1   Information Measures

1.1   Independence and Markov Chain

Definition 2.1 (Independence)   Two random variables X and Y are independent, denoted by $X \perp Y$, if

$$p(x, y) = p(x)p(y)$$

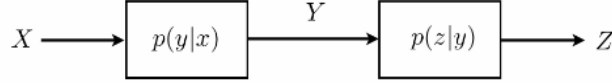for all x and y (i.e., for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$).

$$X \longrightarrow \boxed{p(y|x)} \xrightarrow{\;Y\;} \boxed{p(z|y)} \longrightarrow Z$$

Figure 1: Conceptually, when $X \perp Z \mid Y$, X, Y, Z are related as above.

Definition 2.2 (Mutual Independence)   For $n \geq 3$, random variables $X_1, X_2, \cdots, X_n$ are mutually independent if, for all $x_1, x_2, \cdots, x_n$

$$p(x_1, x_2, \cdots, x_n) = p(x_1) \, p(x_2) \cdots p(x_n)$$

Definition 2.3 (Pairwise Independence)   For $n \geq 3$, random variables $X_1, X_2, \cdots, X_n$ are pairwise independent if X_i and X_j are independent for all $1 \leq i < j \leq n$

Definition 2.4 (Conditional Independence)   For random variables X, Y, and Z, X is independent of Z conditioning on Y, denoted by $X \perp Z \mid Y$, if

$$p(x, y, z) = \begin{cases} \frac{p(x,y)p(y,z)}{p(y)} = p(x, y)p(z \mid y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Proposition 2.5   For random variables X, Y, and $Z, X \perp Z \mid Y$ if and only if

$$p(x, y, z) = a(x, y)b(y, z)$$

for all x, y, and z such that $p(y) > 0$

Proof A.   'Only if' part.   Assume $p(x, y, z)$ takes the form in Definition 2.4. For all x and for all y such that $p(y) > 0$, let

$$a(x, y) = \frac{p(x, y)}{p(y)} \qquad b(y, z) = p(y, z)$$

Proof B.   'If' part.
1. Assume that for all x, y, and z such that $p(y) > 0$,

$$p(x, y, z) = a(x, y)b(y, z)$$

2. Then for such x, y, and z, we have

$$p(x, y) = \sum_z p(x, y, z) = \sum_z a(x, y)b(y, z) = a(x, y) \sum_z b(y, z)$$

$$p(y, z) = \sum_x p(x, y, z) = \sum_x a(x, y)b(y, z) = b(y, z) \sum_x a(x, y)$$

3. Furthermore,

$$p(y) = \sum_z p(y, z) = \left(\sum_x a(x, y) \right) \left(\sum_z b(y, z) \right) > 0$$

4. Therefore,

$$\frac{p(x, y)p(y, z)}{p(y)} = \frac{(a(x, y) \sum_z b(y, z)) (b(y, z) \sum_x a(x, y))}{(\sum_x a(x, y)) (\sum_z b(y, z))} = a(x, y)b(y, z) = p(x, y, z)$$

5. And for x, y, and z such that $p(y) = 0$, since

$$0 \le p(x, y, z) \le p(y) = 0 \quad \rightarrow \quad p(x, y, z) = 0$$

Definition 2.6 (Markov Chain) For random variables $X_1, X_2, \cdots, X_n$, where $n \ge 3$, $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if

$$p(x_1, x_2, \cdots, x_n) = \begin{cases} p(x_1, x_2) \, p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1}) & \text{if } p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Remark. $X_1 \to X_2 \to X_3$ is equivalent to $X_1 \perp X_3 \mid X_2$

Proposition 2.7 $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if $X_n \to X_{n-1} \to \cdots \to X_1$ forms a Markov chain.

Proposition 2.8 $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if

$$X_1 \to X_2 \to X_3$$
$$(X_1, X_2) \to X_3 \to X_4$$
$$\vdots$$
$$(X_1, X_2, \cdots, X_{n-2}) \to X_{n-1} \to X_n$$

form Markov chains.

Proposition 2.9 $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if

$$p(x_1, x_2, \cdots, x_n) = f_1(x_1, x_2) f_2(x_2, x_3) \cdots f_{n-1}(x_{n-1}, x_n)$$

for all $x_1, x_2, \cdots, x_n$ such that $p(x_2), p(x_3), \cdots, p(x_{n-1})$

Proposition 2.10 (Markov subchains) Let $\mathcal{N}_n = \{1, 2, \cdots, n\}$ and let $X_1 \to X_2 \to \cdots \to X_n$ form a Markov chain. For any subset α of $\mathcal{N}_n$, denote $(X_i, i \in \alpha)$ by X_α. Then for any disjoint subsets $\alpha_1, \alpha_2, \cdots, \alpha_m$ of $\mathcal{N}_n$ such that

$$k_1 < k_2 < \cdots < k_m$$

for all $k_j \in \alpha_j, j = 1, 2, \cdots, m$

$$X_{\alpha_1} \to X_{\alpha_2} \to \cdots \to X_{\alpha_m}$$

forms a Markov chain. That is, a subchain of $X_1 \to X_2 \to \cdots \to X_n$ is also a Markov chain. (Exercise)
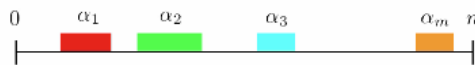

Figure 2: Markov subchains.

1.2   Shannon's Information Measures

- Entropy
- Conditional entropy
- Mutual information
- Conditional mutual information

Definition 2.13 (Entropy.)   The entropy $H(X)$ of a random variable X is defined as

$$H(X) = -\sum_x p(x) \log p(x)$$

- Convention: summation is taken over $\mathcal{S}_X$.

- When the base of the logarithm is α, write $H(X)$ as $H_\alpha(X)$.

- Entropy measures the uncertainty of a discrete random variable.

- The unit for entropy is

$$\begin{array}{ll} \text{bit} & \text{if } \alpha = 2 \\ \text{nat} & \text{if } \alpha = e \\ D - \text{it} & \text{if } \alpha = D \end{array}$$

Example (Binary Entropy Function).   For $0 \le \gamma \le 1$, define the binary entropy function

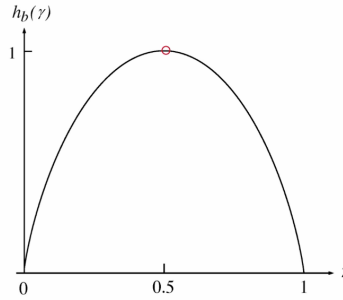$$h_b(\gamma) = -\gamma \log \gamma - (1 - \gamma) \log(1 - \gamma)$$

Figure 3: Binary Entropy Function.

Definition 2.14 (Joint Entropy)   The joint entropy $H(X, Y)$ of a pair of random variables X and Y is defined as

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) = -E \log p(X, Y)$$

Definition 2.15 (Conditional Entropy)   For random variables X and Y, the conditional entropy of Ygiven X is defined as

$$H(Y \mid X) = -\sum_{x,y} p(x, y) \log p(y \mid x) = -E \log p(Y \mid X)$$

Proposition 2.16

$$H(X, Y) = H(X) + H(Y \mid X)$$

$$H(X, Y) = H(Y) + H(X \mid Y)$$

Definition 2.17 (Mutual Information) For random variables X and Y, the mutual information between X and Y is defined as

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = E \log \frac{p(X,Y)}{p(X)p(Y)}$$

Remark $I(X;Y)$ is symmetrical in X and Y.
Remark Alternatively, we can write

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x \mid y)}{p(x)} = E \log \frac{p(X \mid Y)}{p(X)}$$

Proposition 2.18 The mutual information between a random variable X and itself is equal to the entropy of X, i.e., $I(X;X) = H(X)$

Proposition 2.19

$$I(X;Y) = H(X) - H(X \mid Y)$$
$$I(X;Y) = H(Y) - H(Y \mid X)$$
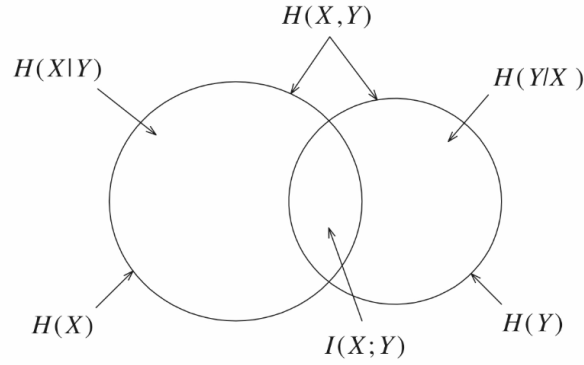
and

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Figure 4: Information Diagram.

Definition 2.20 For random variables X, Y and Z, the mutual information between X and Y conditioning on Z is defined as

$$I(X;Y \mid Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} = E \log \frac{p(X,Y \mid Z)}{p(X \mid Z)p(Y \mid Z)}$$

1.3   Continuity of Shannon's Information Measures for Fixed Finite Alphabets

Definition 2.23 (Variational Distance)   Let p and q be two probability distributions on a common alphabet $\mathcal{X}$. 'l'he variational distance between p ind q is defined as

$$V(p,q) = \sum_{x \in \mathcal{X}'} p(x) - q(x) \mid$$

The entropy funetion is continuous at p if

$$\lim_{p' \to p} H\left(p'\right) - H\left(\lim_{p' \to p'} p'\right) - H(p)$$

or equivalently, for any $\epsilon > 0$, there exists $\delta \to 0$ such that

$$\mid H(p) - H(q) \mid < \epsilon$$

for all $q \in \mathcal{P}_x$ satisfying

$$V(p,q) < \delta$$

1.4   Chain Rules

Proposition 2.24(Chain Rule for Entropy)

$$H\left(X_1, X_2, \cdots, X_n\right) = \sum_{i=1}^{n} H\left(X_i \mid X_1, \cdots, X_{i-1}\right)$$

Proposition 2.25 (Chain Rule for Conditional Entropy)

$$H\left(X_1, X_2, \cdots, X_n \mid Y\right) = \sum_{i=1}^{n} H\left(X_i \mid X_1, \cdots, X_{i-1}, Y\right)$$

Proposition 2.26 (Chain Rule for Mutual Information)

$$I\left(X_1, X_2, \cdots, X_n; Y\right) = \sum_{i=1}^{n} I\left(X_i; Y \mid X_1, \cdots, X_{i-1}\right)$$

Proposition 2.27 (Chain Rule for Conditional Mutual Information)

$$I\left(X_1, X_2, \cdots, X_n; Y \mid Z\right) = \sum_{i=1}^{n} I\left(X_i; Y \mid X_1, \cdots, X_{i-1}, Z\right)$$

Alternative Proof of Proposition 2.25

$$
\begin{aligned}
H\left(X_1, X_2, \cdots, X_n \mid Y\right) &= \sum_{y} p(y) H\left(X_1, X_2, \cdots, X_n \mid Y = y\right) \\
&= \sum_{y} p(y) \sum_{i=1}^{n} H\left(X_i \mid X_1, \cdots, X_{i-1}, Y = y\right) \\
&= \sum_{i=1}^{n} \sum_{y} p(y) H\left(X_i \mid X_1, \cdots, X_{i-1}, Y = y\right) \\
&= \sum_{i=1}^{n} H\left(X_i \mid X_1, \cdots, X_{i-1}, Y\right)
\end{aligned}
$$

Remark   This alternative proof explains why Proposition 2.25 can be obtained from Proposition 2.24 by conditioning on Y.

1.5 Information Divergence

Definition 2.28 (Information Divergence) The informational divergence between two probability distributions p and q on a common alphabet $\mathcal{X}$ is defined as

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}$$

where E_p denotes expectation with respect to p

Convention:

- Summation is over $\mathcal{S}_p$, i.e., $\sum_{x \in \mathcal{S}_p}$

- $c\log \frac{c}{0} = \infty$ for $c > 0$

- If $D(p\|q) < \infty$, then $p(x) > 0 \Rightarrow q(x) > 0$, or $\mathcal{S}_p \subset \mathcal{S}_q$

- $D(p\|q)$ measures the "distance" between p and q

- $\boldsymbol{D}(p\|q)$ is not symmetrical in p and q, so $D(\cdot\|\cdot)$ is not a true metric.

- $D(\cdot\|\cdot)$ does not satisfy the triangular inequality.

Lemma 2.29 (Fundamental Inequality) For any $a > 0$,

$$\ln a \leq a - 1$$

with equality if and only if $a = 1$

Corollary 2.30 For any $a > 0$,

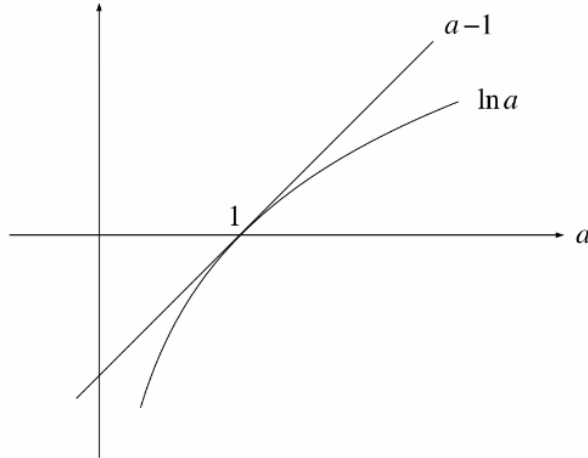$$\ln a \geq 1 - \frac{1}{a}$$

with equality if and only if $a = 1$.

