1 Introduction

2 Information Measures

2.1 Independence and Markov Chains

Definition 2.1 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 $(x,y) \in \mathcal{X} \times \mathcal{Y}$).

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

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

for all $x_1, x_2, \cdots, x_n$

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

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

for all x, y, and z, or equivalently,

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

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 $p(y) > 0$

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

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

for all $x_1, x_2, \cdots, x_n$, or equivalently,

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

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

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

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 $p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0$

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.

Proposition 2.12 Let X_1, X_2, X_3, and X_4 be random variables such that $p(x_1, x_2, x_3, x_4)$ is strictly positive. Then

$$\left.\begin{array}{l} X_1 \perp X_4 \mid (X_2, X_3) \\ X_1 \perp X_3 \mid (X_2, X_4) \end{array}\right\} \Rightarrow X_1 \perp (X_3, X_4) \mid X_2$$

- Not true if p is not strictly positive
- Let $X_1 = Y, X_2 = Z$, and $X_3 = X_4 = (Y, Z)$, where $Y \perp Z$
- Then $X_1 \perp X_4 \,|(X_2, X_3), X_1 \perp X_3| (X_2, X_4)$, but $X_1 \not\perp (X_3, X_4) \mid X_2$
- p is not strictly positive because $p(x_1, x_2, x_3, x_4) = 0$ if $x_3 \neq (x_1, x_2)$ or $x_4 \neq (x_1, x_2)$

2.2 Shannon's Information Measures

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

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

- $H(X)$ depends only on the distribution of X but not on the actual value taken by X, hence also write $H(p)$
- Convention

$$E[g(X)] = \sum_x p(x) g(x)$$

  where summation is over $\mathcal{S}_X$

- linearity

$$E[f(X) + g(X)] = Ef(X) + Eg(X)$$

- Can write

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

- **(Binary Entropy Function)** For $0 \leq \gamma \leq 1$, define the binary entropy function

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

with the convention $0 \log 0 = 0$

- For $X \sim \{\gamma, 1-\gamma\}$

$$H(X) = h_b(\gamma)$$

- $h_b(\gamma)$ achieves the maximum value 1 when $\gamma = \frac{1}{2}$.

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

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

Definition (Entropy of Y conditioning on x)

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

- Write

$$H(Y \mid X) = \sum_{x} p(x) \left[-\sum_{y} p(y \mid x) \log p(y \mid x) \right]$$

- The inner sum is the entropy of Y conditioning on a fixed $x \in \mathcal{S}_X$, denoted by

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

- Thus

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

- Similarly,

$$H(Y \mid X, Z) = \sum_{z} p(z) H(Y \mid X, Z = z)$$

where

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

Proposition 2.16

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

and

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

Proof.

$$\begin{aligned} H(X,Y) &= -E\log p(X,Y) \\ &\overset{a)}{=} -E\log[p(X)p(Y \mid X)] \\ &\overset{b)}{=} -E\log p(X) - E\log p(Y \mid X) \\ &= H(X) + H(Y \mid X) \end{aligned}$$

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.

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

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

and

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

provided that all the entropies and conditional entropies are finite.

Definition 2.20 (Conditional Mutual Information) For random variables X, Y and Z, the mutual information between X and Y conditioning on Z is defined as

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

Remark: $I(X;Y \mid Z)$ is symmetrical in X and Y.
Similar to entropy, we have

$$I(X;Y \mid Z) = \sum_{z} p(z) I(X;Y \mid Z = z)$$

where

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

Proposition 2.21 The mutual information between a random variable X and itself conditioning on a random variable Z is equal to the conditional entropy of X given Z, i.e., , $I(X;X \mid Z) = H(X \mid Z)$

Proposition 2.22

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

and

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

provided that all the conditional entropies are finite. Remark All Shannon's information measures are special cases of conditional mutual information.

Continuity of Shannon's Information Measures for Fixed Finite Alphabets

- All Shannon's information measures are continuous when the alphabets are fixed and finite.
- For countable alphabets, Shannon's information measures are everywhere discontinuous.

