Review Summary

- **Definitions:**

$$
\begin{aligned}
H(X) &= E_p \log \frac{1}{p(X)} \\
H(X,Y) &= E_p \log \frac{1}{p(X,Y)} \\
H(X|Y) &= E_p \log \frac{1}{p(X|Y)} \\
I(X;Y) &= E_p \log \frac{p(X,Y)}{p(X)p(Y)} \\
D(p\|q) &= E_p \log \frac{p(X)}{q(X)} \\
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y).
\end{aligned}
$$

Review Summary

- **Chain rules:**

  Entropy:

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

  Mutual information:

  $I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \ldots, X_{i-1}).$

  Relative entropy:

  $D\big(p(x, y) \| q(x, y)\big) = D\big(p(x) \| q(x)\big) + D\big(p(y|x) \| q(y|x)\big).$

Inequalities related to D and I

1. $D(p\|q) \geq 0$ with equality iff $p(x) = q(x)$, for all $x \in \mathcal{X}$ (*information inequality*).
2. $I(X;Y) = D(p(x,y)\|p(x)p(y)) \geq 0$, with equality iff $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent).
3. If $|\mathcal{X}| = m$, and u is the uniform distribution over $\mathcal{X}$, then $D(p\|u) = \log m - H(p)$.

Jensen's Inequality

If f is a convex function, then $E[f(X)] \geq f(E[X])$.

Data-processing inequality

If $X \to Y \to Z$ forms a Markov chain, then $I(X;Y) \geq I(X;Z)$.

Review Summary

Theorem (AEP)

"Almost all events are almost equally surprising." Specifically, if $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

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

Definition

The *typical set* $A_\epsilon^{(n)}$ is the set of sequences $x_1, x_2, \ldots, x_n$ satisfying

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

Review Summary

Properties of the typical set

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then
   $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$.
2. $\Pr[A_\epsilon^{(n)}] > 1 - \epsilon$ for n sufficiently large.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the cardinality of the set A.
4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Theorem

Let X^n be i.i.d. $\sim p(x)$. There exists a code that one-to-one maps sequences x^n of length n into binary strings with

$$E[\frac{1}{n}\ell(X^n)] \leq H(X) + \epsilon$$

for n sufficiently large.

Review Summary

- **Classes of codes**

  Prefix codes $\Rightarrow$ Uniquely decodable codes $\Rightarrow$ Nonsingular codes

- **Kraft inequality**

  Prefix codes $\Leftrightarrow \sum D^{-\ell_i} \leq 1$.

- **McMillan inequality**

  Uniquely decodable codes $\Leftrightarrow \sum D^{-\ell_i} \leq 1$.

- **Huffman code**

$$L^* = \min_{\sum D^{-\ell_i} \leq 1} \sum p_i \ell_i$$

$$H_D(X) \leq L^* < H_D(X) + 1.$$

Review Summary

- **Entropy rate**. Two definitions of entropy rate for a stochastic process are

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n),$$
$$H^{'}(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2} \ldots, X_1).$$

For a **stationary** stochastic process, $H(\mathcal{X}) = H^{'}(\mathcal{X})$.

- Entropy rate of a stationary Markov chain.

$$H(\mathcal{X}) = -\sum_{i,j} \mu_i P_{ij} \log P_{ij}.$$

Review

- **Channel capacity.** The logarithm of the number of distinguishable inputs is given by

$$C = \max_{p(x)} I(X;Y).$$

- **Examples**
  - Binary symmetric channel: $C = 1 - H(p)$
  - Binary erasure channel: $C = 1 - \alpha$
  - Symmetric channel: $C = \log|\mathcal{Y}| - H$ (row of trans. matrix)

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Joint Typical Set

- Joint typicality. Given two i.i.d. random variable sequences X^n and Y^n, the set of jointly typical sequences is

$$A_\epsilon^{(n)} = \Bigg\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \\ \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon \\ \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \Bigg\}$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$.

Joint AEP

- **Joint AEP** Let (X^n, Y^n) be the sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$, then:

  1. $\Pr\left[(X^n, Y^n) \in A_\epsilon^{(n)}\right] \to 1$ as $n \to \infty$.
  2. $\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\epsilon)}$.
  3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

  $$\Pr\left[\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_\epsilon^{(n)}\right] \leq 2^{-n(I(X;Y)-3\epsilon)}.$$

  Please refer to p196 for the proof (proof of Theorem 7.6.1)

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Channel Coding Theorem

Theorem (Channel coding theorem)

For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.

Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Achievability: when $R < C$, there exists zero-error code.
Converse: zero-error codes must have $R \leq C$.

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Differential Entropy - 2

Theorem

The capacity of a Gaussian channel with power constraint P and noise variance N is

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N}\right) \quad \textit{bits per transmission}$$

Proof.

Use the same ideas as in the proof of the channel coding theorem in the discrete case to prove:
1) achievability; 2) converse

□

Two main differences:
1) the power constraint P;
2) the variables are continuous