Figure 5: Information Diagram.

> **Theorem 2.31 (Divergence Inequality)** For any two probability distributions p and q on a common alphabet $\mathcal{X}$
> $$D(p\|q) \geq 0$$
> with equality if and only if $p = q$

Proof. For simplicity, assume $\mathcal{S}_p = \mathcal{S}_q$, consider

$$
\begin{aligned}
D(p\|q) &= \sum_{x \in S_p} p(x) \log \frac{p(x)}{q(x)} \\
&= (\log c) \sum_{x \in \mathcal{S}_p} p(x) \ln \frac{p(x)}{q(x)} \\
&\geq (\log e) \sum_{x \in \mathcal{S}_p} p(x) \left(1 - \frac{q(x)}{p(x)} \right) \\
&= (\log c) \left[\sum_{x \in \mathcal{S}_p} p(x) - \sum_{x \in \mathcal{S}_p} q(x) \right] \\
&= (\log e) \left[\sum_{x \in \mathcal{S}_p} p(x) - \sum_{x \in \mathcal{S}_q} q(x) \right] \\
&= 0.
\end{aligned}
$$

For equality to hold, further require

$$
\frac{p(x)}{q(x)} = 1 \text{ or } p(x) = q(x) \quad \text{for all } x \in \mathcal{S}_p
$$

> **Theorem 2.32 (Log-Sum Inequality)** For positive numbers $a_1, a_2, \cdots$ and nonnegative numbers $b_1, b_2, \cdots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$
> $$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}$$
> with the convention that $\log \frac{a_i}{0} = \infty$. Moreover, equality holds if and only if $\frac{a_i}{b_i} = $ constant for all i
>
> **Example:**
> $$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}$$

Proof. Let $a'_i = a_i / \sum_j a_j$ and $b'_i = b_i / \sum_j b_j$. Using the divergence inequality, we have

$$0 \leq \sum_i a'_i \log \frac{a'_i}{b'_i}$$

$$= \sum_i \frac{a_i}{\sum_j a_j} \log \frac{a_i / \sum_j a_j}{b_i / \sum_j b_j}$$

$$= \frac{1}{\sum_j a_j} \left[\sum_i a_i \log \frac{a_i / \sum_j a_j}{b_i / \sum_j b_j} \right]$$

$$= \frac{1}{\sum_j a_j} \left[\sum_i a_i \log \frac{a_i}{b_i} - \sum_i a_i \log \frac{\sum_j a_j}{\sum_j b_j} \right]$$

$$= \frac{1}{\sum_j a_j} \left[\sum_i a_i \log \frac{a_i}{b_i} - \left(\sum_i a_i \right) \log \frac{\sum_j a_j}{\sum_j b_j} \right]$$

Remark Divergence Inequality vs Log-Sum Inequality are equivalent.

Theorem 2.33 (Pinsker's Inequality)

$$D(p\|q) \geq \frac{1}{2\ln 2} V^2(p, q)$$

- If $D(p\|q)$ or $D(q\|p)$ is small, then so is $V(p, q) = V(q, p)$

- For a sequence of probability distributions q_k, as $k \to \infty$, if $D(p\|q_k) \to 0$ or $D(q_k\|p) \to 0$, then $V(p, q_k) = V(q_k, p) \to 0$

- That is, "convergence in divergence" is a stronger notion than "convergence in variational distance."

1.6   Basic Inequalities

Theorem 2.34   For random variables X, Y, and Z,

$$I(X;Y \mid Z) \geq 0$$

with equality if and only if X and Y are independent when conditioning on Z.

Corollary   All Shannon's information measures are nonnegative, because they are all special cases of conditional mutual information.

Proof

$$I(X;Y \mid Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)}$$

$$= \sum_{z} \sum_{x,y} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)}$$

$$= \sum_{z} \sum_{x,y} p(z)p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)}$$

$$= \sum_{z} p(z) \sum_{x,y} p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)}$$

$$= \sum_{z} p(z) D\left(p_{XY|z} \| p_{X|z} p_{Y|z}\right)$$

Proposition 2.35 $H(X) = 0$ if and only if X is deterministic.

Proposition 2.36 $H(Y \mid X) = 0$ if and only if Y is a function of X

$$H(Y \mid X) = \sum_{x} p(x) H(Y \mid X = x)$$

Proposition 2.37 $I(X;Y) = 0$ if and only if X and Y are independent.

1.7 Some Useful Information Inequalities

Theorem 2.38 (Conditioning Does Not Increase Entropy)

$$H(Y \mid X) \leq H(Y)$$

with equality if and only if X and Y are independent.

Proof
$$H(Y \mid X) = H(Y) - I(X;Y) \leq H(Y)$$

with equality if and only if $I(X;Y) = 0$, or X and Y are independent.

- Similarly, $H(Y \mid X, Z) \leq H(Y \mid Z)$

- Warning: $I(X;Y \mid Z) \leq I(X;Y)$ does not hold in general.

Theorem 2.39 (Independence Bound for Entropy)

$$H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if $X_i, i = 1, 2, \cdots, n$ are mutually independent.

Proof By the chain rule for entropy,

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

The inequality is tight iff it is tight for each i, i.e., $H(X_i \mid X_1, \cdots, X_{i-1}) = H(X_i)$

$$\begin{aligned}
p(x_1, x_2, \cdots, x_n) &= p(x_1, x_2, \cdots, x_{n-1}) p(x_n) \\
&= p(x_1, x_2, \cdots, x_{n-2}) p(x_{n-1}) p(x_n) \\
&= p(x_1) p(x_2) \cdots p(x_n)
\end{aligned}$$

Theorem 2.40

$$I(X;Y, Z) \geq I(X;Y)$$

with equality if and only if $X \rightarrow Y \rightarrow Z$ forms a Markov chain.

Proof By the chain rule for mutual information, we have

$$I(X;Y, Z) = I(X;Y) + I(X;Z \mid Y) \geq I(X;Y)$$

The above inequality is tight iff $I(X;Z \mid Y) = 0$ (or $X \rightarrow Y \rightarrow Z$ forms a Markov chain.)

Lemma 2.41 If $X \to Y \to Z$ forms a Markov chain, then

$$I(X;Z) \leq I(X;Y)$$

$$I(X;Z) \leq I(Y;Z)$$

Corollary If $X \to Y \to Z$, then

$$H(X \mid Z) \geq H(X \mid Y)$$

Proof Corollary 1. Assume $X \to Y \to Z$, i.e., $X \perp Z \mid Y$. Then

$$I(X;Z \mid Y) = 0$$

2. Consider

$$
\begin{aligned}
I(X;Z) &\overset{a}{=} I(X;Y,Z) - I(X;Y \mid Z) \\
&\leq I(X;Y,Z) \\
&\overset{b}{=} I(X;Y) + I(X;Z \mid Y) \\
&= I(X;Y)
\end{aligned}
$$

a) Chain rule for mutual information:

$$
\begin{aligned}
I(X;Y,Z) &= I(X;Z) + I(X;Y \mid Z) \\
\Rightarrow \quad I(X;Z) &= I(X;Y,Z) - I(X;Y \mid Z)
\end{aligned}
$$

b) Chain rule for mutual information
3. since $X \to Y \to Z$ is equivalent to $Z \to Y \to X$ we also have proved (2) by symmetry.

Proof

$$
\begin{aligned}
H(X \mid Z) &= H(X) - I(X;Z) \\
&\geq H(X) - I(X;Y) \\
&= H(X \mid Y)
\end{aligned}
$$

Remark Suppose Y is an observation of X. Then further processing of Y can only increase the uncertainty about X on the average.

Theorem 2.42 (Data Processing Theorem) If $U \to X \to Y \to V$ forms a Markov chain, then

$$I(U;V) \leq I(X;Y)$$

Proof For two subchains

$$U \to X \to Y$$
$$U \to Y \to V$$

By applying Lemma 2.41

$$I(U;V) \leq I(U;Y) \leq I(X;Y)$$

1.8 Fano's Inequality

Theorem 2.43 For any random variable X,

$$H(X) \leq \log |\mathcal{X}|$$

where $|\mathcal{X}|$ denotes the size of the alphabet $\mathcal{X}$. This upper bound is tight if and only if X is distributed uniformly on $\mathcal{X}$.

Remark For a random variable X, if the alphabet is finite, then

$$H(X) \leq \log |\mathcal{X}| < \infty$$

i.e., $H(X)$ is finite.

Proof Let u be the uniform distribution on $\mathcal{X}$, i.e.,

$$u(x) = \frac{1}{|\mathcal{X}|} \quad \text{for all } x \in \mathcal{X}$$

Then

$$
\begin{aligned}
\log |\mathcal{X}| - H(X) &= -\sum_{x \in \mathcal{S}_X} p(x) \log \frac{1}{|\mathcal{X}|} + \sum_{x \in \mathcal{S}_X} p(x) \log p(x) \\
&= -\sum_{x \in \mathcal{S}_X} p(x) \log u(x) + \sum_{x \in \mathcal{S}_X} p(x) \log p(x) \\
&= \sum_{x \in \mathcal{S}_X} p(x) \log \frac{p(x)}{u(x)} \\
&= D(p\|u) \geq 0
\end{aligned}
$$

Theorem 2.47 (Fano's Inequality) Let X and $\hat{X}$ be random variables taking values in the same alphabet $\mathcal{X}$. Then
$$H(X \mid \hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1)$$
where $P_e = \Pr\{X \neq \hat{X}\}$ and h_b is the binary entropy function.

Corollary 2.48 $H(X \mid \hat{X}) < 1 + P_e \log |\mathcal{X}|$

1.9 Entropy Rate of a Stationary Source

Definition 2.54 The entropy rate of an information source $\{X_k\}$ is defined as

$$H_X = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

when the limit exists.

2 The I-Measure

$$
\begin{aligned}
H/I \quad &\leftrightarrow \mu^* \\
, \quad &\leftrightarrow \cup \\
; \quad &\leftrightarrow \cap \\
| \quad &\leftrightarrow - (A - B = A \cap B^c)
\end{aligned}
$$

1. Examples

$$
\begin{aligned}
H\left(X_1 \mid X_2\right) &= \mu^*\left(\tilde{X}_1 - \tilde{X}_2\right) \\
H\left(X_2 \mid X_1\right) &= \mu^*\left(\tilde{X}_2 - \tilde{X}_1\right) \\
I\left(X_1; X_2\right) &= \mu^*\left(\bar{X}_1 \cap \tilde{X}_2\right)
\end{aligned}
$$

2. Inclusion-Exclusion formulation in set-theory

$$
\mu^*\left(\tilde{X}_1 \cup \tilde{X}_2\right) = \mu^*\left(\tilde{X}_1\right) + \mu^*\left(\tilde{X}_2\right) - \mu^*\left(\tilde{X}_1 \cap \tilde{X}_2\right)
$$

corresponds to

$$
H\left(X_1, X_2\right) = H\left(X_1\right) + H\left(X_2\right) - I\left(X_1; X_2\right)
$$

in information theory.

2.1 Preliminaries

Definition 3.1 The field $\mathcal{F}_n$ generated by sets $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_n$ is the collection of sets which can be obtained by any sequence of usual set operations (union, intersection, complement, and difference) on $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_n$

Definition 3.2 The atoms of $\mathcal{F}_n$ are sets of the form $\cap_{i=1}^{n} Y_i$, where Y_i is either $\bar{X}_i$ or $\bar{X}_i^c$, the complement of $\bar{X}_i$

Example 3.3

- The sets $\tilde{X}_1$ and $\tilde{X}_2$ generate the field $\mathcal{F}_2$

- There are 4 atoms in $\mathcal{F}_2$:
$$
\tilde{X}_1 \cap \tilde{X}_2, \quad \tilde{X}_1^c \cap \tilde{X}_2, \quad \tilde{X}_1 \cap \tilde{X}_2^c, \quad \tilde{X}_1^c \cap \tilde{X}_2^c
$$

- There are a total of $2^4 = 16$ sets in $\mathcal{F}_2$, formed by the unions of the above 4 atoms.

Definition 3.4 A real function μ defined on $\mathcal{F}_n$ is called a signed measure if it is set-additive, i.e., for disjoint A and B in $\mathcal{F}_n$
$$
\mu(A \cup B) = \mu(A) + \mu(B)
$$

Remark: A signed measure can take positive or negative values. If a signed measure takes only positive values, it is simply called a measure.