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

Definition 2.23 Let p and q be two probability distributions on a common alphabet $\mathcal{X}$. The variational distance between p and q is defined as

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

The entropy function is continuous at p if

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

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

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

for all $q \in \mathcal{P}_{\mathcal{X}}$ satisfying

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

2.4 Chain Rules

Proposition 2.24 (Chain Rule for Entropy)

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

Proof by induction.

$$\begin{aligned} H(X_1, \cdots, X_m, X_{m+1}) &= H(X_1, \cdots, X_m) + H(X_{m+1} \mid X_1, \cdots, X_m) \\ &= \sum_{i=1}^{m} H(X_i \mid X_1, \cdots, X_{i-1}) + H(X_{m+1} \mid X_1, \cdots, X_m) \\ &= \sum_{i=1}^{m+1} H(X_i \mid X_1, \cdots, X_{i-1}) \end{aligned}$$

Proposition 2.25 (Chain Rule for Conditional Entropy)

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

Proof (1).

$$\begin{aligned} H(X_1, X_2, \cdots, X_n \mid Y) &= H(X_1, X_2, \cdots, X_n, Y) - H(Y) \\ &= H((X_1, Y), X_2, \cdots, X_n) - H(Y) \\ &\overset{a}{=} H(X_1, Y) + \sum_{i=2}^{n} H(X_i \mid X_1, \cdots, X_{i-1}, Y) - H(Y) \\ &= H(X_1 \mid Y) + \sum_{i=2}^{n} H(X_i \mid X_1, \cdots, X_{i-1}, Y) \\ &= \sum_{i=1}^{n} H(X_i \mid X_1, \cdots, X_{i-1}, Y) \end{aligned}$$

where a) follows from Proposition 2.24 (chain rule for entropy).

Proof (2).

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

Proposition 2.26 (Chain Rule for Mutual Information)

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

Proof.

$$\begin{aligned} I(X_1, X_2, \cdots, X_n; Y) &= H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n \mid Y) \\ &= \sum_{i=1}^n [H(X_i \mid X_1, \cdots, X_{i-1}) - H(X_i \mid X_1, \cdots, X_{i-1}, Y)] \\ &= \sum_{i=1}^n I(X_i; Y \mid X_1, \cdots, X_{i-1}) \end{aligned}$$

Proposition 2.27 (Chain Rule for Conditional Mutual Information)

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

2.5 Informational Divergence

Definition 2.28 (Informational 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:
  1. summation over $\mathcal{S}_p$
  2. $c \log \frac{c}{0} = \infty$ for $c > 0-$ if $D(p\|q) < \infty$, then $\mathcal{S}_p \subset \mathcal{S}_q$
- $D(p\|q)$ measures the "distance" between p and q.
- $D(p\|q)$ 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.
- Also called relative entropy or the Kullback-Leibler distance.

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 $a = 1$.

Theorem 2.6.2 (Cover: Jensen's inequality) If f is a convex function and X is a random variable, then

$$Ef(X) \geq f(EX)$$

Moreover, if f is strictly convex, then equality in (2.76) implies that $X = EX$ with probability 1, i.e., X is a constant.

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. If $q(x) = 0$ for some $x \in \mathcal{S}_p$, then $D(p\|q) = \infty$ and the theorem is trivially true. Therefore, we assume that $q(x) > 0$ for all $x \in \mathcal{S}_p$. Consider

$$\begin{aligned} D(p\|q) &= (\log e) \sum_{x \in \mathcal{S}_p} p(x) \ln \frac{p(x)}{q(x)} \\ &\geq (\log e) \sum_{x \in \mathcal{S}_p} p(x) \left(1 - \frac{q(x)}{p(x)}\right) \quad \text{(Corollary 2.30)} \\ &= (\log e) \left[\sum_{x \in \mathcal{S}_p} p(x) - \sum_{x \in \mathcal{S}_p} q(x)\right] \\ &\geq 0 \end{aligned}$$

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:

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}$$

- The divergence inequality implies the log-sum inequality.
  Proof. Let $a'_i = a_i / \sum_j a_j$ and $b'_i = b_i / \sum_j b_j$. Then $\{a'_i\}$ and $\{b'_i\}$ are probability distributions. Using the

divergence inequality,

$$\begin{aligned} 0 \le \sum_i a_i' \log \frac{a_i'}{b_i'} &= D(a||b) \\ &= \sum_i \frac{a_i}{\sum_j a_j} \log \frac{a_i/\sum_j a_j}{b_i/\sum_j b_j} \\ &= \frac{1}{\sum_j a_j}\left[\sum_i a_i \log \frac{a_i}{b_i} - \left(\sum_i a_i\right) \log \frac{\sum_j a_j}{\sum_j b_j}\right] \end{aligned}$$

- The log-sum inequality also implies the divergence inequality.

$$\begin{aligned} D(p\|q) &= \sum p(x) \log \frac{p(x)}{q(x)} \\ &\ge \left(\sum p(x)\right) \log \frac{(\sum p(x))}{(\sum q(x))} \\ &= 1 \log \frac{1}{1} = 0 \end{aligned}$$

- The two inequalities are equivalent.

Theorem 2.33 (Pinsker's Inequality)

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

- If $D(p\|q)$ or $D(q\|p)$ is small, then so is $V(p,q)$.
- 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 $V(p,q_k) \to 0$
- "Convergence in divergence" is a stronger notion than "convergence in variational distance."

2.6 The Basic Inequalities

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

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

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

Proof. Observe that

$$\begin{aligned} I(X;Y \mid Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \\ &= \sum_z p(z) \sum_{x,y} p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \\ &= \sum_z p(z) D\left(p_{XY|z} \| p_{X|z} p_{Y|z}\right) \\ &\ge 0 \end{aligned}$$

Finally, we see from Theorem 2.31 that $I(X;Y \mid Z) = 0$ if and only if for all $z \in \mathcal{S}_z$

$$p(x,y \mid z) = p(x \mid z)p(y \mid z)$$

or

$$p(x,y,z) = p(x,z)p(y \mid z)$$

for all x and y. Therefore, X and Y are independent conditioning on Z. The proof is accomplished. □

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

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

2.7 Some Useful Information Inequalities

Theorem 2.38 (Conditioning Does Not Increase Entropy)

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

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

Similarly, $H(Y \mid X, Z) \le H(Y \mid Z)$.
Warning: $I(X;Y \mid Z) \le I(X;Y)$ does not hold in general.

Theorem 2.39 (Independence Bound for Entropy)

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, \cdots, X_{i-1}) \le \sum_{i=1}^{n} H(X_i) \quad \text{(Theorem 2.38)}$$

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

The inequality is tight if and only if it is tight for each i,

$$H(X_i \mid X_1, \cdots, X_{i-1}) = H(X_i)$$

for $1 \le i \le n$. From the last theorem, this is equivalent to X_i being independent of $X_1, X_2, \cdots, X_{i-1}$ for each i. Then

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

for all $x_1, x_2, \cdots, x_n$, i.e., $X_1, X_2, \cdots, X_n$ are mutually independent.

Alternatively, we can prove the theorem by considering

$$\begin{aligned} \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \cdots, X_n) &= -\sum_{i=1}^{n} E \log p(X_i) + E \log p(X_1, X_2, \cdots, X_n) \\ &= -E \log [p(X_1)\, p(X_2) \cdots p(X_n)] + E \log p(X_1, X_2, \cdots, X_n) \\ &= E \log \frac{p(X_1, X_2, \cdots, X_n)}{p(X_1)\, p(X_2) \cdots p(X_n)} \\ &= D(p_{X_1 X_2 \cdots X_n} \| p_{X_1} p_{X_2} \cdots p_{X_n}) \\ &\ge 0 \end{aligned}$$

where equality holds if and only if

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

for all $x_1, x_2, \cdots, x_n$, i.e., $X_1, X_2, \cdots, X_n$ are mutually independent.

Theorem 2.40 By Chain Rule for Mutual Information

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

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

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

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

and

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

Proof.

$$\begin{aligned} I(X;Z) &= I(X;Y,Z) - I(X;Y \mid Z) && \text{(Chain Rule)} \\ &\leq I(X;Y,Z) \\ &= I(X;Y) + I(X;Z \mid Y) && (I(X;Z \mid Y) = 0 \text{ by Markov Chain}) \\ &= I(X;Y) \end{aligned}$$

since $X \to Y \to Z$ is equivalent to $Z \to Y \to X$, $I(X;Z) \leq I(Y;Z)$ also proved.

