotic Equipartition Property:

  $\frac{1}{n}\log\frac{1}{P(X_1,X_2,\cdots X_n)}$ is close to the entropy H, where $X_1,X_2,\ldots,X_n$ are i.i.d. random variables and $P(X_1,X_2,\ldots,X_n)$ is the probability of observing the sequence $X_1,X_2,\ldots,X_n$.

  Thus, the probability $P(X_1,X_2,\ldots,X_n)$ assigned to an observed sequence will be close to 2^{-nH}

2

- Almost all events are almost equally surprising:

$$P_r\left\{(X_1, X_2, \cdots, X_N) : P(X_1, X_2, \cdots, X_N) = 2^{-n(H \pm \varepsilon)}\right\} \cong 1$$

if $X_1, X_2, \ldots, X_n$ are i.i.d. ~ P(x)

- Thm: (AEP): If $X_1, X_2, \ldots$, are i.i.d. ~ P(X), then $-1/n \log P(X_1, X_2, \ldots, X_n) \rightarrow H(X)$ in probability

Proof: Since the X_i are i.i.d., so are $\log P(X_i)$. By the weak law of law of large numbers,

$$-\frac{1}{n}\log P(X_1, X_2, \cdots, X_n) = -\frac{1}{n}\sum \log P(X_i)$$
$$\rightarrow -E \log P(X) \text{ in probability}$$
$$= H(X)$$

- The typical set $A_\varepsilon^{(n)}$ w.r.t. P(X) is the set of sequences $(X_1, X_2, \ldots, X_n) \in \boldsymbol{X}^n$ with the following property:

$$2^{-n(H(X)+\varepsilon)} \leq P((X_1, X_2, \ldots, X_n) \leq 2^{-n(H(X)-\varepsilon)}$$

The properties of the typical set $A_\varepsilon^{(n)}$:

i. If $(x_1, x_2, \ldots, x_n) \in A_\varepsilon^{(n)}$, then $H(X) - \varepsilon \leq -\frac{1}{n}\log P(X_1, X_2, \cdots X_n) \leq H(X) + \varepsilon$

ii. $P_r\{A_\varepsilon^{(n)}\} > 1 - \varepsilon$ for n sufficiently large.

iii. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$

iv. $|A_\varepsilon^{(n)}| \geq (1 - \varepsilon)\, 2^{n(H(X)-\varepsilon)}$ for n sufficiently large.

Thus, the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical is nearly 2^{nH}.

proof :

(i) From the definition of typical set

$$2^{-n(H(x)+\varepsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(x)-\varepsilon)}$$

Taking logarithm, Dividing by n →

$$H(x) - \varepsilon \le -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \le H(x) + \varepsilon$$

(ii) Since the prob. of the event $(x_1, x_2, \ldots, x_n) \in A_\varepsilon^{(n)}$ tends to 1 as n→∞. Thus for any δ > 0, there exists an n_0, such that for all n ≧n_0, we have

$$P_r\left\{\left|-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) - H(x)\right| < \varepsilon\right\} > 1 - \delta$$

setting $\delta = \varepsilon$, we have $P_r\left\{A_\varepsilon^{(n)}\right\} > 1 - \varepsilon$

(iii) $1 = \sum_{\vartheta \in \mathbf{X}^n} P(\vartheta)$

$$\ge \sum_{\vartheta \in A_\varepsilon^{(n)}} P(\vartheta) \ge \sum_{\vartheta \in A_\varepsilon^{(n)}} 2^{-n(H(x)+\varepsilon)} = 2^{-n(H(x)+\varepsilon)} \left|A_\varepsilon^{(n)}\right|,$$

hence $\left|A_\varepsilon^{(n)}\right| \le 2^{n(H(x)+\varepsilon)}$

(iv) For sufficiently large n, $P_r\left\{A_\varepsilon^{(n)}\right\} > 1 - \varepsilon$ (from (ii))

$$1 - \varepsilon < P_r\left\{A_\varepsilon^{(n)}\right\} \le \sum_{\vartheta \in A_\varepsilon^{(n)}} 2^{-n(H(x)-\varepsilon)}$$

$$= 2^{-n(H(x)-\varepsilon)} \left|A_\varepsilon^{(n)}\right|$$

hence $\left|A_\varepsilon^{(n)}\right| \ge (1-\varepsilon) 2^{n(H(x)-\varepsilon)}$