2.2 Construction of the I-Measure μ^*

Notations For nonempty subset G of $\mathcal{N}_n$: $\boldsymbol{X}_G = (X_i, i \in G)$ $\qquad \tilde{X}_G = \cup_{i \in G} \tilde{X}_i$

Theorem 3.6 Let
$$\mathcal{B} = \left\{ \tilde{X}_G : G \text{ is a nonempty subset of } \mathcal{N}_n \right\}$$
Then a signed measure μ on $\mathcal{F}_n$ is completely specified by $\{\mu(B), B \in \mathcal{B}\}$, which can be any set of real numbers.

Remark We have seen that a signed measure μ on $\mathcal{F}_n$ is completely specified by $\{\mu(A), A \in \mathcal{A}\}$, the set of values of μ on the nonempty atoms. This theorem says that μ can instead be specified by $\{\mu(B), B \in \mathcal{B}\}$, the set of values of μ on the unions.

The Inclusion-Exclusion Formula

$$\mu \left(\bigcup_{k=1}^m A_k \right) = \sum_{1 \le i \le m} \mu(A_i) - \sum_{1 \le i < j \le m} \mu(A_i \cap A_j) + \cdots$$
$$+ (-1)^{m+1} \mu(A_1 \cap A_2 \cap \cdots \cap A_m)$$

Theorem 3.19 (Variation of the Inclusion-Exclusion Formula)

$$\mu \left(\bigcap_{k=1}^m A_k - B \right) = \sum_{1 \le i \le m} \mu(A_i - B) - \sum_{1 \le i < j \le m} \mu(A_i \cup A_j - B) + \cdots$$
$$+ (-1)^{m+1} \mu(A_1 \cup A_2 \cup \cdots \cup A_m - B)$$

Proof of Lemma 3.7

$$\begin{aligned}
\mu(A \cap B - C) &= \mu(A - C) + \mu(B - C) - \mu(A \cup B - C) \\
&= (\mu(A \cup C) - \mu(C)) + (\mu(B \cup C) - \mu(C)) - (\mu(A \cup B \cup C) - \mu(C)) \\
&= \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C)
\end{aligned}$$

Proof of Lemma 3.8

$$\begin{aligned}
I(X; Y \mid Z) &= H(X \mid Z) - H(X \mid Y, Z) \\
&= H(X, Z) - H(Z) - (H(X, Y, Z) - H(Y, Z)) \\
&= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z)
\end{aligned}$$

Construction of the I -Measure μ^* on $\mathcal{F}_n$ Define μ^* by setting

$$\mu^* \left(\tilde{X}_G \right) = H(X_G)$$

for all nonempty subsets G of $\mathcal{N}_n$ - That is, the value of μ^* on the union of a collection G of set variables is equal to the joint entropy of the collection G of random variables.

Theorem 3.9 μ^* is the unique signed measure on $\mathcal{F}_n$ which is consistent with all Shannon's information measures.

Implications Can formally regard Shannon's information measures for n r.v.'s as the unique signed measure μ^* defined on $\mathcal{F}_n$ Can employ set-theoretic tools to manipulate expressions of Shannon's information measures.

2.3 μ^* can be Negative

For $n = 3$, the values of μ^* on the nonempty atoms of $\mathcal{F}_3$ all correspond to Shannon's information measures, except

$$\mu^* \left(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \right) = I \left(X_1; X_2; X_3 \right)$$

We will show that it is possible to construct r.v.'s X_1, X_2, and X_3 such that $\mu^* \left(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \right) < 0$

Example 3.10 Let X_1 and X_2 be independent binary random variables with uniform distribution, i.e.,

$$\Pr \{ X_i = 0 \} = \Pr \{ X_i = 1 \} = 0.5, \qquad i = 1, 2$$

Let

$$X_3 = (X_1 + X_2) \bmod 2$$

It is easy to check that X_3 also has a uniform distribution. Thus, $H(X_i) = 1$ for $i = 1, 2, 3$. It is also easy to check that X_1, X_2, and X_3 are pairwise independent. Therefore,

$$H(X_i, X_j) = 2 \qquad I(X_i; X_j) = 0$$

for $1 \le i < j \le 3$ We see from (1) that X_3 is a function of X_1 and X_2, so that

$$H(X_3 \mid X_1, X_2) = 0$$

Then by the chain rule for entropy, we have

$$H(X_1, X_2, X_3) = H(X_1, X_2) + H(X_3 \mid X_1, X_2) = 2 + 0 = 2$$

Now for distinct $1 \le i, j, k \le 3$

$$
\begin{aligned}
I(X_i; X_j \mid X_k) &= H(X_i, X_k) + H(X_j, X_k) - H(X_1, X_2, X_3) - H(X_k) \\
&= 2 + 2 - 2 - 1 = 1
\end{aligned}
$$

It then follows that

$$
\begin{aligned}
\mu^* \left(\tilde{X}_1 \cap \bar{X}_2 \cap \tilde{X}_3 \right) &= \mu^* \left(\bar{X}_1 \cap \bar{X}_2 \right) - \mu^* \left(\bar{X}_1 \cap \tilde{X}_2 - \bar{X}_3 \right) \\
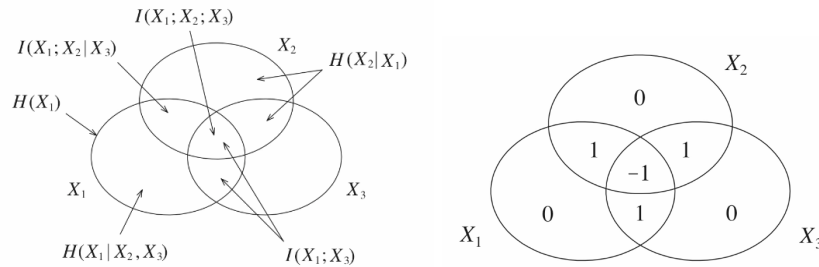&= I(X_1; X_2) - I(X_1; X_2 \mid X_3) \\
&= 0 - 1 < 0
\end{aligned}
$$

Figure 6: μ^* can be Negative.

> **Theorem 3.11** If there is no constraint on $X_1, X_2, \cdots, X_n$, then μ^* can take any set of nonnegative values on the nonempty atoms of $\mathcal{F}_n$.

Evidently, we can take $\mu^*(A) = H(Y_A)$ for all $A \in \mathcal{A}$. By the uniqueness of μ^* (Theorem 3.9), this is also the only possibility for μ^*

2.4 Information Diagrams for Markov Chains

- If $X_1 \to X_2 \to \cdots \to X_n$ form a Markov chain, then the structure of μ^* is much simpler and hence the information diagram can be simplified.

- For $n = 3$, $X_1 \to X_2 \to X_3$ iff $I(X_1; X_3 \mid X_2) = 0$, or $\mu^*\left(\tilde{X}_1 \cap \tilde{X}_3 - \tilde{X}_2\right) = 0$

- So the atom $\tilde{X}_1 \cap \tilde{X}_3 - \tilde{X}_2$ can be suppressed in the information diagram.
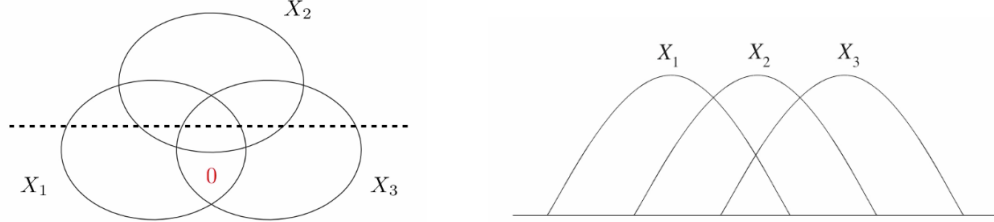

Figure 7: Suppressed information diagram.

Illustration: μ^* for $X_1 \to X_2 \to X_3$
In this information diagram,

$$I(X_1; X_3 \mid X_2) = \mu^*\left(\tilde{X}_1 \cap \tilde{X}_3 - \tilde{X}_2\right)$$
$$= \mu^*(\emptyset)$$
$$= 0$$

Also,

$$\mu^*\left(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3\right) = \mu^*\left(\tilde{X}_1 \cap \tilde{X}_3\right)$$
$$= I(X_1; X_3)$$
$$\geq 0$$

Since the values of μ^* on all the remaining atoms correspond to Shannon's information measures and hence are nonnegative, we conclude that μ^* is a measure.
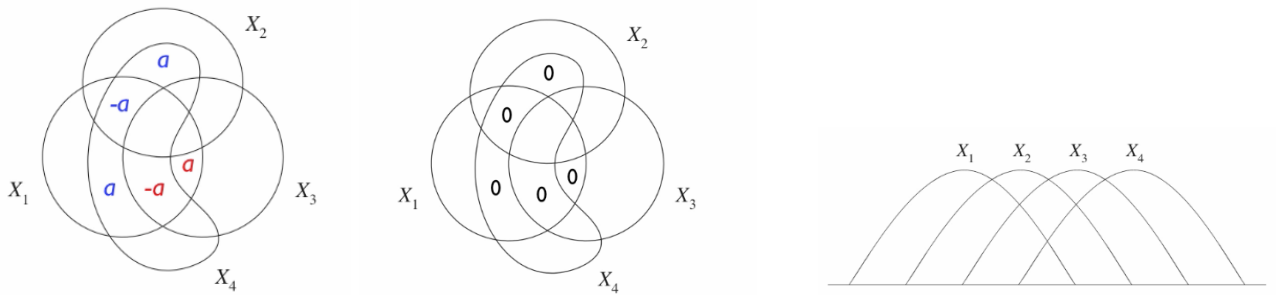

Figure 8: Suppressed information diagram.

Illustration: Structure of μ^* for $X_1 \to X_2 \to X_3 \to X_4$
1. The Markov subchain $X_1 \to X_2 \to X_3$ implies $0 = I(X_1; X_3 \mid X_2) = I(X_1; X_3; X_4 \mid X_2) + I(X_1; X_3 \mid X_2, X_4)$
Let $I(X_1; X_3 \mid X_2, X_4) = a \geq 0$. Then

$$I(X_1; X_3; X_4 \mid X_2) = -a$$

2. The Markov subchain $X_1 \to X_2 \to X_4$ implies $0 = I(X_1; X_4 \mid X_2) = I(X_1; X_3; X_4 \mid X_2) + I(X_1; X_4 \mid X_2, X_3)$
since $I(X_1; X_3; X_4 \mid X_2) = -a$

$$I(X_1; X_4 \mid X_2, X_3) = a$$

17

3. The Markov subchain $X_1 \to X_3 \to X_4$ implies $0 = I(X_1; X_4 \mid X_3) = I(X_1; X_2; X_4 \mid X_3) + I(X_1; X_4 \mid X_2, X_3)$ since $I(X_1; X_4 \mid X_2, X_3) = a$

$$I(X_1; X_2; X_4 \mid X_3) = -a$$

4. The Markov subchain $X_2 \to X_3 \to X_4$ implies $0 = I(X_2; X_4 \mid X_3) = I(X_1; X_2; X_4 \mid X_3) + I(X_2; X_4 \mid X_1, X_3)$ since $I(X_1; X_2; X_4 \mid X_3) = -a$

$$I(X_2; X_4 \mid X_1, X_3) = a$$

5. The Markov subchain $(X_1, X_2) \to X_3 \to X_4$ implies

$$0 = I(X_1, X_2; X_4 \mid X_3) = I(X_1; X_4 \mid X_2, X_3) + I(X_1; X_2; X_4 \mid X_3) + I(X_2; X_4 \mid X_1, X_3)$$

Then

$$0 = a - a + a = a$$

Therefore $a = 0$, and so μ^* vanishes on the corresponding 5 atoms as shown in the information diagram.

2.5 Examples of Applications

> **Example 3.12 (Concavity of Entropy)**   Let $X_1 \sim p_1(x)$ and $X_2 \sim p_2(x)$, and
> $$X \sim p(x) = \lambda p_1(x) + \bar{\lambda} p_2(x)$$
> where $0 \le \lambda \le 1$ and $\bar{\lambda} = 1 - \lambda$. Show that $H(X) \ge \lambda H\left(X_1\right) + \bar{\lambda} H\left(X_2\right)$

Proof.
$$
\begin{aligned}
H(X) \ge & H(X \mid Z) \\
= & \Pr\{Z = 1\} H(X \mid Z = 1) + \Pr\{Z = 2\} H(X \mid Z = 2) \\
= & \lambda H\left(X_1\right) + \bar{\lambda} H\left(X_2\right)
\end{aligned}
$$

This shows that $H(X)$ is a concave functional of $p(x)$

Interpretation   The entropy of a mixture of distributions is at least the mixture of the corresponding entropies.

