1 Review From Previous Lecture

1.1 Typical Set Properties

- Individual probability:
  \[ x \in T^n_\varepsilon \Rightarrow -nH(X) - n\varepsilon \leq \log p(x) \leq -nH(X) + n\varepsilon \]

- Total Probability:
  \[ P(x \in T^n_\varepsilon) > 1 - \varepsilon, \quad \text{for } n > N_\varepsilon \]

- Size:
  \[ (1 - \varepsilon)2^{n(H(X) - \varepsilon)} < |T^n_\varepsilon| \leq 2^{n(H(X) + \varepsilon)} \]

1.2 Asymptotic Equipartition Principle

Theorem 1.1 For any \( \varepsilon \), and \( n > N_\varepsilon \), almost any event is almost equally surprising.

\[ P(x \in T^n_\varepsilon) > 1 - \varepsilon, \quad \text{for } n > N_\varepsilon \]

Figure 1: Typical Set
From the Figure 1, we can see that a typical set can be small compared to the set of all the possible sequence, but it contains almost all the probability mass.

1.3 Source Coding and Data Compression

- $x \in T^n_{\varepsilon}$: "0" + at most $1 + n(H(X) + \varepsilon)$ bits.
- $x \notin T^n_{\varepsilon}$: "1" + at most $1 + n \log |X|$ bits.
- $L_n \leq (1 - \varepsilon)(1 + n(H(X) + \varepsilon)) + \varepsilon(1 + n \log |X|) = n(H(X) + \varepsilon + \varepsilon \log |X| + \frac{2}{n})$

2 Typical Set

2.1 Choice of $N_{\varepsilon}$

What $N_{\varepsilon}$, to ensure that $P(x^n \in T^n_{\varepsilon}) > 1 - \varepsilon$?

Answer: From Weak Law of Large Number (WLLN), suppose we have

$$Var(-\log p(x_i)) = \sigma^2$$

then for any $n$ and $\varepsilon$,

$$\varepsilon^2 P(|\frac{1}{n}\sum_i - \log p(x_i) - H(X)| > \varepsilon) \leq \frac{\sigma^2}{n} \quad \text{(according to Chevyshev inequality)}$$

if we choose $N_{\varepsilon} = \sigma^2 \varepsilon^{-3}$,

$$P(x \notin T^n_{\varepsilon}) \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow P(x \in T^n_{\varepsilon}) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

if $N_{\varepsilon} = \sigma^2 \varepsilon^{-3}$, then we have for $n > N_{\varepsilon}$,

$$P(x \in T^n_{\varepsilon}) \geq 1 - \varepsilon$$

for this choice of $N_{\varepsilon} = \sigma^2 \varepsilon^{-3}$,

$$-nH(X) - \sigma^2 \varepsilon^{-2} \leq \log p(x) \leq -nH(X) + \sigma^2 \varepsilon^{-2}$$

Within the typical set, the probability of a typical sequence can vary up to by a factor of $2\sigma^2 \varepsilon^{-2}$. 
2.2 Smallest High Probability Set

- $T_n^\varepsilon$ is a small subset of $X^n$ containing most of the probability mass.

- The way we show this is to show that if we pick a set $S(n)^\varepsilon$, where $|S(n)^\varepsilon| < 2^{n(H(X) - 2\varepsilon)}$, then $P(x \in S(n)^\varepsilon)$ (when $\varepsilon$ is small)

\[
P(x \in S(n)) = P(x \in S(n)^\varepsilon \cap T_n^{\varepsilon}) + P(x \in S(n)^\varepsilon \cap \bar{T}_n^{\varepsilon})
\]

\[
< \frac{|S(n)^\varepsilon|}{p(x \in T_n^{\varepsilon})} + P(x \notin T_n^{\varepsilon})
\]

\[
< 2^{n(H(X) - 2\varepsilon)} 2^{-n(H(X) - \varepsilon)} + \varepsilon 
\]

\[
= 2^{-n\varepsilon} + \varepsilon < 2\varepsilon 
\]

(for $n > N_\varepsilon$)

2.3 Summary for Typical Set

- Typical Set
  - Individual probability:

\[
\mathbb{P}(x \in T_n^{\varepsilon}) \Rightarrow -nH(X) - n\varepsilon \leq \log p(x) \leq -nH(X) + n\varepsilon
\]

  - Total Probability:

\[
P(x \in T_n^{\varepsilon}) > 1 - \varepsilon, \text{ for } n > N_\varepsilon = \sigma^2\varepsilon^{-3}
\]

  - Size:

\[
(1 - \varepsilon)2^{n(H(X) - \varepsilon)} \leq |T_n^{\varepsilon}| \leq 2^{n(H(X) + \varepsilon)}
\]

- No other high probability set can be much smaller than $T_n^{\varepsilon}$.