Consequences of the AEP : Data Compression

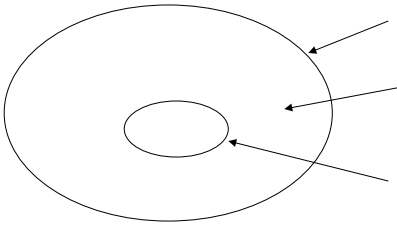
Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables~p(x)

Data compression / Compaction :

to find "short Descriptions" for such sequences of rv's.

We divide all sequences in $\mathbf{X}^n$ into two sets :

the typical set $A_\varepsilon^{(n)}$ and its complement $A_\varepsilon^{(n)c}$



We order all elements in each set according to some order (say lexicographic order).⇨each sequence of $A_\varepsilon^{(n)}$ can be represented by giving the index of the sequence in the set.

- Since there are $\leq 2^{n(H(x)+\varepsilon)}$ sequences in $A_\varepsilon^{(n)}$, the indexing requires no more than $n(H(x)+\varepsilon)+1$ bits.(The extra bit may be necessary because $n(H(x)+\varepsilon)$ may not be an integer.)

  We prefix all these sequences by a '0', giving a total length of ≦ $n(H(x)+\varepsilon)+2$ bits to represent each sequence in $A_\varepsilon^{(n)}$.

- Similarly, we can index each sequence in $A_\varepsilon^{(n)c}$ by using not more than $n\log|\mathbf{X}|+1$ bits.

  Prefixing these indices by '1', we have a code for all the sequences in $\boldsymbol{X}^n$.

Some features of the above coding scheme

- The code is one-to-one and easily decodable.

  The initial bit acts as a flag bit to indicate the length of the codeword that follows.
- We have used a brute force enumeration of the atypical set $A_\varepsilon^{(n)c}$ without taking into account the fact that the number of elements in $A_\varepsilon^{(n)c}$ is less than the number of elements in X^n. (Surprisingly, this is good enough to yield an efficient description.)

- The typical sequences have short description of length ≈ nH(x)

$$X^n \to \{x_1, x_2, \ldots, x_n\}$$

$\ell(x^n)$: the cordword length corresponding to x^n

$$E(\ell(x^n)) = \sum_{x^n} p(x^n)\ell(x^n) = \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)\ell(x^n) + \sum_{x^n \in A_\varepsilon^{(n)c}} p(x^n)\ell(x^n)$$

$$\leq \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)\left[n(H+\varepsilon)+2\right] + \sum_{x^n \in A_\varepsilon^{(n)c}} p(x^n)\left[n\log|\mathbf{X}|+2\right]$$

$$= P_r\{A_\varepsilon^{(n)}\}\left[n(H+\varepsilon)+2\right] + P_r\{A_\varepsilon^{(n)c}\}\left[n\log|\mathbf{X}|+2\right]$$

$$\approx 1 \qquad\qquad \approx \in$$

$$\leq n(H+\varepsilon) + \varepsilon\, n(\log|\mathbf{X}|) + 2 = n(H+\varepsilon')$$

where $\varepsilon' = \varepsilon + \varepsilon\log|\mathbf{X}| + \frac{2}{n}$ can be made arbitrarily small by an appropriate choice of ε and n.

Theorem :

Let X^n be i.i.d ~p(x). Let ε>0.

There exists a code which maps sequences X^n of length n into binary strings such that mapping is one-to-one (and therefore invertible) and

$$E\left[\frac{1}{n}\ell\left(x^n\right)\right] \le H(x)+\varepsilon$$

for sufficiently large n.

One can represent sequences X^n using nH(x) bits on the average!!