> **Example 3.13/3.14 (Convexity/Concavity of Mutual Information)**   Let
> $$(X, Y) \sim p(x, y) = p(x) p(y \mid x)$$
> Show that for fixed $p(x)$, $I(X; Y)$ is a convex functional of $p(y \mid x)$.
> Show that for fixed $p(y \mid x)$, $I(X; Y)$ is a concave functional of $p(x)$

Proof 3.13
$$
\begin{aligned}
I(X; Y) = & I(X; Y \mid Z) + I(X; Y; Z) \\
\le & I(X; Y \mid Z) \\
= & \Pr\{Z = 1\} I(X; Y \mid Z = 1) + \Pr\{Z = 2\} I(X; Y \mid Z = 2) \\
= & \lambda I\left(p(x), p_1(y \mid x)\right) + \bar{\lambda} I\left(p(x), p_2(y \mid x)\right)
\end{aligned}
$$

Interpretation   For a fixed input distribution $p(x)$, the mutual information between the input and the output of the system as shown, which is obtained by mixing 2 channels $p_1(y \mid x)$ and $p_2(y \mid x)$, is at most the mixture of the 2 mutual informations corresponding to $p_1(y \mid x)$ and $p_2(y \mid x)$, respectively

Proof 3.14
$$
\begin{aligned}
I(X; Y) \ge & I(X; Y \mid Z) \\
= & \Pr\{Z = 1\} I(X; Y \mid Z = 1) + \Pr\{Z = 2\} I(X; Y \mid Z = 2) \\
= & \lambda I\left(p_1(x), p(y \mid x)\right) + \bar{\lambda} I\left(p_2(x), p(y \mid x)\right)
\end{aligned}
$$

This shows that for fixed $p(y \mid x)$, $I(X; Y)$ is a concave functional of $p(x)$

Interpretation   For a fixed channel, by mixing the input distribution, the mutual information is at least equal to the mixture of the corresponding mutual informations.
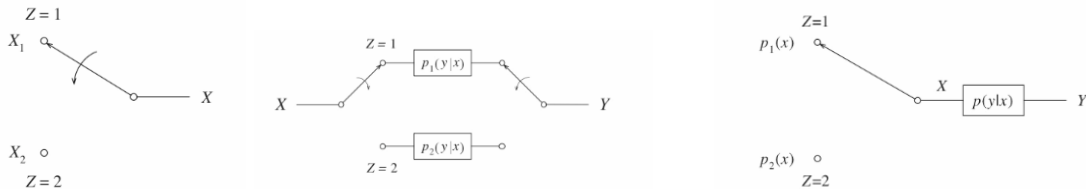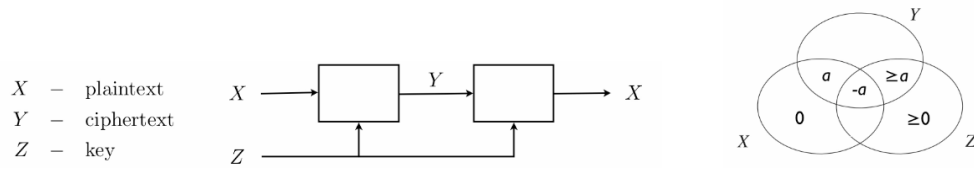

Figure 9: Systems

Figure 10: Shannon's Perfect Secrecy Theorem

Shannon's Perfect Secrecy Theorem

- Perfect Secrecy: $I(X;Y) = 0$

- Decipherability: $H(X \mid Y, Z) = 0$

- These implies $H(Z) \geq H(X)$, i.e., the length of the key is at least the same as the length of the plaintext.

- Shannon (1949) gave a combinatorial proof. Can readily be proved by an information diagram.

Example 3.15 (Imperfect Secrecy Theorem)   Let X be the plain text, Y be the cipher text, and Z be the key in a secret key cryptosystem. since X can be recovered from Y and Z, we have

$$H(X \mid Y, Z) = 0$$

Show that this constraint implies

$$I(X;Y) \geq H(X) - H(Z)$$

Remark $I(X;Y)$ measures the "leakage of information." When $I(X;Y) = 0$, it reduces Shannon's perfect secrecy theorem.

Example 3.17 (Data Processing Theorem)   If $X \to Y \to Z \to T$, then

- $I(X;T) \leq I(Y;Z)$

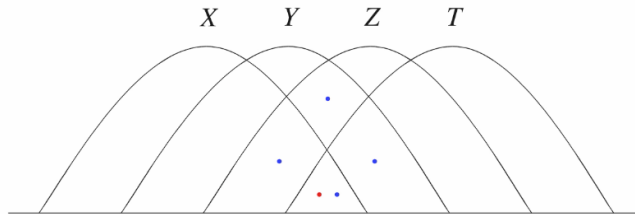- $I(Y;Z) = I(X;T) + I(X;Z \mid T) + I(Y;T \mid X) + I(Y;Z \mid X,T)$

Figure 11: Shannon's Perfect Secrecy Theorem

3   Zero-Error Data Compression

- Why $H(X)$ measures the amount of information in X?

- A first look at data compression: Prefix codes

- How to construct optimal prefix codes - Huffman codes?

3.1   The Entropy Bound

Definition 4.1   A D-ary source code $\mathcal{C}$ for a source random variable X is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$, the set of all finite length sequences of symbols taken from a D -ary code alphabet.

Definition 4.2   A code $\mathcal{C}$ is uniquely decodable if for any finite source sequence, the sequence of code symbols corresponding to this source sequence is different from the sequence of code symbols corresponding to any other (finite) source sequence.

Example 4.3   Let $\mathcal{X} = \{A, B, C, D\}$. Consider the code $\mathcal{C}$ defined by

x	$\mathcal{C}(x)$
A	0
B	1
C	01
D	10

$$
\begin{aligned}
AAD & \rightarrow 0010 \\
ACA & \rightarrow 0010 \\
AABA & \rightarrow 0010
\end{aligned}
$$

Therefore, $\mathcal{C}$ not uniquely decodable.

Theorem 4.4 (Kraft Inequality)   Let $\mathcal{C}$ be a D -ary source code, and let $l_1, l_2, \cdots, l_m$ be the lengths of the codewords. If $\mathcal{C}$ is uniquely decodable, then

$$\sum_{k=1}^{m} D^{-l_k} \leq 1$$

Proof   1. Without loss of generality, assume

$$l_1 \leq l_2 \leq \cdots \leq l_m$$

2. Let N be an arbitrary positive integer, and consider

$$\left(\sum_{k=1}^{m} D^{-l_k}\right)^N = \sum_{k_1=1}^{m} \sum_{k_2=1}^{m} \cdots \sum_{k_N=1}^{m} D^{-\left(l_{k_1} + l_{k_2} + \cdots + l_{k_N}\right)}$$

3. By collecting terms of the same degree on the RHS, we write

$$\left(\sum_{k=1}^{m} D^{-l}k\right)^N = \sum_{i=1}^{Nlm} A_i D^{-i}$$

where A_i is the coefficient of D^{-i} on the LHS.

4. Now observe that A_i gives the total number of sequences of N codewords with a total length of i code symbols. Since the code is uniquely decodable, these code sequences must be distinct, and therefore

$$A_i \leq D^i$$

21

because there are D^i distinct sequences of i code symbols. 5. Substitute and we have

$$\left(\sum_{k=1}^m D^{-l_k} \right)^N \leq \sum_{i=1}^{Nlm} D^i D^{-i} = \sum_{i=1}^{Nl_m} 1 = Nl_m$$

or

$$\sum_{k=1}^m D^{-l}k \leq (Nl_m)^{1/N}$$

since this inequality holds for any N, upon letting $N \to \infty$, we obtain (1), completing the proof.

Theorem 4.6 (Entropy Bound)   Let $\mathcal{C}$ be a D-ary uniquely decodable code for a source random variable X with entropy $H_D(X)$. Then the expected length of $\mathcal{C}$ is lower bounded by $H_D(X)$, i.e.

$$L \geq H_D(X)$$

This lower bound is tight if and only if $l_i = -\log_D p_i$ for all i

---

Proof   1. since $\mathcal{C}$ is uniquely decodable, the lengths of its codewords satisfy the Kraft inequality. Write

$$L = \sum_i p_i l_i = \sum_i p_i \log_D D^{l_i}$$

and recall that

$$H_D(X) = -\sum_i p_i \log_D p_i$$

Then

$$L - H_D(X) = \sum_i p_i \left(\log_D p_i + \log_D D^l i \right)$$

$$= \sum_i p_i \log_D \left(p_i D^l i \right)$$

$$= (\ln D)^{-1} \sum_i p_i \ln \left(p_i D^l i \right)$$

$$\geq (\ln D)^{-1} \sum_i p_i \left(1 - \frac{1}{p_i D^l i} \right) \qquad \text{(fundamental inequality:} \quad \ln a \geq 1 - \frac{1}{a} \quad (a = p_i D^l i))$$

$$= (\ln D)^{-1} \sum_i \left(p_i - D^{-l} i \right)$$

$$\geq (\ln D)^{-1} \left[1 - \sum_i D^{-l} i \right]$$

$$= 0$$

The inequality bounds hold tight if and only if $p_i D^{l_i} = 1$, or $l_i = -\log_D p_i$ for all i. If this holds, we have

$$\sum_i D^{-l_i} = \sum_i D^{\log_D p_i} = \sum_i p_i = 1$$

Corollary 4.7 (Theorem 2.43) $H(X) \leq \log |\mathcal{X}|$.

Proof Let $\mathcal{X} = \{0, 1, \cdots, |\mathcal{X}| - 1\}$. Let $\mathcal{C}$ be the identity code, i.e.,

x	0	1	$\cdots$	$|\mathcal{X}| - 1$
$\mathcal{C}(x)$	0	1	$\cdots$	$|\mathcal{X}| - 1$

Evidently, $\mathcal{C}$ is an $|\mathcal{X}|$-ary uniquely decodable code, with expected length equals 1. By the entropy bound, we have

$$1 = L \geq H_{|\mathcal{X}|}(X)$$

Leaving the base unspecified, we have

$$H(X) \leq \log |\mathcal{X}|$$

Definition 4.8 The redundancy R of a D-ary uniquely decodable code is the difference between the expected length of the code and the entropy of the source. By the entropy bound,

$$R = L - H_D(X) \geq 0$$

3.2 Prefix Codes

Definition 4.9 A code is called a prefix-free code if no codeword is a prefix of any other codeword. For brevity, a prefix-free code will be referred to as a prefix code.

Code Tree for Prefix Code

- A D-ary tree is a graphical representation of a collection of finite sequences of D -ary symbols.

- A node is either an internal node or a leaf.

- The tree representation of a prefix code is called a code tree.

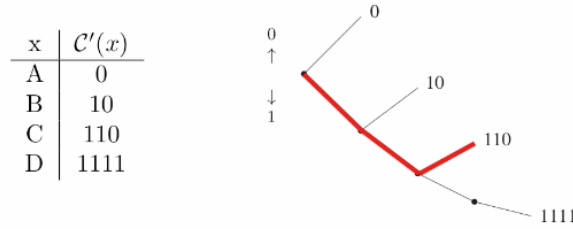$$1011011110110\cdots \rightarrow 10, 110, 1111, 0, \underline{110}, \cdots$$

x	$\mathcal{C}'(x)$
A	0
B	10
C	110
D	1111

Figure 12: Instantaneous Decoding

> **Theorem 4.11** There exists a D-ary prefix code with codeword lengths $l_1\ l_2, \cdots, l_m$ if and only if the Kraft inequality
> $$\sum_{k=1}^{m} D^{-l_k} \leq 1$$
> is satisfied.

Proof Direct part follows because a prefix code is uniquely decodable and hence satisfies Kraft's inequality.

(Converse) 1. We need to prove the existence of a D-ary prefix code with codeword lengths $l_1, l_2, \cdots, l_m$ if these lengths satisfy the Kraft inequality. Without loss of generality, assume that

$$l_1 \leq l_2 \leq \cdots \leq l_m$$

2. Consider all the D-ary sequences of lengths less than or equal to l_m and regard them as the nodes of the full D-ary tree of depth l_m. We will refer to a sequence of length l as a node of order l.

3. There are $D^{l_1} > 1$ (since $l_1 \geq 1$) nodes of order l_1 which can be chosen as the first codeword. Thus choosing the first codeword is always possible.

4. Assume that the first i codewords have been chosen successfully, where $1 \leq i \leq m - 1$, and we want to choose a node of order l_{i+1} as the $(i+1)$ st codeword such that it is not prefixed by any of the previously chosen codewords.

5. Since all the previously chosen codewords are not prefixes of each other, their descendants of order l_{i+1} do not overlap. The $(i+1)$ st node to be chosen cannot be a descendant of any of the previously chosen codewords. Therefore, the number of nodes which can be chosen as the $(i+1)$ st codeword is

$$D^{l_i+1} - D^{l_i+1} - l_1 - D^{l_i+1} - l_2 - \ldots - D^{l_i+1} - l_i$$

24

6. If $l_1, l_2, \cdots, l_m$ satisfy the Kraft inequality, we have

$$D^{-l_1} + \cdots + D^{-l_i} + D^{-l_{i+1}} \leq 1$$

7. Multiplying by $D^l i + 1$, we have

$$D^{l_{i+1} - l_1} + \cdots + D^{l_i + 1} - l_i + D^{l_{i+1} - l_{i+1}} \leq D^{l_i + 1}$$

Or

$$D^{l_{i+1}} - D^{l_{i+1} - l_1} - \cdots - D^{l_{i+1} - l_i} \geq 1$$

Thus we have shown by induction the existence of a prefix code with codeword lengths $l_1, l_2, \cdots, l_m$ completing the proof.