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

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

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. $I(U;V) \leq I(U;Y) \leq I(X;Y)$ by Lemma 2.41

Theorem 2.43 For any random variable X,

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

where $|\mathcal{X}|$ denotes the size of the alphabet $\mathcal{X}$. The bound tight iff X is distributed uniformly on $\mathcal{X}$.

Proof. Let u be the uniform distribution on $\mathcal{X}$, i.e., $u(x) = |\mathcal{X}|^{-1}$ for all $x \in \mathcal{X}$. Then

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

This upper bound is tight iff only $D(p\|u) = 0$, which from Theorem 2.31 is equivalent to $p(x) = u(x)$ for all $x \in \mathcal{X}$, completing the proof. □

Corollary 2.44 The entropy of a random variable may take any nonnegative real value.

Proof. For any value $0 < b < \log |X|$, by the intermediate value theorem of continuous functions, there exists a distribution for X such that $H(X) = b$. Then we see that $H(X)$ can take any positive value by letting $|X|$ be sufficiently large.

Remark Let $|\mathcal{X}| = D$, or the random variable X is a D -ary symbol. When the base of the logarithm is D, (2.162) becomes

$$H_D(X) \leq 1$$

Recall that the unit of entropy is the D-it when the logarithm is in the base D. This inequality says that a D-ary symbol can carry at most 1 D-it of information. This maximum is achieved when X has a uniform distribution.

Remark The entropy of a random variable

- is finite if its alphabet is finite.
- can be finite or infinite if its alphabet is finite (see Examples 2.45 and 2.46)

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 h_b is the binary entropy function.

Proof. Define a random variable

$$Y = \begin{cases} 0 \text{ if } X = \hat{X} \\ 1 \text{ if } X \neq \hat{X} \end{cases}$$

The random variable Y is an indicator of the error event $\{X \neq \hat{X}\}$, with $\Pr\{Y = 1\} = P_e$ and $H(Y) = h_b(P_e)$. since Y is a function X and $\hat{X}$

$$H(Y \mid X, \hat{X}) = 0$$

Then

$$\begin{aligned}
H(X \mid \hat{X}) &= H(X \mid \hat{X}) + H(Y \mid X, \hat{X}) \\
&= H(X, Y \mid \hat{X}) \\
&= H(Y \mid \hat{X}) + H(X \mid \hat{X}, Y) \\
&\leq H(Y) + H(X \mid \hat{X}, Y) \\
&= H(Y) + \sum_{\hat{x} \in \mathcal{X}} [\Pr\{\hat{X} = \hat{x}, Y = 0\} H(X \mid \hat{X} = \hat{x}, Y = 0) + \Pr\{\hat{X} = \hat{x}, Y = 1\} H(X \mid \hat{X} = \hat{x}, Y = 1)] \\
&\leq h_b(P_e) + \left(\sum_{\hat{x} \in \mathcal{X}} \Pr\{\hat{X} = \hat{x}, Y = 1\} \right) \log(|\mathcal{X}| - 1) \\
&= h_b(P_e) + \Pr\{Y = 1\} \log(|\mathcal{X}| - 1) \\
&= h_b(P_e) + P_e \log(|\mathcal{X}| - 1)
\end{aligned}$$

- For <finite> alphabet, if $P_e \to 0$, then $H(X \mid \hat{X}) \to 0$
- This may NOT hold for <countably infinite> alphabet (see Example 2.49).

Corollary 2.48

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

3 The I-Measure

3.1 Preliminaries

Example 1.

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

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.