Markov's and Chebyshev's inequalities :

(i) Markov's inequality :

For any non-negative rv. X and any δ>0, then

$$P_r\{X \ge \delta\} \le \frac{EX}{\delta}$$

(ii) Chebyshev's inequality :

Let Y be a random variable with mean μ and variance σ^2. By letting $X=(Y-\mu)^2$, then for any ε>0,

$$P_r\{|Y-\mu| > \varepsilon\} \le \frac{\sigma^2}{\varepsilon^2}$$

Entropy Rate of a Stochastic Process

The prescribed AEP establishes that nH(X) bits suffice on the average to describe n independent and identically distributed (i.i.d.) random variables.

- But what if the random variables are dependent ?
  We will show, just as in i.i.d. case, that the entropy $H(X_1,X_2,\ldots,X_n)$ grows (asymptotically) linear with n at a rate $H(\boldsymbol{X})$, which will be called the entropy rate of the process.

Markov Chain

A stochastic process is an indexed sequence of rv's. In general, there can be an arbitrary dependence among them. The process is characterized by the joint Probability Mass Functions

$$P_r\{(X_1,X_2,\ldots,X_n)=(x_1,x_2,\ldots,x_n)\}=p(x_1,x_2,\ldots,x_n),$$
$$(x_1,x_2,\ldots,x_n)\in \mathbf{X}^n \quad \text{for n}=1,2,\ldots$$

Definition : A stochastic process is said be stationary if the joint distribution of any subset of the sequence of rv's is invariant w.r.t. shifts in the time index, i.e.,

$$P_r\{X_1=x_1,X_2=x_2,\ldots,X_n=x_n\}$$
$$=P_r\{X_{1+\ell}=x_1,X_{2+\ell}=x_2,\ldots,X_{n+\ell}=x_n\}$$
$$\text{for every shift } \ell \text{ and for all } x_1,x_2,\ldots,x_n\in\mathbf{X}$$

A simple example of a stochastic process with dependence is one in which each r.v. depends on the one preceding it and is conditionally independent of all the other preceding rv's. Such a process is said to be Markov.

Definition:

A discrete stochastic process $X_1, X_2, \ldots$ is said to be a Markov chain or a Markov process if, for n=1,2,…,

$$P_r\left(X_{n+1} = x_{n+1} \middle| X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\right)$$
$$= P_r\left(X_{n+1} = x_{n+1} \middle| X_n = x_n\right)$$
$$\text{for all} \quad x_1, x_2, \ldots, x_n, x_{n+1} \in \mathbf{X}$$

In this case, the joint probability mass function of the rv's can be written as :

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_n|x_{n-1})$$

Definition :

The Markov chain is said to be time invariant if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n, i.e., for n=1,2,…,

$$P_r\{X_{n+1} = b | X_n = a\} = P_r\{X_2 = b | X_1 = a\},$$
$$\text{for all a, b} \in \mathbf{X}$$

Recall:

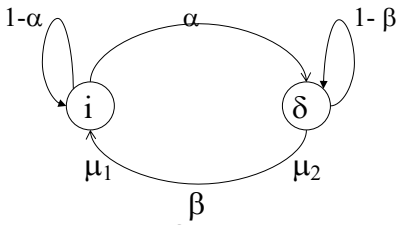
If $\{X_i\}$ is a Markov chain, then X_n is called the state at time n. A time invariant Markov chain is characterized by its initial state and a probability transition matrix $P=[P_{ij}]$, $i,j \in \{1,2,\ldots,m\}$, where $P_{ij}=P_r\{X_{n+1}=j|X_n=I\}$

- If it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps, then the Markov chain is said to be Irreducible. If the largest common factor of the lengths of different paths from a state to itself is 1, the Markov chain is said to aperiodic.
- If the pmf of the rv. at time n is $p(x_n)$, then the pmf at time n+1 is $P(x_{n+1}) = \sum_{x_n} p(x_n) P_{X_n X_{n+1}}$

- A distribution on the states such that the distribution at time n+1 is the same as the distribution at time n is called a stationary distribution.
- If the initial state of a Markov chain is drawn according to a stationary distribution, then the Markov chain forms a stationary process.
- If the finite state Markov chain is irreducible and aperiodic, then the stationary distribution is unique, and from any starting distribution, the distribution of X_n tends to the stationary distribution as $n \to \infty$.