Definition (D-adic distribution)

- $p_i = D^{-t_i}$ for all i, where t_i is integer

- dyadic when $D = 2$

> **Corollary 4.12** There exists a D-ary prefix code which achieves the entropy bound for a distribution $\{p_i\}$ if and only if $\{p_i\}$ is D-adic.

3.3 Huffman Codes

A simple construction of optimal prefix codes.

- Binary Case: Keep merging the two smallest probability masses until one probability mass (i.e., 1) is left.

- D-ary Case: Insert zero probability masses until there are $D + k(D - 1)$ masses (if necessary). Keep merging the D smallest probability masses until one probability mass (i.e., 1) is left.

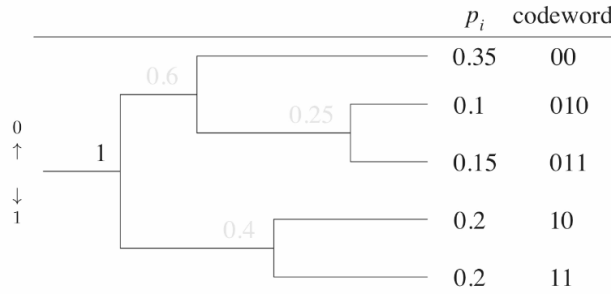- In general there can be more than one Huffman code.

Figure 13: Instantaneous Decoding

Optimality of Huffman Codes

- Without loss of generality, assume $p_1 \geq p_2 \geq \cdots \geq p_m$

- Denote the codeword assigned to p_i by c_i, and its length by l_i

Theorem 4.17 The Huffman procedure produces an optimal prefix code.

Lemma 4.15 In an optimal code, shorter codewords are assigned to larger probabilities, i.e.,

$$l_1 \leq l_2 \leq \cdots \leq l_m$$

Lemma 4.16 There exists an optimal code in which the codewords assigned to the two smallest probabilities are siblings, i.e., the two codewords have the same length and they differ only in the last symbol.

Proof 1. Consider a probability distribution

$$\{p_1, \cdots, p_i, \cdots, p_j, \cdots p_m\}$$

such that $p_i > p_j$. Assume that in a particular code, the codewords c_i and c_j are such that $l_i > l_j$, i.e., a shorter codeword is assigned to a smaller probability. 2. Intuitively, by exchanging c_i and c_j, the expected length of the code should be improved. 3. Specifically, let

$$L = \sum_k p_k l_k = \sum_{k \neq i,j} p_k l_k + (p_i l_i + p_j l_j)$$

be the expected length of the code, and

$$L' = \sum_{k \neq i,j} p_k l_k + (p_i l_j + p_j l_i)$$

be the expected length of the code obtained by exchanging c_i and c_j
    4. Comparing L' and L, we see that

$$
\begin{aligned}
L' - L &= (p_i l_j + p_j l_i) - (p_i l_i + p_j l_j) \\
&= (p_i l_j - p_i l_i) - (p_j l_j - p_j l_i) \\
&= p_i (l_j - l_i) - p_j (l_j - l_i) \\
&= (p_i - p_j)(l_j - l_i)
\end{aligned}
$$

This is negative because $p_i > p_j$ and $l_i > l_j$. Therefore, $L' < L$ 5. Since the original code can be improved, it is not an optimal code. 6. Therefore, for an optimal code, shorter codewords are assigned to larger probabilities. The lemma is proved.

4 Strong Typicality

4.1 Strong AEP

Definition 6.1 The strongly typical set $T^n_{[X]\delta}$ with respect to $p(x)$ is the set of sequences $x = (x_1, x_2, \cdots, x_n) \in X_n$ such that $N(x; \mathbf{x}) = 0$ for $x \notin \mathcal{S}_X$ and

$$\sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| \leq \delta \tag{1}$$

where $N(x; \mathbf{x})$ is the number of occurrences of x in the sequence $\mathbf{x}$ and δ is an arbitrarily small positive real number. The sequences in $T^n_{[X]\delta}$ are called strongly δ-typical sequences.

Theorem 6.2 (Strong AEP) There exists $\eta > 0$ such that $\eta \to 0$ as $\delta \to 0$, and the following hold:
1) If $\mathbf{x} \in T^n_{[X]\delta}$, then

$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}$$

2) For n sufficiently large,

$$\Pr \left\{ \mathbf{X} \in T^n_{[X]\delta} \right\} > 1 - \delta$$

3) For n sufficiently large,

$$(1 - \delta) 2^{n(H(X)-\eta)} \leq \left| T^n_{[X]\delta} \right| \leq 2^{n(H(X)+\eta)}$$

Proof.

1. To prove Property 1, for $\mathbf{x} \in T^n_{[X]\delta}$, we write

$$p(\mathbf{x}) = \prod_x p(x)^{N(x;\mathbf{x})}$$

$$\log p(\mathbf{x}) = \sum_x N(x; \mathbf{x}) \log p(x)$$

$$= \sum_x (N(x; \mathbf{x}) - np(x) + np(x)) \log p(x)$$

$$= n \sum_x p(x) \log p(x) - n \sum_x \left(\frac{1}{n} N(x; \mathbf{x}) - p(x) \right) (-\log p(x))$$

$$= -n \left[H(X) + \sum_x \left(\frac{1}{n} N(x; \mathbf{x}) - p(x) \right) (-\log p(x)) \right]$$

Since $\mathbf{x} \in T^n_{[X]\delta}$

$$\sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| \leq \delta$$

which implies

$$\left| \sum_x \left(\frac{1}{n} N(x; \mathbf{x}) - p(x) \right) (-\log p(x)) \right| \leq \sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| (-\log p(x))$$

$$\leq -\log \left(\min_x p(x) \right) \sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right|$$

$$\leq -\delta \log \left(\min_x p(x) \right)$$

$$= \eta > 0$$

Therefore,

$$-\eta \leq \sum_x \left(\frac{1}{n} N(x; \mathbf{x}) - p(x) \right) (-\log p(x)) \leq \eta$$

It then follows from (6.9) that

$$-n(H(X) + \eta) \leq \log p(\mathbf{x}) \leq -n(H(X) - \eta)$$
$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}$$

where $\eta \to 0$ as $\delta \to 0$, proving Property 1.

2. To prove Property 2, we write $N(x; \mathbf{X}) = \sum_{k=1}^n B_k(x)$

$$B_k(x) = \begin{cases} 1 \text{ if } X_k = x \\ 0 \text{ if } X_k \neq x \end{cases}$$

Then $B_k(x), k = 1, 2, \cdots, n$ are i.i.d. random variables with

$$\Pr\{B_k(x) = 1\} = p(x)$$

and

$$\Pr\{B_k(x) = 0\} = 1 - p(x)$$

Note that

$$EB_k(x) = (1 - p(x)) \cdot 0 + p(x) \cdot 1 = p(x)$$

By the weak law of large numbers, for any $\delta > 0$ and for any $x \in \mathcal{X}$

$$\Pr\left\{ \left| \frac{1}{n} \sum_{k=1}^n B_k(x) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \right\} < \frac{\delta}{|\mathcal{X}|}$$

for n sufficiently large. Then

$$\Pr\left\{ \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x \right\} = \Pr\left\{ \left| \frac{1}{n} \sum_{k=1}^n B_k(x) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x \right\}$$

$$= \Pr\left\{ \bigcup_x \left\{ \left| \frac{1}{n} \sum_{k=1}^n B_k(x) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \right\} \right\}$$

$$\leq \sum_x \Pr\left\{ \left| \frac{1}{n} \sum_{k=1}^n B_k(x) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \right\}$$

$$< \sum_x \frac{\delta}{|\mathcal{X}|} = \delta$$

where we have used the union bound $(\Pr\{A \cup B\} \leq \Pr\{A\} + \Pr\{B\})$ to obtain (6.27). since

$$\sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| > \delta$$

implies

$$\left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \quad \text{for some } x \in \mathcal{X}$$

we have

$$\Pr\left\{ \mathbf{X} \in T_{[X]\delta}^n \right\} = \Pr\left\{ \sum_x \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| \leq \delta \right\}$$

$$= 1 - \Pr\left\{ \sum_x \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| > \delta \right\}$$

$$\geq 1 - \Pr\left\{ \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x \in \mathcal{X} \right\}$$

$$> 1 - \delta$$

proving Property 2 .

Homework   2. Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$, where X_k are i.i.d. with generic random variable X. Prove that

$$\Pr\left\{\mathbf{X} \in T^n_{[X]\delta}\right\} \geq 1 - \frac{|\mathcal{X}|^3}{n\delta^2}$$

for any n and $\delta > 0$. This shows that $\Pr\left\{\mathbf{X} \in T^n_{[X]\delta}\right\} \to 1$ as $\delta \to 0$ and $n \to \infty$ if $\sqrt{n}\delta \to \infty$

Proof.

$$\Pr\left\{\mathbf{X} \in T^n_{[X]\delta}\right\} = \Pr\left\{\sum_x \left|\frac{1}{n}N(x;\mathbf{X}) - p(x)\right| \leq \delta\right\}$$

$$= 1 - \Pr\left\{\sum_x \left|\frac{1}{n}N(x;\mathbf{X}) - p(x)\right| > \delta\right\}$$

$$\geq 1 - \Pr\left\{\left|\frac{1}{n}N(x;\mathbf{X}) - p(x)\right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x \in \mathcal{X}\right\}$$

$$\Pr\left\{\left|\frac{1}{n}N(x;\mathbf{X}) - p(x)\right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x\right\} = \Pr\left\{\left|\frac{1}{n}\sum_{k=1}^n B_k(x) - p(x)\right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x\right\}$$

$$= \Pr\left\{\bigcup_x \left\{\left|\frac{1}{n}\sum_{k=1}^n B_k(x) - p(x)\right| > \frac{\delta}{|\mathcal{X}|}\right\}\right\}$$

$$\leq \sum_x \Pr\left\{\left|\frac{1}{n}\sum_{k=1}^n B_k(x) - p(x)\right| > \frac{\delta}{|\mathcal{X}|}\right\} \qquad \text{(i.i.d)}$$

$$\leq \sum_x \Pr\left\{|B_k(x) - p(x)| > \frac{\delta}{|\mathcal{X}|}\right\}$$

$$< \sum_x \frac{\sigma^2|\mathcal{X}|^2}{n\delta^2}$$

$$< \frac{|\mathcal{X}|^3}{n\delta^2}$$

Using Chebyshev's inequality

$$\mathrm{P}\left(\left|\bar{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Therefore we proved

$$\Pr\left\{\mathbf{X} \in T^n_{[X]\delta}\right\} \geq 1 - \frac{|\mathcal{X}|^3}{n\delta^2}$$

Theorem 6.3.   For sufficiently large n, there exists $\varphi(\delta) > 0$ such that

$$\Pr\left\{\mathbf{X} \notin T^n_{[X]\delta}\right\} < 2^{-n\varphi(\delta)}$$

The proof of this theorem is based on the Chernoff bound [66] which we prove in the next lemma.
Apply Lemma 6.4.

$$\log \Pr\left\{\sum_{k=1}^n B_k(x) \geq n(p(x) + \delta)\right\} \leq -sn(p(x) + \delta) + \log E\left[2^{s\sum_{k=1}^n B_k(x)}\right]$$

$$\stackrel{a)}{=} -sn(p(x) + \delta) + \log\left(\prod_{k=1}^n E\left[2^{sB_k(x)}\right]\right)$$

$$\stackrel{b)}{=} -sn(p(x) + \delta) + n\log\left(1 - p(x) + p(x)2^s\right)$$

$$\stackrel{c)}{\leq} -sn(p(x) + \delta) + n(\ln 2)^{-1}\left(-p(x) + p(x)2^s\right)$$

$$= -n\left[s(p(x) + \delta) + (\ln 2)^{-1}p(x)\left(1 - 2^s\right)\right]$$

29

where
(a) follows because $B_k(x)$ are mutually independent;
(b) is a direct evaluation of the expectation from the definition of $B_k(x)$ in (6.20)
(c) follows from the fundamental inequality $\ln a \leq a - 1$
In (6.48), upon defining

$$\beta_x(s, \delta) = s(p(x) + \delta) + (\ln 2)^{-1} p(x) \left(1 - 2^s\right)$$

we have

$$\log \Pr \left\{ \sum_{k=1}^{n} B_k(x) \geq n(p(x) + \delta) \right\} \leq -n\beta_x(s, \delta)$$