Definition 3.1 (Field) 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 (Atoms) The atoms of $\mathcal{F}_n$ are sets of the form

$$\cap_{i=1}^{n} Y_i,$$

where Y_i is either $\tilde{X}_i$ or $\tilde{X}_i^c$, the complement of $\tilde{X}_i$

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 $\mu(\emptyset) = 0$

Example 3.5

- A signed measure μ on $\mathcal{F}_2$ is completely specified by the values on the atoms

$$\mu\left(\tilde{X}_1 \cap \tilde{X}_2\right), \mu\left(\tilde{X}_1^c \cap \tilde{X}_2\right), \mu\left(\tilde{X}_1 \cap \tilde{X}_2^c\right), \mu\left(\tilde{X}_1^c \cap \tilde{X}_2^c\right)$$

- The value of μ on other sets in $\mathcal{F}_2$ are obtained by set-additivity.

3.2 The I-Measure for Two Random Variables

3.3 Construction of the I-Measure μ^*

- Let $\tilde{X}$ be a set corresponding to a r.v. X.
- $\mathcal{N}_n = \{1, 2, \cdots, n\}$
- Universal set to be the union of the sets $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_n$,

$$\Omega = \bigcup_{i \in \mathcal{N}_n} \tilde{X}_i$$

- Empty atom of $\mathcal{F}_n$

$$A_0 = \bigcap_{i \in \mathcal{N}_n} \tilde{X}_i^c$$

- $\mathcal{A}$ is the set of other atoms of $\mathcal{F}_n$, called non-empty atoms. $|\mathcal{A}| = 2^n - 1$
- A signed measure μ on $\mathcal{F}_n$ is completely specified by the values of μ on the nonempty atoms of $\mathcal{F}_n$.
- Notations: For nonempty subset G of $\mathcal{N}_n$:

$$X_G = (X_i, i \in G) \qquad \tilde{X}_G = \cup_{i \in G} \tilde{X}_i$$

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.

Lemma 3.7

$$\mu(A \cap B - C) = \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C)$$

Lemma 3.8

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

Construct the I -Measure μ^* on $\mathcal{F}_n$ using Theorem 3.6 by defining

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

for all nonempty subsets G of $\mathcal{N}_n$. In order for μ^* to be meaningful, it has to be consistent with all Shannon's information measures (via the substitution of symbols in (3.19)). In that case, the following must hold for all (not necessarily disjoint) subsets G, G', G'' of $\mathcal{N}_n$ where G and G' are nonempty:

$$\mu^* \left(\tilde{X}_G \cap \tilde{X}_{G'} - \tilde{X}_{G''} \right) = I\left(X_G; X_{G'} \mid X_{G''}\right)$$

When $G'' = \emptyset$, (3.41) becomes

$$\mu^* \left(\tilde{X}_G \cap \tilde{X}_{G'} \right) = I\left(X_G; X_{G'}\right)$$

When $G = G'$, (3.41) becomes

$$\mu^* \left(\tilde{X}_G - \tilde{X}_{G''} \right) = H\left(X_G \mid X_{G''}\right)$$

When $G = G'$ and $G'' = \emptyset$, (3.41) becomes

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

Thus (3.41) covers all the four cases of Shannon's information measures, and it is the necessary and sufficient condition for μ^* to be consistent with all Shannon's information measures.

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

- 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.

3.4 μ^* Can Be Negative

For $n = 3$, $\mu^* \left(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \right)$ does not correspond to a Shannon's information measure. $\mu^* \left(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \right)$ can actually be negative.

Example 3.10 Let all entropies be in the base 2. Let X_1 and X_2 be independent binary random variables with

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

Let

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

It is easy to check that X_3 has the same marginal distribution as X_1 and X_2. Thus,

$$H(X_i) = 1$$

for $i = 1, 2, 3$. Moreover, X_1, X_2, and X_3 are pairwise independent. Therefore,

$$H(X_i, X_j) = 2$$

and

$$I(X_i; X_j) = 0$$

for $1 \le i < j \le 3$. We further see from, $X_3 = (X_1 + X_2) \bmod 2$, that each random variable is a function of the other two random variables. Then by the chain rule for entropy, we have

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1, X_2) + H(X_3 \mid X_1, X_2) \\ &= 2 + 0 = 2 \end{aligned}$$

Now for $1 \le i < j < k \le 3$

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

where we have invoked Lemma 3.8. It then follows that

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

Thus μ^* takes a negative value on the atom $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3$.
The mutual information,

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

For this example, $I(X_1; X_2; X_3) < 0$, which implies

$$I(X_1; X_2 \mid X_3) > I(X_1; X_2)$$

Therefore, unlike entropy, the mutual information between two random variables can be increased by conditioning on a third random variable.