Example: Consider a two-state Markov chain with a probability transition matrix

$$P_{ij} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$



The stationary probability can be found by solving the equation $\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ more simply, by balancing probabilities.

19

For the stationary distribution, the net probability flow across any cut-set in the state transition graph is 0.

Applying this property to the above state transition graph,

we obtain

$$\mu_1\alpha = \mu_2\beta$$

Since μ1+μ2=1, the stationary distribution is

$$\mu_1 = \frac{\beta}{\alpha+\beta}, \mu_2 = \frac{\alpha}{\alpha+\beta}$$

20

If the Markov chain has an initial state drawn according to the stationary distribution, the resulting process will be stationary. The entropy of the state X_n at time n is

$$H(X_n) = H\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right)$$

However, this is not the rate at which entropy grows for $H(x_1, x_2, \ldots, x_n)$

Entropy Rate :

If we have a sequence of n rv's, a natural question to ask is "how does the entropy of the sequence grow with n" .

Definition :

The Entropy Rate of a stochastic process $\{X_i\}$ is defined by

$$H(X) = \lim \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

when the limit exists.

Examples :

1. Typewriter

A typewriter that has m equally likely output letters. The typewriter can produce m^n sequences of length n, all of them equally likely. Hence

$H(X_1,X_2,\ldots X_n)=\log m^n$, and the entropy rate is

$$H(X)=\lim\frac{1}{n}H(X_1,X_2,\ldots,X_n)=\log m \ \ {}^{\text{bits}}\!/_{\text{symbol}}$$

2. $X_1,X_2,\ldots,X_n$ are i.i.d. rv's. Then

$$H(X)=\lim\frac{1}{n}H(X_1,X_2,\ldots,X_n)=\frac{nH(X_1)}{n}=H(X_1), \ {}^{\text{b}}\!/_{\text{s}}$$

3. Sequence of independent, but not identically distributed rv's.

In this case, $H(X_1,X_2,\ldots,X_n)=\sum_{i=1}^{n}H(X_i)$

but the $H(X_i)$'s are not all equal.

We can choose a sequence of distributions on $X_1,X_2,\ldots$ such that the limit of $\frac{1}{n}\sum H(X_i)$ does not exist.

An example of such a sequence is a random binary sequence where $P_i=P(X_i=1)$ is not constant , but a function of i, chosen carefully so that the limit in $\lim\frac{1}{n}H(X_1,X_2,\ldots,X_n)$ does not exist.

For example, let

$$P_i = \begin{cases} 0.5 & , \text{ if } 2k < \log\log i \le 2k+1 \\ 0 & , \text{ if } 2k+1 < \log\log i \le 2k+2 \end{cases}$$

Then there are arbitrarily long stretches where $H(x_i)=1$, followed by exponentially longer segments, where $H(x_i)=0$.

Hence the running average of the $H(x_i)$ will oscillate between 0 and 1 and will not have a limit.

Thus $H(\boldsymbol{X})$ is not definded for this process.

Now, let's define a related quantity for entropy rate : $H'(X) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$ when the limit exists.

Note that :

$H(X) = \lim \frac{1}{n} H(X_1, X_2, \ldots, X_n)$

: the per symbol entropy of the rv's.

$H'(X) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$

: the conditional entropy of the last rv. given the past

For stationary processes both limits exist and are equal.

Theorem :

For a stationary stochastic process,

$H(\boldsymbol{X})=H'(\boldsymbol{X})$

we will first prove that lim $H(X_n|X_{n-1},\ldots,X_1)$ exists.