Or

$$\Pr \left\{ \sum_{k=1}^{n} B_k(x) \geq n(p(x) + \delta) \right\} \leq 2^{-n\beta_x(s,\delta)}$$

$$\Pr \left\{ \left| \frac{1}{n} \sum_{k=1}^{n} B_k(x) - p(x) \right| \geq \delta \right\} = \Pr \left\{ \left| \sum_{k=1}^{n} B_k(x) - np(x) \right| \geq n\delta \right\}$$

$$\leq \Pr \left\{ \sum_{k=1}^{n} B_k(x) \geq n(p(x) + \delta) \right\} + \Pr \left\{ \sum_{k=1}^{n} B_k(x) \leq n(p(x) - \delta) \right\}$$

$$\leq 2^{-n\beta_x(s,\delta)} + 2^{-n\sigma_x(s,\delta)}$$

$$\leq 2 \cdot 2^{-n \min(\beta_x(s,\delta), \sigma_x(s,\delta))}$$

$$= 2^{-n\left[\min(\beta_x(s,\delta), \sigma_x(s,\delta)) - \frac{1}{n}\right]}$$

$$= 2^{-n\varphi_x(\delta)}$$

where

$$\varphi_x(\delta) = \min\left(\beta_x(s, \delta), \sigma_x(s, \delta)\right) - \frac{1}{n}$$

$$\Pr \left\{ \mathbf{X} \in T_{[X]\delta}^n \right\} = \Pr \left\{ \sum_x \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| \leq \delta \right\}$$

$$\geq \Pr \left\{ \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| \leq \frac{\delta}{|\mathcal{X}|} \text{ for all } x \in \mathcal{X} \right\}$$

$$= 1 - \Pr \left\{ \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \text{ for some } x \in \mathcal{X} \right\}$$

$$\geq 1 - \sum_x \Pr \left\{ \left| \frac{1}{n} N(x; \mathbf{X}) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \right\}$$

$$= 1 - \sum_x \Pr \left\{ \left| \frac{1}{n} \sum_{k=1}^{n} B_k(x) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \right\}$$

$$= 1 - \sum_{x:p(x)>0} \Pr \left\{ \left| \frac{1}{n} \sum_{k=1}^{n} B_k(x) - p(x) \right| > \frac{\delta}{|\mathcal{X}|} \right\}$$

$$\geq 1 - \sum_{x:p(x)>0} 2^{-n\varphi_x\left(\frac{\delta}{|x|}\right)}$$

> **Lemma 6.4 (Chernoff Bound).** Let Y be a real random variable and s be any nonnegative real number. Then for any real number a,
>
> $$\log \Pr\{Y \geq a\} \leq -sa + \log E\left[2^{sY}\right]$$
>
> and
>
> $$\log \Pr\{Y \leq a\} \leq sa + \log E\left[2^{-sY}\right]$$

Proof . Let

$$u(y) = \begin{cases} 1 \text{ if } y \geq 0 \\ 0 \text{ if } y < 0 \end{cases}$$

Then for any $s \geq 0$

$$u(y - a) \leq 2^{s(y-a)}$$

Taking expectation on both sides

$$E[u(Y - a)] \leq E\left[2^{s(Y-a)}\right] = 2^{-sa}E\left[2^{sY}\right]$$

since

$$E[u(Y - a)] = \Pr\{Y \geq a\} \cdot 1 + \Pr\{Y < a\} \cdot 0 = \Pr\{Y \geq a\}$$

we see that

$$\Pr\{Y \geq a\} \leq 2^{-sa}E\left[2^{sY}\right] = 2^{-sa+\log E\left[2^{sY}\right]}$$

4.2 Strong Typicality Versus Weak Typicality

We will prove in the next proposition that strong typicality is stronger than weak typicality in the sense that the former implies the latter

Proposition 6.5. For any $\mathbf{x} \in \mathcal{X}^n$, if $\mathbf{x} \in T^n_{[X]\delta}$, then $\mathbf{x} \in W^n_{[X]\eta}$, where $\eta \to 0$ as $\delta \to 0$

Proof. By Property 1 of strong AEP (Theorem 6.2), if $\mathbf{x} \in T^n_{[X]\delta}$, then

$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}$$

Or

$$H(X) - \eta \leq -\frac{1}{n}\log p(\mathbf{x}) \leq H(X) + \eta$$

where $\eta \to 0$ as $\delta \to 0$. Then $\mathbf{x} \in W^n_{[X]\eta}$ by Definition 5.2. The proposition is proved.

4.3 Joint Typicality

Consider a bivariate information source $\{(X_k, Y_k), k \geq 1\}$ where (X_k, Y_k) are i.i.d. with distribution $p(x, y)$. We use (X, Y) to denote the pair of generic random variables.

Definition 6.6. The strongly jointly typical set $T^n_{[XY]\delta}$ with respect to $p(x, y)$ is the set of $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that $N(x, y; \mathbf{x}, \mathbf{y}) = 0$ for $(x, y) \notin \mathcal{S}_{XY}$ and

$$\sum_x \sum_y \left|\frac{1}{n}N(x, y; \mathbf{x}, \mathbf{y}) - p(x, y)\right| \leq \delta$$

where $N(x, y; \mathbf{x}, \mathbf{y})$ is the number of occurrences of (x, y) in the pair of sequences $(\mathbf{x}, \mathbf{y})$ and δ is an arbitrarily small positive real number. A pair of sequences $(\mathbf{x}, \mathbf{y})$ is called strongly jointly δ -typical if it is in $T^n_{[XY]\delta}$

Theorem 6.7 (Consistency). If $(\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}$, then $\mathbf{x} \in T^n_{[X]\delta}$ and $\mathbf{y} \in T^n_{[Y]\delta}$

Proof. If $(\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}$, then

$$\sum_x \sum_y \left| \frac{1}{n} N(x, y; \mathbf{x}, \mathbf{y}) - p(x, y) \right| \leq \delta$$

Upon observing that

$$N(x; \mathbf{x}) = \sum_y N(x, y; \mathbf{x}, \mathbf{y})$$

we have

$$
\begin{aligned}
\sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| &= \sum_x \left| \frac{1}{n} \sum_y N(x, y; \mathbf{x}, \mathbf{y}) - \sum_y p(x, y) \right| \\
&= \sum_x \left| \sum_y \left(\frac{1}{n} N(x, y; \mathbf{x}, \mathbf{y}) - p(x, y) \right) \right| \\
&\leq \sum_x \sum_y \left| \frac{1}{n} N(x, y; \mathbf{x}, \mathbf{y}) - p(x, y) \right| \\
&\leq \delta
\end{aligned}
$$

Therefore, $\mathbf{x} \in T^n_{[X]\delta^*}$ Similarly, $\mathbf{y} \in T^n_{[Y]\delta}$. The theorem is proved.

Theorem 6.8 (Preservation). Let $Y = f(X)$. If

$$\mathbf{x} = (x_1, x_2, \cdots, x_n) \in T^n_{[X]\delta}$$

then

$$f(\mathbf{x}) = (y_1, y_2, \cdots, y_n) \in T^n_{[Y]\delta}$$

where $y_i = f(x_i)$ for $1 \leq i \leq n$

Proof. Consider $\mathbf{x} \in T^n_{[X]\delta}$, i.e.,

$$\sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| < \delta$$

since $Y = f(X)$

$$p(y) = \sum_{x \in f^{-1}(y)} p(x)$$

for all $y \in \mathcal{Y}$. On the other hand,

$$N(y; f(\mathbf{x})) = \sum_{x \in f^{-1}(y)} N(x; \mathbf{x})$$

for all $y \in \mathcal{Y}$. Then

$$
\begin{aligned}
\sum_y \left| \frac{1}{n} N(y; f(\mathbf{x})) - p(y) \right| &= \sum_y \left| \sum_{x \in f^{-1}(y)} \left(\frac{1}{n} N(x; \mathbf{x}) - p(x) \right) \right| \\
&\leq \sum_y \sum_{x \in f^{-1}(y)} \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| \\
&= \sum_x \left| \frac{1}{n} N(x; \mathbf{x}) - p(x) \right| \\
&< \delta
\end{aligned}
$$

Therefore, $f(\mathbf{x}) \in T^n_{[Y]\delta}$, proving the lemma.

For a bivariate i.i.d. source $\{(X_k, Y_k)\}$, we have the strong joint asymptotic equipartition property (strong JAEP), which can readily be obtained by applying the strong AEP to the source $\{(X_k, Y_k)\}$.

Theorem 6.9 (Strong JAEP). Let

$$(\mathbf{X}, \mathbf{Y}) = ((X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n))$$

where (X_i, Y_i) are i.i.d. with generic pair of random variables (X, Y). Then there exists $\lambda > 0$ such that $\lambda \to 0$ as $\delta \to 0$, and the following hold:

1) If $(\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}$, then

$$2^{-n(H(X,Y)+\lambda)} \le p(\mathbf{x}, \mathbf{y}) \le 2^{-n(H(X,Y)-\lambda)}$$

2) For n sufficiently large,

$$\Pr\left\{(\mathbf{X}, \mathbf{Y}) \in T^n_{[XY]\delta}\right\} > 1 - \delta$$

3) For n sufficiently large,

$$(1-\delta)2^{n(H(X,Y)-\lambda)} \le \left|T^n_{[XY]\delta}\right| \le 2^{n(H(X,Y)+\lambda)}$$

From the strong JAEP, we can see the following. since there are approximately $2^{nH(X,Y)}$ typical $(\mathbf{x}, \mathbf{y})$ pairs and approximately $2^{nH(X)}$ typical $\mathbf{x}$, for a typical $\mathbf{x}$, the number of $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ is jointly typical is approximately

$$\frac{2^{nH(X,Y)}}{2^{nH(X)}} = 2^{nH(Y|X)}$$

on the average. The next theorem reveals that this is not only true on the average, but it is in fact true for every typical $\mathbf{x}$ as long as there exists at least one $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ is jointly typical.

Theorem 6.10 (Conditional Strong AEP). For any $\mathbf{x} \in T^n_{[X]\delta}$, define

$$T^n_{[Y|X]\delta}(\mathbf{x}) = \left\{\mathbf{y} \in T^n_{[Y]\delta} : (\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}\right\}$$

If $\left|T^n_{[Y|X]\delta}(\mathbf{x})\right| \ge 1$, then

$$2^{n(H(Y|X)-\nu)} \le \left|T^n_{[Y|X]\delta}(\mathbf{x})\right| \le 2^{n(H(Y|X)+\nu)}$$

where $\nu \to 0$ as $n \to \infty$ and $\delta \to 0$

Proof. Assume that $\left|T^n_{[Y|X]\delta}(\mathbf{x})\right| \ge 1$. We now prove the lower bound on $\left|T^n_{[Y|X]\delta}(\mathbf{x})\right|$. Let

$$\{K(x,y), (x,y) \in \mathcal{X} \times \mathcal{Y}\}$$

be any set of nonnegative integers such thats

$$\sum_y K(x,y) = N(x; \mathbf{x})$$

for all $x \in \mathcal{X}$, and for any $\mathbf{y} \in \mathcal{Y}^n$, if

$$N(x, y; \mathbf{x}, \mathbf{y}) = K(x,y)$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, then $(\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}$ Then by Definition 6.6, $\{K(x,y)\}$ satisfies

$$\sum_x \sum_y \left|\frac{1}{n}K(x,y) - p(x,y)\right| \le \delta$$

which implies that for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$,

$$\left|\frac{1}{n}K(x,y) - p(x,y)\right| \le \delta$$

Or

$$p(x, y) - \delta \le \frac{1}{n} K(x, y) \le p(x, y) + \delta$$

Straightforward combinatorics reveals that the number of $\mathbf{y}$ which satisfy the constraints in (6.128) is equal to

$$M(K) = \prod_x \frac{N(x; \mathbf{x})!}{\prod_y K(x, y)!}$$

and it is readily seen that

$$\left| T^n_{[Y|X]\delta}(\mathbf{x}) \right| \ge M(K)$$

Lemma 6.11. For any $n > 0$,

$$n \ln n - n < \ln n! < (n + 1) \ln(n + 1) - n$$

Proof. First, we write

$$\ln n! = \ln 1 + \ln 2 + \cdots + \ln n$$

since $\ln x$ is a monotonically increasing function of x, we have

$$\int_{k-1}^k \ln x dx < \ln k < \int_k^{k+1} \ln x dx$$

Summing over $1 \le k \le n$, we have

$$\int_0^n \ln x dx < \ln n! < \int_1^{n+1} \ln x dx$$

Or

$$n \ln n - n < \ln n! < (n + 1) \ln(n + 1) - n$$

The lemma is proved.

The above theorem says that for any typical $\mathbf{x}$, as long as there is one typical $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ is jointly typical, there are approximately $2^{nH(Y|X)}$ $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ is jointly typical. This theorem has the following corollary that the number of such typical x grows with n at almost the same rate as the total number of typical $\mathbf{x}$.

Corollary 6.12. For a joint distribution $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, let $S^n_{[X]\delta}$ be the set of all sequences $\mathbf{x} \in T^n_{[X]\delta}$ such that $T^n_{[Y|X]\delta}(\mathbf{x})$ is nonempty. Then

$$\left| S^n_{[X]\delta} \right| \ge (1 - \delta) 2^{n(H(X) - \psi)}$$

where $\psi \to 0$ as $n \to \infty$ and $\delta \to 0$.