3.5 Information Diagrams

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 $\mathcal{F}_n$.

3.6 Examples of Applications

Example 3.12 (Concavity of Entropy) Let $X_1 \sim p_1(x)$ and $X_2 \sim p_2(x)$. Let

$$X \sim p(x) = \lambda p_1(x) + \bar{\lambda} p_2(x)$$

where $0 \le \lambda \le 1$ and $\bar{\lambda} = 1 - \lambda$.

$$H(X) \ge \lambda H(X_1) + \bar{\lambda} H(X_2)$$

$$
\begin{aligned}
H(X) &\geq H(X \mid Z) \\
&= \Pr\{Z=1\}H(X \mid Z=1) + \Pr\{Z=2\}H(X \mid Z=2) \\
&= \lambda H(X_1) + \bar{\lambda} H(X_2)
\end{aligned}
$$

Example 3.13 (Convexity of Mutual Information) Let

$$(X,Y) \sim p(x,y) = p(x)p(y \mid x)$$

For fixed $p(x), I(X;Y)$ is a convex functional of $p(y \mid x)$

$$
\begin{aligned}
I(X;Y) &= I(X;Y \mid Z) + I(X;Y;Z) \\
&= I(X;Y \mid Z) - a \\
&\leq I(X;Y \mid Z) \\
&= \Pr\{Z=1\}I(X;Y \mid Z=1) + \Pr\{Z=2\}I(X;Y \mid Z=2) \\
&= \lambda I(p(x), p_1(y \mid x)) + \bar{\lambda} I(p(x), p_2(y \mid x))
\end{aligned}
$$

Example 3.14 (Concavity of Mutual Information) Let

$$(X,Y) \sim p(x,y) = p(x)p(y \mid x)$$

For fixed $p(y \mid x), I(X;Y)$ is a concave functional of $p(x)$

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

4 Zero-Error Data Compression

4.1 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 (Uniquely Decodable Codes) 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.

In the next theorem, we prove that for any uniquely decodable code, the lengths of the codewords have to satisfy an inequality called the Kraft inequality.

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. Let N be an arbitrary positive integer, and consider

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

By collecting terms on the right-hand side, we write

$$\left(\sum_{k=1}^{m} D^{-l_k}\right)^N = \sum_{i=1}^{Nl_{\max}} A_i D^{-i}$$

where

$$l_{\max} = \max_{1\le k\le m} l_k$$

and A_i is the coefficient of D^{-i} in $\left(\sum_{k=1}^{m} D^{-l_k}\right)^N$. Now observe that A_i gives the total number of sequences of N codewords with a total length of i code symbols (see Example 4.5 below). since the code is uniquely decodable, these code sequences must be distinct, and therefore

$$A_i \le D^i$$

because there are D^i distinct sequences of i code symbols. Substituting this inequality into (4.4), we have

$$\left(\sum_{k=1}^{m} D^{-l_k}\right)^N \le \sum_{i=1}^{Nl_{\max}} 1 = Nl_{\max}$$

or

$$\sum_{k=1}^{m} D^{-l_k} \le (Nl_{\max})^{1/N}$$

Note that using L'hopital's Rule

$$\lim_{n\to\infty} n^{1/n} = 1$$

since this inequality holds for any N, upon letting $N \to \infty$, we obtain (4.2) completing the proof. □

Expected length of $\mathcal{C}$

$$L = \sum_i p_i l_i$$

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.,

$$L \ge H_D(X)$$

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

Proof.

$$\begin{aligned}
L - H_D(X) &= \sum_i p_i \log_D \left(p_i D^{l_i}\right) \\
&= (\ln D)^{-1} \sum_i p_i \ln \left(p_i D^{l_i}\right) \\
&\ge (\ln D)^{-1} \sum_i p_i \left(1 - \frac{1}{p_i D^{l_i}}\right) \\
&= (\ln D)^{-1} \left[\sum_i p_i - \sum_i D^{-l_i}\right] \\
&\ge (\ln D)^{-1}(1-1) \\
&= 0
\end{aligned}$$

where we have invoked the fundamental inequality in (4.19) and the Kraft inequality in (4.21). This proves (4.14). In order for this lower bound to be tight, both (4.19) and (4.21) have to be tight simultaneously. Now (4.19) is 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

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

i.e., (4.21) is also tight. This completes the proof of the theorem.