Lemma 1: For a stationary stochastic process, $H(X_n|X_{n-1},\ldots,X_1)$ is decreasing in n and has a limit H'(***X***).

proof :

$$H\left(X_{n+1}\middle|X_n,X_{n-1},\ldots,X_1\right)\le H\left(X_{n+1}\middle|X_n,\ldots,X_2\right)$$
$$=H\left(X_n\middle|X_{n-1},\ldots,X_1\right)$$

Since $H(X_n|X_{n-1},\ldots,X_1)$ is a decreasing sequence of non-negative numbers, it has a limit H'(***X***)

27

Lemma 2 : (Ces'aro mean) :

$$\text{If } a_n\to a \text{ and } b_n=\frac{1}{n}\sum_{i=1}^{n}a_i$$

$$\text{then } b_n\to a$$

proof :

Since $a_n\to a$,there exists a number N(ε) such that $|a_n-a|\leqq\varepsilon$ for all $n\geqq N(\varepsilon)$.Hence

$$|b_n-a|=\left|\frac{1}{n}\sum_{i=1}^{n}a_i-a\right|=\left|\frac{1}{n}\sum_{i=1}^{n}(a_i-a)\right|$$
$$\le\frac{1}{n}\sum_{i=1}^{n}|(a_i-a)|$$
$$\le\frac{1}{n}\sum_{i=1}^{N(\varepsilon)}|a_i-a|+\frac{n-N(\varepsilon)}{n}\varepsilon$$
$$\le\frac{1}{n}\sum_{i=1}^{N(\varepsilon)}|a_i-a|+\varepsilon$$

28

for all n≧N(ε). Since the first term goes to zero as n→∞, we can make $|b_n-a|\leqq 2\varepsilon$ by taking n large enough. Hence $b_n \to a$ as $n\to\infty$.

Proof of the theorem : By the chain rule,

$$\frac{1}{n}H(X_1,X_2,\ldots,X_n)=\frac{1}{n}\sum_{i=1}^{n}H(X_i|X_{i-1},\ldots,X_1)$$

i.e., the entropy rate is the time average of the conditional entropies. But we know that the conditional entropies tend to have a limit H'(***X***). Hence by lemma 2, their running average has a limit, which is equal to H'(***X***). Thus,

$$H(X)=\lim\frac{1}{n}H(X_1,X_2,\ldots,X_n)=\lim H(X_n|X_{n-1},\ldots,X_1)=H'(X)$$

The significance for the entropy rate of a stochastic process arises from the AEP for a stationary Ergodic process.

remark :

For any stationary ergodic process,

$$-\frac{1}{n}\log p(X_1,X_2,\ldots,X_n)\to H(X)$$

with probability 1.

Markov chains :

For a stationary Markov chain, the entropy rate is given by

$H(\boldsymbol{X})=H'(\boldsymbol{X})=\lim H(X_n|X_{n-1},\ldots,X_1)$

$=\lim H(X_n|X_{n-1})=H(X_2|X_1)$

Theorem :

Let {Xi} be a stationary Markov chain with stationary distribution μ and transition matrix p. Then the entropy rate is :

$$H(X) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}$$

proof : $H(X) = H(X_2|X_1) = \sum_i \mu_i \left(\sum_j - P_{ij} \log P_{ij} \right)$

Therefore, the entropy rate of the two-state Markov chain example should be :

$$H(X) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta} H(\alpha) + \frac{\alpha}{\alpha+\beta} H(\beta)$$

Concluding Remark :

If the Markov chain is irreducible and aperiodic, then it has a unique stationary distribution on the states, and any initial distribution tends to the stationary distribution as n→∞.

Hidden Markov Models :

Let $X_1,X_2,\ldots,X_n,\ldots$ be a stationary Markov chain, and let $Y_i=\Phi(X_i)$ be a process, each term of which is a function of the corresponding state in the Markov chain.

Such functions of Markov chains occur often in practice!!

It would simplify matters greatly if $Y_1,Y_2,\ldots,Y_n$ also formed a Markov chain, but in many cases this is not true.

However, since the Markov chain is stationary, so is $Y_1,\ldots,Y_n$, and the Entropy rate is well defined.

However, if we wish to compute H(Y), we might compute $H(Y_n|Y_{n-1},\ldots,Y_1)$ for each n and find the limit. Since the convergence can be arbitrarily slow, we will never know how close we are to the limit; we will not know when to stop. (we can't look at the change between the values at n and n+1, since this difference may be small even when we are far away from the limit!!)