Proof. By the consistency of strong typicality (Theorem 6.7), if $(\mathbf{x}, \mathbf{y}) \in T^n_{[XY]\delta}$, then $\mathbf{x} \in T^n_{[X]\delta}$. In particular, $\mathbf{x} \in S^n_{[X]\delta}$. Then

$$T^n_{[XY]\delta} = \bigcup_{\mathbf{x} \in S^n_{[X]\delta}} \left\{ (\mathbf{x}, \mathbf{y}) : \mathbf{y} \in T^n_{[Y|X]\delta}(\mathbf{x}) \right\}$$

Using the lower bound on $\left| T^n_{[XY]\delta} \right|$ in Theorem 6.9 and the upper bound on $\left| T^n_{[Y|X]\delta}(\mathbf{x}) \right|$ in the last theorem, we have

$$(1 - \delta) 2^{n(H(X,Y) - \lambda)} \le \left| T^n_{[XY]\delta} \right| \le \left| S^n_{[X]\delta} \right| 2^{n(H(Y|X) + \nu)}$$

which implies

$$\left| S^n_{[X]\delta} \right| \ge (1 - \delta) 2^{n(H(X) - (\lambda + \nu))}$$

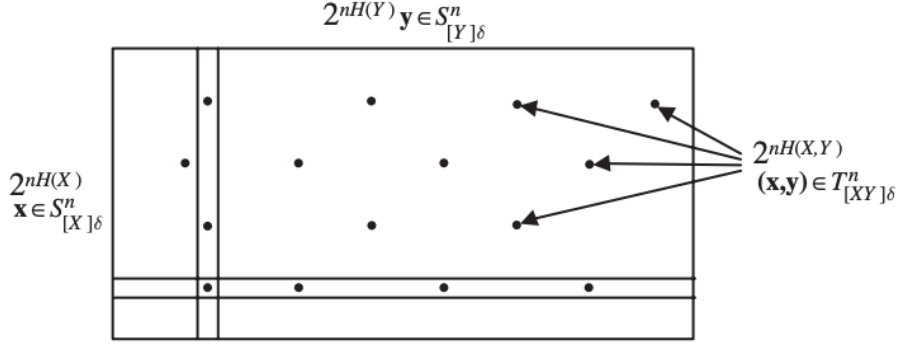The theorem is proved upon letting $\psi = \lambda + \nu$.

Fig. 6.2. A two-dimensional strong joint typicality array.

Figure 14: In this array, the rows and the columns are the typical sequences $\mathbf{x} \in S_{[X]\delta}^n$ and $\mathbf{y} \in S_{[Y]\delta}^n$ respectively. The total number of rows and columns are approximately equal to $2^{nH(X)}$ and $2^{nH(Y)}$, respectively. An entry indexed by $(\mathbf{x}, \mathbf{y})$ receives a dot if $(\mathbf{x}, \mathbf{y})$ is strongly jointly typical. The total number of dots is approximately equal to $2^{nH(X,Y)}$. The number of dots in each row is approximately equal to $2^{nH(Y|X)}$, while the number of dots in each column is approximately equal to $2^{nH(X|Y)}$

Proposition 6.13. With respect to a joint distribution $p(x,y)$ on $\mathcal{X} \times \mathcal{Y}$, for any $\delta > 0$

$$\Pr\left\{\mathbf{X} \in S_{[X]\delta}^n\right\} > 1 - \delta$$

for n sufficiently large.

Homework (Proof.) By Theorem 6.7. If $(\mathbf{x}, \mathbf{y}) \in T_{[XY]\delta}^n$, then $\mathbf{x} \in T_{[X]\delta}^n$

$$\sum_x \left|\frac{1}{n}N(x;\mathbf{x}) - p(x)\right| = \sum_x \left|\frac{1}{n}\sum_y N(x,y;\mathbf{x},\mathbf{y}) - \sum_y p(x,y)\right|$$

$$\sum_x \left|\sum_y \left(\frac{1}{n}N(x,y;\mathbf{x},\mathbf{y}) - p(x,y)\right)\right|$$

$$\leq \sum_x \sum_y \left|\frac{1}{n}N(x,y;\mathbf{x},\mathbf{y}) - p(x,y)\right|$$

By Weak Law of Large Number

$$\Pr\left\{\left|\frac{1}{n}N(x,y;\mathbf{x},\mathbf{y}) - p(x,y)\right| > \frac{\delta}{|\mathcal{X}||\mathcal{Y}|}\right\} < \frac{\delta}{|\mathcal{X}||\mathcal{Y}|}$$

Then

$$\sum_x \left|\frac{1}{n}N(x;\mathbf{x}) - p(x)\right| < \delta$$

4.4 An Interpretation of the Basic Inequalities

Consider random variables X, Y, and Z and a fixed $\mathbf{z} \in S_{[Z]\delta}^n$, so that $T_{[XY|Z]\delta}^n(\mathbf{z})$ is nonempty. By the consistency of strong typicality, if $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in T_{[XYZ]\delta}^n$, then $(\mathbf{x}, \mathbf{z}) \in T_{[XZ]\delta}^n$ and $(\mathbf{y}, \mathbf{z}) \in T_{[YZ]\delta}^n$, or $\mathbf{x} \in T_{[X|Z]\delta}^n(\mathbf{z})$ and $\mathbf{y} \in T_{[Y|Z]\delta}^n(\mathbf{z})$, respectively. Thus

$$T_{[XY|Z]\delta}^n(\mathbf{z}) \subset T_{[X|Z]\delta}^n(\mathbf{z}) \times T_{[Y|Z]\delta}^n(\mathbf{z})$$

which implies

$$\left|T_{[XY|Z]\delta}^n(\mathbf{z})\right| \leq \left|T_{[X|Z]\delta}^n(\mathbf{z})\|T_{[Y|Z]\delta}^n(\mathbf{z})\right|$$

35

Applying the lower bound in Theorem 6.10 to $T^n_{[XY|Z]\delta}(\mathbf{z})$ and the upper bound to $T^n_{[X|Z]\delta}(\mathbf{z})$ and $T^n_{[Y|Z]\delta}(\mathbf{z})$, we have

$$2^{n(H(X,Y|Z)-\zeta)} \leq 2^{n(H(X|Z)+\gamma)} 2^{n(H(Y|Z)+\phi)}$$

where $\zeta, \gamma, \phi \to 0$ as $n \to \infty$ and $\delta \to 0$. Taking logarithm to the base 2 and dividing by n, we obtain

$$H(X,Y \mid Z) \leq H(X \mid Z) + H(Y \mid Z)$$

upon letting $n \to \infty$ and $\delta \to 0$. This inequality is equivalent to

$$I(X;Y \mid Z) \geq 0$$

Thus we have proved the nonnegativity of conditional mutual information. Since all Shannon's information measures are special cases of conditional mutual information, we have proved the nonnegativity of all Shannon's information measures, namely the basic inequalities.

Homework   Show that $(\mathbf{x}, \mathbf{y}) \in T^n_{[X,Y]\delta}$ and $(\mathbf{y}, \mathbf{z}) \in T^n_{[Y,Z]\delta}$ do not imply $(\mathbf{x}, \mathbf{z}) \in T^n_{[X,Z]\delta}$

Counter Example   In the following problems, for a sequence $\mathbf{x} \in \mathcal{X}^n$, let $q_{\mathbf{x}}$ be the empirical distribution of $\mathbf{x}$, i.e., $q_{\mathbf{x}}(x) = n^{-1}N(x; \mathbf{x})$ for all $x \in \mathcal{X}$. Similarly, for a pair of sequences $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$, let $q_{\mathbf{x},\mathbf{y}}$ be the joint empirical distribution of $(\mathbf{x}, \mathbf{y})$, i.e., $q_{\mathbf{x},\mathbf{y}}(x,y) = n^{-1}N(x,y; \mathbf{x}, \mathbf{y})$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$

5. Alternative definition of weak typicality.   Let $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ be an i.i.d. sequence whose generic random variable X is distributed with $p(x)$. Let $q_{\mathbf{x}}$ be the empirical distribution of the sequence $\mathbf{x}$, i.e., $q_{\mathbf{x}}(x) = n^{-1}N(x; \mathbf{x})$ for all $x \in \mathcal{X}$, where $N(x; \mathbf{x})$ is the number of occurrence of x in $\mathbf{x}$

$$|D\left(q_{\mathbf{x}}\|p\right) + H\left(q_{\mathbf{x}}\right) - H(p)| \le \epsilon$$

Homework Alternative definition of strong typicality.   Show that (6.1)

$$\sum_x \left| \frac{1}{n}N(x; \mathbf{x}) - p(x) \right| \le \delta$$

is equivalent to

$$V\left(q_{\mathbf{x}}, p\right) \le \delta$$

where $V(\cdot, \cdot)$ denotes the variational distance.

$$V(p, q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$$

Thus strong typicality can be regarded as requiring the empirical distribution of a sequence to be close to the probability distribution of the generic random variable in variational distance. Also compare the result here with the alternative definition of weak typicality (Problem 5 in Chapter 5).

Proof.

$$V\left(q_{\mathbf{x}}, p\right) = \sum_{x \in \mathcal{X}} |q_{\mathbf{x}} - p(x)| = \sum_{x \in \mathcal{X}} \left| \frac{1}{n}N(x; \mathbf{x}) - p(x) \right| \le \delta$$

We can see that strong typicality is stronger than weak in the sense that weak typicality only requires the closeness in entropy. Note that $d(q_x, p) = |D\left(q_{\mathbf{x}}\|p\right) + H\left(q_{\mathbf{x}}\right) - H(p)| = 0$ when $q_x = p$

Homework   8. The empirical distribution $q_{\mathbf{x}}$ of the sequence x is also called the type of $\mathbf{x}$. Assuming that $\mathcal{X}$ is finite, show that there are a total of $\begin{pmatrix} n + |\mathcal{X}| - 1 \\ n \end{pmatrix}$ distinct types $q_{\mathbf{x}}$. Hint: There are $\begin{pmatrix} a + b - 1 \\ a \end{pmatrix}$ ways to distribute a identical balls in b boxes.

Proof.   The original problem may be reformulated as arranging $k-1$ bars and the n balls, by selecting n positions for balls out of $n + k - 1$ locations.

$$\underbrace{* * *}_{n \text{ balls}} \underbrace{[|||||||]}_{k-1 \text{ bars}}$$

Directly apply this idea to get emperical distribution q_x, treat as for assigning each sample x into $|\mathcal{X}|$ boxes. Which gives the result

$$\begin{pmatrix} n + |\mathcal{X}| - 1 \\ n \end{pmatrix}$$

Homework   6. Let p be any probability distribution over a finite set $\mathcal{X}$ and η be a real number in $(0, 1)$. Prove that for any subset A of $\mathcal{X}^n$ with $p^n(A) \ge \eta$

$$\left| A \cap T^n_{[X]\delta} \right| \ge 2^{n\left(H(p) - \delta'\right)}$$

where $\delta' \to 0$ as $\delta \to 0$ and $n \to \infty$

Proof   Recall for $x \in T_{[X]\delta}$,
$$2^{-n(H(X)+\delta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\delta)}$$

then for $y \in A \cap T_{[X]\delta}$, we have $p(\mathbf{y}) = p(\mathbf{x})p(A) = \eta p(\mathbf{x})$
(Note that given in question, $p(A) \geq \eta$, for simplicity, we fix $p(A) = \eta, \eta \in (0,1)$)
By using the upper bound $p(\mathbf{x}) \leq 2^{-n(H(X)-\delta)}$

$$p(\mathbf{y}) \leq \eta 2^{-n(H(X)-\delta)} = 2^{-n(\frac{\log(\eta)}{n}+H(X)-\delta)}$$

As $n \to \infty$, $\frac{\log(\eta)}{n}H(X) - \delta \to H(X) - \delta$

$$\left| A \cap T_{[X]\delta}^n \right| 2^{-n(H(X)-\delta)} \geq 1$$

$$\left| A \cap T_{[X]\delta}^n \right| \geq 2^{n(H(X)-\delta)}$$

5 Discrete Memoryless Channels

5.1 7.1 Definition and Capacity

Definition 7.1. Let $\mathcal{X}$ and $\mathcal{Y}$ be discrete alphabets, and $p(y \mid x)$ be a transition matrix from $\mathcal{X}$ to $\mathcal{Y}$. A discrete channel $p(y \mid x)$ is a single input-single output system with input random variable X taking values in $\mathcal{X}$ and output random variable Y taking values in $\mathcal{Y}$ such that

$$\Pr\{X = x, Y = y\} = \Pr\{X = x\}p(y \mid x)$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$

Define random variables Z_x with $\mathcal{Z}_x = \mathcal{Y}$ for $x \in \mathcal{X}$ such that

$$\Pr\{Z_x = y\} = p(y \mid x)$$

for all $y \in \mathcal{Y}$. We assume that $Z_x, x \in \mathcal{X}$ are mutually independent and also independent of X. Further define the random variable

$$Z = (Z_x : x \in \mathcal{X})$$

called the noise variable. Note that Z is independent of X. Now define a random variable taking values in $\mathcal{Y}$ as

$$Y = Z_x \quad \text{if } X = x$$

Evidently, Y is a function of X and Z. Then for $x \in \mathcal{X}$ such that $\Pr\{X = x\} > 0$, we have

$$
\begin{aligned}
\Pr\{X = x, Y = y\} &= \Pr\{X = x\} \Pr\{Y = y \mid X = x\} \\
&= \Pr\{X = x\} \Pr\{Z_x = y \mid X = x\} \\
&= \Pr\{X = x\} \Pr\{Z_x = y\} \\
&= \Pr\{X = x\}p(y \mid x)
\end{aligned}
$$

Definition 7.2. Let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{Z}$ be discrete alphabets. Let $\alpha : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and Z be a random variable taking values in $\mathcal{Z}$, called the noise variable. A discrete channel (α, Z) is a single input-single output system with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. For any input random variable X, the noise variable Z is independent of X, and the output random variable Y is given by
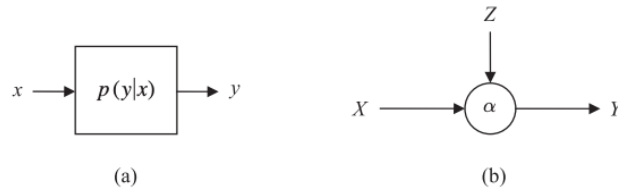
$$Y = \alpha(X, Z)$$

Figure 15: Illustrations of (a) a discrete channel $p(y|x)$ and (b) a discrete channel (α, Z).