- Asymptotic equipartition principle: Almost all event sequences are equally surprising.

3 Source and Channel Coding

![Source and Channel Coding](image)

The input could be raw image or raw video. Common compression format for raw image would be JPEG. MPEG is a common compression format for raw video.
• Source Coding
  Compress data to remove redundancy.

• Channel Coding
  Add redundancy to protect against channel errors.

3.1 Discrete Memoryless Channel (D.M.C.)

![Simple communication system diagram](image)

- Discrete input, discrete output
  
  \[ x \in \mathcal{X}, \quad y \in \mathcal{Y} \]

- Channel matrix \( Q \)

  For entry \( ij \) in the matrix \( Q \), we have
  
  \[ Q_{ij} = P(Y = y_j | X = x_i) \]

  and
  
  \[ P(Y = y_i) = P(X = x_i)P(Y = y_j | X = x_i) \]

  or
  
  \[ p_Y = Q^T p_X \] (Think of this as in matrix form)

  where \( Q \in \mathbb{R}^{m \times n} \), \( p_Y = [y_1 \ y_2 \ \cdots \ y_n]^T \), and \( p_X = [x_1 \ x_2 \ \cdots \ x_m]^T \), also the sum of each row of \( Q = 1 \); i.e., \( \sum_j Q_{ij} = 1 \)

- Memoryless

  \[ p(y_n | x^{(n)}, y^{(n-1)}) = p(y_n | x_n) \]

  where \( y^{(n-1)} = (y_1, y_2, \cdots, y_{n-1}) \)
3.2 Binary Channels

3.2.1 Binary Symmetric Channel

In binary symmetric channel, $x = [0, 1], y = [0, 1]$. The channel characteristic is shown in Figure 4.

From Figure 4, we could derive the following Table 1.

<table>
<thead>
<tr>
<th>$Y$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$1 - p$</td>
<td>$p$</td>
</tr>
<tr>
<td>1</td>
<td>$p$</td>
<td>$1 - p$</td>
</tr>
</tbody>
</table>

Table 1: Binary Symmetric Channel Characteristic

From Table 1, we could easily find that $Q = \begin{pmatrix} 1 - p & p \\ p & 1 - p \end{pmatrix}$.

3.2.2 Binary Erasure Channel

In binary erasure channel, $x = [0, 1], y = [0, ?, 1]$. The channel characteristic is shown in Figure 5.

From Figure 5, we could derive the following Table 2.
Table 2: Binary Erasure Channel Characteristic

From Table 2, we could easily find that \( Q = \begin{pmatrix} 1 - p & p & 0 \\ 0 & p & 1 - p \end{pmatrix} \).

### 3.2.3 Z Channel

In Z channel, \( x = [0, 1], y = [0, 1] \). The channel characteristic is shown in Figure 6.

![Figure 6: Diagram of Z Channel](image)

From Figure 6, we could derive the following Table 3.

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 - p</td>
<td>p</td>
</tr>
<tr>
<td>1</td>
<td>p</td>
<td>1 - p</td>
</tr>
</tbody>
</table>

Table 3: Z Channel Characteristic

From Table 3, we could easily find that \( Q = \begin{pmatrix} 1 & 0 \\ p & 1 - p \end{pmatrix} \).

### 3.3 Weakly Symmetric Channels

1. All columns of \( Q \) have the same sum
   
   If \( X \) is uniform distributed \( p(x) = |X|^{-1} \), then \( Y \) is also uniform. 
   
   \[
   p(y) = \sum_{x \in X} p(y|x)p(x) = |X|^{-1}|X||Y|^{-1} = |Y|^{-1}
   \]

2. Rows of \( Q \) are permutation of each other
   
   Entropy of each row will be the same. 
   
   \[
   H(Y|X) = \sum_{x \in X} p(x)H(Y|X = x) = H(Q_1)
   \]
3.4 Symmetric Channel

1. Rows of \( Q \) are permutations of each other
2. Columns of \( Q \) are permutations of each other

Side note: Symmetry \( \implies \) Weakly symmetric

Example \( Q_1 = \begin{pmatrix} 1/3 & 1/6 & 1/2 \\ 1/2 & 1/3 & 1/6 \end{pmatrix} \), \( Q_2 = \begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.3 & 0.4 & 0.3 \\ 0.4 & 0.3 & 0.3 \end{pmatrix} \). Are \( Q_1 \) and \( Q_2 \) weakly symmetric?