Remark: (bounded from 2 sides is better than from one side only)

It would be useful computationally to have upper and lower bounds converging to the limit from above and below!

We can halt the computation when the difference between the upper bound and the lower bound is small, and we will then have a good estimate of the limit.

recall : For a stationary stochastic process, $H(X_n|X_{n-1},\ldots,X_1)$ is decreasing in n and has a limit H'(***X***).

$\rightarrow H(Y_n|Y_{n-1},\ldots,Y_1)$ converges monotonically to H(y) from above. $\Rightarrow H(Y_n|Y_{n-1},\ldots,Y_1) \geqq H(y)$

For a lower bound, we will use $H(Y_n|Y_{n-1}, \ldots,Y_1,X_1)$.

This is a neat trick based on the idea that X_1 contains as much information about Y_n as Y_1, $Y_0,Y_{-1},\ldots$

Lemma : $H(Y_n|Y_{n-1},\ldots,Y_2,X_1) \leqq H(y)$

proof : we have, for k=1,2,…,

$$\overset{(a)}{H}(Y_n|Y_{n-1},\ldots,Y_2,X_1)$$
$$\overset{(b)}{=} H(Y_n|Y_{n-1},\ldots,Y_2,Y_1,X_1)$$
$$\overset{(c)}{=} H(Y_n|Y_{n-1},\ldots,Y_1,X_1,X_0,X_{-1},\ldots,X_{-k})$$
$$\overset{(d)}{=} H(Y_n|Y_{n-1},\ldots,Y1,X_1,X_0,X_{-1},\ldots X_{-k},Y_0,\ldots,Y_{-k})$$
$$\leqq H(Y_n|Y_{n-1},\ldots,Y_1,Y_0,\ldots,Y_{-k})$$
$$= H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1) \quad \text{(stationary)}$$

(a) Y_1 is a function of X_1
(b) Markovity of X
(c) Yi is a function of X_i
(d) conditioning reduces entropy

Since the inequality is true for all k, it is true in the limit. Thus,

$$H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \le \lim_k H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1)$$
$$= H(y)$$

The next lemma shows that the interval between the upper and the lower bounds decreases in length.

Lemma :

$H(Y_n|Y_{n-1},\ldots,Y_1)-H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \rightarrow 0$

37

proof : The interval length can be written as

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

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

By the properties of mutual information,

$I(X_1;Y_1,Y_2,\ldots,Y_n) \le H(X_1)$ and $I(X_1;Y_1,Y_2,\ldots,Y_n)$ increases with n

and hence

$$\lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n) \text{ existed and } \lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n) \le H(X_1)$$

By the chain rule,

$$H(X) \ge \lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n) = \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1;Y_i|Y_{i-1},\ldots,Y_1)$$
$$= \sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\ldots,Y_1)$$

Since this infinite sum is finite and the terms are non-negative, the terms must tend to 0, i.e.,

$\lim I(X_1;Y_n|Y_{n-1},\ldots,Y_1)=0$

38

Combining the previous two lemmas, we have the following theorem:

Theorem :

If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain,

and $Y_i = \Phi(X_i)$

then

$H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \leqq H(y) \leqq H(Y_n|Y_{n-1},\ldots,Y_1)$

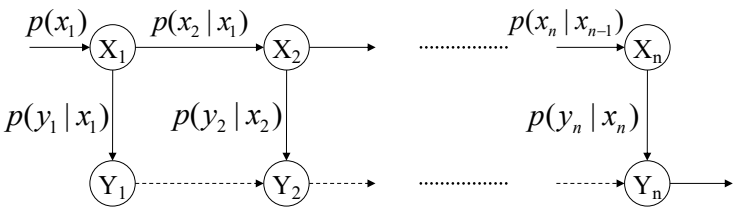
and

$\lim H(Y_n|Y_{n-1},\ldots,Y_1,X_1) = H(y) = \lim H(Y_n|Y_{n-1},\ldots,Y_1)$

In general, we could also consider the case where Y_i is a stochastic function as opposed to a deterministic function of X_i.