The next definition gives the condition for the equivalence of the two specifications of a discrete channel according to Definitions 7.1 and 7.2, respectively.

Definition 7.3. Two discrete channels $p(y \mid x)$ and (α, Z) defined on the same input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ are equivalent if

$$\Pr\{\alpha(x, Z) = y\} = p(y \mid x)$$

for all x and y

Definition 7.5. *A discrete memoryless channel (α, Z) is a sequence of replicates of a generic discrete channel (α, Z). These discrete channels are indexed by a discrete-time index i, where $i \geq 1$, with the ith channel being available for transmission at time i. Transmission through a channel is assumed to be instantaneous. Let X_i and Y_i be, respectively, the input and the output of the DMC at time i, and let T_{i-} denote all the random variables that are generated in the system before X_i. The noise variable Z_i for the transmission at time i is a copy of the generic noise variable Z and is independent of (X_i, T_{i-}). The output of the DMC at time i is given by*

$$Y_i = \alpha \left(X_i, Z_i \right)$$

We will prove subsequently that C is in fact the maximum rate at which information can be communicated reliably through a DMC.

Definition 7.6. The capacity of a discrete memoryless channel $p(y \mid x)$ is defined as

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

where X and Y are, respectively, the input and the output of the generic discrete channel, and the maximum is taken over all input distributions $p(x)$.

From the above definition, we see that $C \geq 0$ because $I(X;Y) \geq 0$ for all input distributions $p(x)$. By Theorem 2.43 , we have

$$C = \max_{p(x)} I(X;Y) \leq \max_{p(x)} H(X) = \log |\mathcal{X}|$$

Likewise, we have $C \leq \log |\mathcal{Y}|$. Therefore,

$$C \leq \min(\log |\mathcal{X}|, \log |\mathcal{Y}|)$$

5.2   The Channel Coding Theorem

Definition 7.9. An (n, M) code for a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined by an encoding function

$$f : \{1, 2, \cdots, M\} \rightarrow \mathcal{X}^n$$

and a decoding function

$$g : \mathcal{Y}^n \rightarrow \{1, 2, \cdots, M\}$$

The set $\{1, 2, \cdots, M\}$, denoted by $\mathcal{W}$, is called the message set. The sequences $f(1), f(2), \cdots, f(M)$ in $\mathcal{X}^n$ are called codewords, and the set of codewords is called the codebook.

Definition 7.10. For all $1 \leq w \leq M$, let

$$\lambda_w = \Pr\{\hat{W} \neq w \mid W = w\} = \sum_{\mathbf{y} \in \mathcal{Y}^n : g(\mathbf{y}) \neq w} \Pr\{\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = f(w)\}$$

be the conditional probability of error given that the message is w.

Definition 7.11. The maximal probability of error of an (n, M) code is defined as

$$\lambda_{\max} = \max_{w} \lambda_w$$

Definition 7.12. The average probability of error of an (n, M) code is defined as

$$P_e = \Pr\{\hat{W} \neq W\}$$

$$P_e = \Pr\{\hat{W} \neq W\}$$
$$= \sum_w \Pr\{W = w\} \Pr\{\hat{W} \neq W \mid W = w\}$$
$$= \sum_w \frac{1}{M} \Pr\{\hat{W} \neq w \mid W = w\}$$
$$= \frac{1}{M} \sum_w \lambda_w$$
$$\leq \lambda_{\max}$$

Definition 7.13. The rate of an (n, M) channel code is $n^{-1} \log M$ in bits per use.

Definition 7.14. A rate R is asymptotically achievable for a discrete memoryless channel if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\frac{1}{n} \log M > R - \epsilon$$

and

$$\lambda_{\max} < \epsilon$$

For brevity, an asymptotically achievable rate will be referred to as an achievable rate.

Theorem 7.15 (Channel Coding Theorem). A rate R is achievable for a discrete memoryless channel if and only if $R \leq C$, the capacity of the channel.
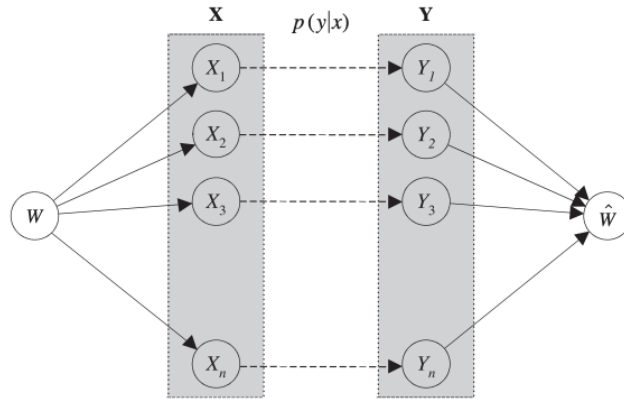
5.3 The Converse

Figure 16: The dependency graph for a channel code without feedback.

From the dependency graph, we see that for all $(w, \mathbf{x}, \mathbf{y}, \hat{w}) \in \mathcal{W} \times \mathcal{X}^n \times \mathcal{Y}^n \times \hat{\mathcal{W}}$ such that $q(\mathbf{x}) > 0$ and $q(\mathbf{y}) > 0$

$$q(w, \mathbf{x}, \mathbf{y}\hat{w}) = q(w) \left(\prod_{i=1}^n q(x_i \mid w) \right) \left(\prod_{i=1}^n p(y_i \mid x_i) \right) q(\hat{w} \mid \mathbf{y})$$

Note that $q(w) > 0$ for all w so that $q(x_i \mid w)$ are well defined, and $q(x_i \mid w)$ and $q(\hat{w} \mid \mathbf{y})$ are both deterministic. Denote the set of nodes $X_1, X_2, \cdots, X_n$ by $\mathbf{X}$ and the set of nodes $Y_1, Y_2, \cdots, Y_n$ by $\mathbf{Y}$. We notice the following

structure in the dependency graph: all the edges from W end in $\mathbf{X}$, all the edges from $\mathbf{X}$ end in Y, and all the edges from Y end in $\hat{W}$. This suggests that the random variables $W, \mathbf{X}, \mathbf{Y}$, and $\hat{W}$ form the Markov chain

$$W \to \mathbf{X} \to \mathbf{Y} \to \hat{W}$$

The validity of this Markov chain can be formally justified by applying Proposition 2.9 to (7.97), so that for all $(w, \mathbf{x}, \mathbf{y}, \hat{w}) \in \mathcal{W} \times \mathcal{X}^n \times \mathcal{Y}^n \times \hat{\mathcal{W}}$ such that $q(\mathbf{x}) > 0$ and $q(\mathbf{y}) > 0$, we can write

$$q(w, \mathbf{x}, \mathbf{y}, \hat{w}) = q(w)q(\mathbf{x} \mid w)q(\mathbf{y} \mid \mathbf{x})q(\hat{w} \mid \mathbf{y})$$

Consider a channel code whose probability of error is arbitrarily small. since $W, \mathbf{X}, \mathbf{Y}$, and $\hat{W}$ form the Markov chain in (7.98), the information diagram for these four random variables is as shown in Figure 7.10. Moreover, $\mathbf{X}$ is a function of W, and $\hat{W}$ is a function of $\mathbf{Y}$. These two relations are equivalent to $H(\mathbf{X} \mid W) = 0$ and $H(\hat{W} \mid \mathbf{Y}) = 0$ respectively. Since the probability of error is arbitrarily small, W and $\hat{W}$ are essentially identical. To gain insight into the problem, we assume for the time being that W and $\hat{W}$ are equivalent, so that

$$H(\hat{W} \mid W) = H(W \mid \hat{W}) = 0$$

since the I -Measure μ^* for a Markov chain is nonnegative, the constraints in $(7.102) - (7.104)$ imply that μ^* vanishes on all the atoms in Figure 7.10 marked with a 0. Immediately, we see that

$$H(W) = I(\mathbf{X}; \mathbf{Y})$$

That is, the amount of information conveyed through the channel is essentially the mutual information between the input sequence and the output sequence of the channel.

For a single transmission, we see from the definition of channel capacity that the mutual information between the input and the output cannot exceed the capacity of the channel, i.e., for all $1 \le i \le n$

$$I\left(X_i; Y_i\right) \le C$$

Summing i from 1 to n, we have

$$\sum_{i=1}^{n} I\left(X_i; Y_i\right) \le nC$$

Lemma 7.16. For a discrete memoryless channel used with a channel code without feedback, for any $n \ge 1$,

$$I(\mathbf{X}; \mathbf{Y}) \le \sum_{i=1}^{n} I\left(X_i; Y_i\right)$$

where X_i and Y_i are, respectively, the input and the output of the channel at time i.

Then the converse of the channel coding theorem then follows from

$$\frac{1}{n}\log M = \frac{1}{n}H(W) = \frac{1}{n}I(\mathbf{X}; \mathbf{Y})$$
$$\le \frac{1}{n}\sum_{i=1}^{n} I\left(X_i; Y_i\right)$$
$$\le C$$

We now formally prove the converse of the channel coding theorem. Let R be an achievable rate, i.e., for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\frac{1}{n}\log M > R - \epsilon, \qquad \lambda_{\max} < \epsilon$$

Consider

$$\log M \stackrel{a}{=} H(W) - H(W \mid \hat{W}) + I(W; \hat{W})$$
$$\stackrel{a}{\le} H(W \mid \hat{W}) + I(\mathbf{X}; \mathbf{Y})$$
$$\le H(W \mid \hat{W}) + \sum_{i=1}^{n} I\left(X_i; Y_i\right)$$
$$\le H(W \mid \hat{W}) + nC \qquad (7.126)$$

where (a) follows from (7.80) (b) follows from the data processing theorem since $W \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \hat{W}$ (c) follows from Lemma 7.16 (d) follows from (7.107) From (7.87) and Fano's inequality (cf. Corollary 2.48), we have

$$H(W \mid \hat{W}) < 1 + P_e \log M$$

Therefore, from (7.126)

$$\begin{aligned} \log M &< 1 + P_e \log M + nC \\ &\leq 1 + \lambda_{\max} \log M + nC \\ &< 1 + \epsilon \log M + nC \end{aligned}$$

where we have used (7.92) and (7.121), respectively, to obtain the last two inequalities. Dividing by n and rearranging the terms, we have

$$\frac{1}{n} \log M < \frac{\frac{1}{n} + C}{1 - \epsilon}$$

and from (7.120), we obtain

$$R - \epsilon < \frac{\frac{1}{n} + C}{1 - \epsilon}$$

For any $\epsilon > 0$, the above inequality holds for all sufficiently large n. Let $n \rightarrow \infty$ and then $\epsilon \rightarrow 0$, we conclude that

$$R \leq C$$

From the above proof, we can obtain an asymptotic bound on P_e when the rate of the code $\frac{1}{n} \log M$ is greater than C. Consider (7.128) and obtain

$$P_e \geq 1 - \frac{1 + nC}{\log M} = 1 - \frac{\frac{1}{n} + C}{\frac{1}{n} \log M}$$

Then

$$P_e \geq 1 - \frac{\frac{1}{n} + C}{\frac{1}{n} \log M} \approx 1 - \frac{C}{\frac{1}{n} \log M}$$

when n is large. This asymptotic bound on P_e, which is strictly positive if $\frac{1}{n} \log M > C$. The converse of the channel coding theorem we have proved is called the weak converse. A stronger version of this result called the strong converse can be proved, which says that $P_e \rightarrow 1$ as $n \rightarrow \infty$ if there exists an $\epsilon > 0$ such that $\frac{1}{n} \log M \geq C + \epsilon$ for all n.

5.4   Achievability

5.5   A Discussion

5.6   Feedback Capacity

5.7   Separation of Source and Channel Coding