Answer: \( Q_1 \) is not a weakly symmetric channel because not all columns of \( Q_1 \) have the same sum. \( Q_2 \) is a weakly symmetric channel because all columns of \( Q_2 \) have the same sum.

4 Channel Capacity

4.1 Capacity of Discrete Memoryless Channel

Definition 4.1 We can define the channel capacity of a discrete memoryless channel as

\[
C = \max_{p(x)} I(X; Y)
\]

- The maximum over all possible input distribution of \( p(x) \).
- \( \exists \) only one maximum \( I(X; Y) \) since \( I(X; Y) \) is concave in \( p(x) \) given \( p(y|x) \).
- We want to find \( p(x) \) that maximize \( I(X; Y) \).
- Limit on \( C \): \( 0 \leq C \leq \min(H(X), H(Y)) \leq \min(\log |X|, \log |Y|) \).

4.2 Capacity of \( n \) Uses of Channel

Definition 4.2 We can define the channel capacity of \( n \) uses channel as:

\[
C^n = \frac{1}{n} \max_{p(x)} I(X_1, X_2, X_3, \ldots, X_n; Y_1, Y_2, Y_3, \ldots, Y_n)
\]

4.3 Mutual Information

Example From Figure 6 and 7, they show the Binary Symmetric Channel with an input random variable \( x \), and \( x \) follows Bernoulli distribution. \( y \) represents for the output of this channel. In this case, we want to find the mutual information \( I(X; Y) \).
\[ I(X;Y) = H(Y) - H(Y|X) \]

From Figure 7 and Figure 8, we can see the probability of \( Y = 0 \) and \( Y = 1 \)

\[
\begin{align*}
P(Y = 0) &= pf + (1 - p)(1 - f) \\
P(Y = 1) &= (1 - p)f + p(1 - f)
\end{align*}
\]

\[
\Rightarrow H(Y) = H(1 + 2pf - p - f) = H(f + p - 2pf)
\]
\[
\Rightarrow H(Y | X) = H(Y | X = 0)P(X = 0) + H(Y | X = 1)P(X = 1)
\]
\[
\Rightarrow (1 - p)H(f) + H(f)p = H(f) = I(X;Y)
\]

**4.4 Mutual Information is Concave In \( p(x) \)**

Mutual information is concaved in \( p(x) \) for a fixed \( p(y|x) \).
**Proof** Let $U$ and $V$ be two random variables. We are trying to prove $u$ and $v$ respectively. Define:

$$X = \begin{cases} U & Z = 1, \text{ with probability } \lambda \\ V & Z = 0, \text{ with probability } 1 - \lambda \end{cases}$$

Here, $z$ acts as a switch with probability $\lambda$ to be 1 and with probability $1 - \lambda$ to be 0 as shown in Figure 9.

![Figure 9: z Acting as a Switch](image)

Thus, it is easy to see that

$$p(x) = \lambda u + (1 - \lambda)v$$

Figure 10 is a visual example of a concave function with random variable $U$ and $V$. We will show that,

$$I(X; Y) = I(\lambda u + (1 - \lambda)v; Y) \geq I(U; Y) + (1 - \lambda)I(V; Y)$$

$$I(X, Z; Y) = I(X; Y) + I(Z; Y|X) = I(Z; Y) + I(X; Y|Z)$$

(1)

(2)
(Note: equation (1) represents using conditional mutual information on $X$, and (2) represents using conditional mutual information on $Z$.)

But,

$$I(Z; Y|X) = H(Y|X) - H(Y|Z, X) = H(Y|X) - H(Y|X) = 0 \quad (3)$$

From (1), (2), (3), we have

$$I(X; Y) \geq I(X; Y|Z) = I(X; Y|Z = 1)P(Z = 1) + I(X; Y|Z = 0)P(Z = 0)$$

$$\geq \lambda I(U; Y) + (1 - \lambda)I(V; Y)$$

So we proved that,

$$I(X; Y) \geq \lambda I(U; Y) + (1 - \lambda)I(V; Y)$$