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

$$p(x^n, y^n) = p(x_1)\prod_{i=1}^{n-1} p(x_{i+1} \mid x_i)\prod_{i=1}^{n} p(y_i \mid x_i)$$



Such a process, called a hidden Markov model (HMM), is used extensively in speech recognition, handwriting recognition, computer vision, music analysis, and so on.

The same argument as that used for functions of a Markov chain carry over to HMM's, and we can lower bound the entropy rate of a HMM by conditioning it on the underlying Markov states.

41

Rate Distortion Functions

- Recall:

  For discrete-amplitude memoryless sources,

$$H(X) = E[I(X)] = -\sum_{k=1}^{K} P(x_k)\log_2 P(x_k)$$

  For discrete-amplitude source with memory,

$$H(X) = \lim_{N\to\infty} H_N(X)$$

$$= \lim_{N\to\infty}\left[-\frac{1}{N}\sum\sum\cdots\sum_{\text{all x}} P(X)\log_2 P(X)\right]$$

$$H(X)|_{\text{with memory}} < H(X)|_{\text{without memory}} \le \log_2 K$$

42

Differential Entropy

Source Redundancy $\stackrel{\Delta}{=} \log_2 K - H(X)$

- Non-uniform distribution of $P(x_K)$
- The present of memory

Non-Gaussian pdf for continuous source

For continuous-amplitude memoryless sources

, $$h(X) = E[-\log_2 P_X(X)] = -\int_{-\infty}^{\infty} P_x(x)\log_2 P_x(x)dx$$

For continous - amplitude sources with memory,

$$h(X) = \lim_{N\to\infty} h_N(X) = \lim_{N\to\infty}\left[-\frac{1}{N}\int\int_{-\infty}^{\infty}\cdots\int P_x(x)\log_2 P_x(x)dx\right]$$

$$h(X)|_{\text{with memory}} < h(X)|_{\text{without memory}} \le \frac{1}{2}\log_2\left(2\pi e\sigma_x^2\right)$$

Gaussian distribution

Source and Channel Coding Theorems

- According to the noisy channel coding theorem, information can be transmitted reliably (i.e., without error) over a noisy channel, at any source rate, say R, below a so-called capacity C of the channel

  $R < C$ for reliable transmission

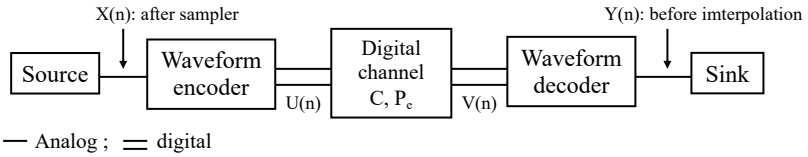
- According to the source coding theorem, there exists a mapping from source waveform to codewords such that for a given distortion D, R(D) bits (per source sample) are sufficient to enable waveform reconstruction with an average distortion that is arbitrarily close to D.

  The actual rate R has to obey:

  $R \ge R(D)$ for fidelity given by D

  The function R(D) is called the rate distortion function. Its inverse, D(R), is called the distortion rate function.

Digital Communication of Waveforms



— Analog ; ═ digital

Both C and R(D) are related to the concept of mutual information.

Let X and Y refer to sequences of coder input and decoder output; and let U and V refer to sequences of channel input and channel output.

I(X;Y) : the average information transmitted to the destination per sample

Distortion D and I(X;Y) depend on the type of source coding. There is a minimum of I(X;Y), that is needed for a reconstruction at the destination if the average distortion must not exceed the specified upper limit D.

This minimum value of I(X;Y) is R(D)!

The channel capacity C is related to the average mutual information I(U;V) per sample that characterizes channel input statistics and input-output mappings across the channel, as described by appropriate conditional probabilities.

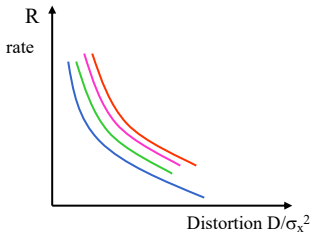
For a given channel, C is the maximum of I(U;V) over all channel input statistics.

- Information Transmission theorem:

  $C \geq R(D)$ for reliable transmission and fidelity D.

  : waveform reconstruction with fidelity D is possible after transmission over a noisy channel, provided that its capacity is greater than R(D).

- The significance of the information transmission theorem is that it justifies the separation of source coding (source representation) and channel coding (information transmission) functions in waveform communication and provides a framework wherein the channel controls the rate, but not necessarily the accuracy, of waveform reproduction.
  Thus the only property of the channel that should concern the source coder is the single parameter C and the only requirement for the efficient utilization of the channel is that its input statistics (statistics at output of source coder) should be so as to maximize I(U;V).



— : information theory bound
— : high complexity coder
— : medium complexity coder
— : low complexity coder

- The rate distortions are monotonically non-increasing (higher information-rate representations should lead to smaller average distortions)
- The distortion at Rate R=0 should be finite if the input variance is finite.
- For discrete-amplitude sources R(0) = H(X)

The Power of Lagrange Multiplier

Formal Proof of:

1. Uniform distribution of finite discrete sources gives maximal entropy

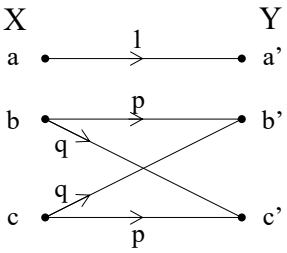
cost function: $Max\sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i}$ subject to: $\sum_{i=1}^{N} p_i = 1$

$$f(p_1, p_2, \ldots, p_N) = \frac{1}{\log_e 2}\sum_{i=1}^{N} p_i \log_e \frac{1}{p_i} + \lambda\left(\sum_{i=1}^{N} p_i - 1\right)$$

$$\frac{\partial f}{\partial p_j} = \frac{1}{\log_e 2}\left[\log\left(\frac{1}{p_j}\right) - 1\right] + \lambda = 0 \quad for\ j = 1,2,\ldots,N$$

Since everything in the above equation is a constant, it follows that each p_j is the same. If all the p_j are equal, each has value 1/N, and hence the maximum entropy is $H_2(p) = \log_2 N$.

2.



$p(a) = \mathbb{P}$

$p(b) = p(c) = \mathbb{Q}$

$\mathbb{P} + 2\mathbb{Q} = 1 \quad \leftarrow$ constraint

Let $\alpha = -[p\log p + q\log q]$

$H(X) = -[\mathbb{P}\log\mathbb{P} + 2\mathbb{Q}\log\mathbb{Q}]$

$H(X|Y) = 2\mathbb{Q}\alpha$

In order to find channel capacity, we have to choose $\mathbb{P}$ and $\mathbb{Q}$ in such a way as to maximize $H(X)$-$H(X|Y)$, subject to the constraint $\mathbb{P} + 2\mathbb{Q} = 1$.

Consider $U = -\mathbb{P}\log\mathbb{P} - 2\mathbb{Q}\log\mathbb{Q} - 2\mathbb{Q}\alpha + \lambda(\mathbb{P} + 2\mathbb{Q})$

$$\left.\begin{aligned}\frac{\partial U}{\partial \mathbb{P}} &= -1 - \log\mathbb{P} + \lambda = 0\\ \frac{\partial U}{\partial \mathbb{Q}} &= -2 - 2\log\mathbb{Q} - 2\alpha + 2\lambda = 0\end{aligned}\right\} \begin{array}{l}\text{Eliminating } \lambda \Rightarrow\\ \log\mathbb{P} = \log\mathbb{Q} + \alpha\\ \mathbb{P} = \mathbb{Q}e^{\alpha} = \mathbb{Q}\beta\end{array}$$

$\Rightarrow \mathbb{P} = \frac{\beta}{\beta+2}$ and $\mathbb{Q} = \frac{1}{\beta+2}$

then $c = \log\frac{\beta+2}{\beta}$

$\mathbb{P} + 2\mathbb{Q} = 1$

$\mathbb{Q}\beta + 2\mathbb{Q} = 1$

$\mathbb{Q} = \frac{1}{\beta+2}$

$\mathbb{P} = \mathbb{Q}\beta = \frac{\beta}{\beta+2}$