Corollary 4.7

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

Proof. Consider encoding each outcome of a random variable X by a distinct symbol in $\{1, 2, \cdots, |\mathcal{X}|\}$

$$H_{|\mathcal{X}|}(X) \leq 1$$

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, $R \geq 0$.

4.2 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.

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. We only 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 $l_1 \leq l_2 \leq \cdots \leq l_m$

The number of nodes which can be chosen as the $(i+1)$ st codeword is

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

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

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

Multiplying by $D^{l_{i+1}}$ and rearranging the terms, we have

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

The left-hand side is the number of nodes which can be chosen as the $(i+1)$ st codeword as given in (4.27). Therefore, it is possible to choose the $(i+1)$ st codeword. Thus we have shown the existence of a prefix code with codeword lengths $l_1, l_2, \cdots, l_m$, completing the proof. □

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

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.

- Assume $p_1 \geq p_2 \geq \cdots \geq p_m$
- Denote the codeword assigned to p_i by c_i, and denote its length by l_i.

Lemma 4.5 In an optimal code, shorter codewords are assigned to larger probabilities.

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.

Lemma 4.17 The Huffman procedure produces an optimal prefix code.

Theorem 4.18 The expected length of a Huffman code, denoted by L_{Huff} satisfies

$$L_{\text{Huff}} < H_D(X) + 1$$

Proof. We will construct a prefix code with expected length less than $H(X)+ 1$. Then, because a Huffman code is an optimal prefix code, its expected length L_{Huff} is upper bounded by $H(X) + 1$

Consider constructing a prefix code with codeword lengths $\{l_i\}$, where

$$l_i = \lceil -\log_D p_i \rceil$$

Then

$$-\log_D p_i \le l_i < -\log_D p_i + 1$$

or

$$p_i \ge D^{-l_i} > D^{-1} p_i$$

Thus

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

i.e., $\{l_i\}$ satisfies the Kraft inequality, which implies that it is possible to construct a prefix code with codeword lengths $\{l_i\}$.

It remains to show that L, the expected length of this code, is less than $H(X) + 1$. Toward this end, consider

$$\begin{aligned} L &= \sum_i p_i l_i \\ &< \sum_i p_i \left(-\log_D p_i + 1\right) \\ &= -\sum_i p_i \log_D p_i + \sum_i p_i \\ &= H(X) + 1 \end{aligned}$$

where (4.44) follows from the upper bound in (4.40). Thus we conclude that

$$L_{\text{Huft}} \le L < H(X) + 1$$

To see that this upper bound is the tightest possible, we have to show that there exists a sequence of distributions P_k such that L_{Huff} approaches $H(X) + 1$ as $k \to \infty$. This can be done by considering the sequence of D -ary distributions 4.3 Redundancy of Prefix Codes

$$P_k = \left\{1 - \frac{D-1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right\}$$

where $k \ge D$. The Huffman code for each P_k consists of D codewords of length 1. Thus L_{Hufl} is equal to 1 for all k. As $k \to \infty, H(X) \to 0$, and hence L_{Huff} approaches $H(X) + 1$. The theorem is proved. □

This bound is the tightest among all the upper bounds on L_{Huff} which depend only on the source entropy. Proof.

- Construct a code with codeword lengths $l_i = \lceil -\log_D p_i \rceil$ by showing that the Kraft inequality is satisfied.
- Show that $L = \sum_i p_i l_i < H(X) + 1$
- Then $L_{\text{Huff}} \le L < H(X) + 1$
- For tightness, consider $P_k = \left\{1 - \frac{D-1}{k}, \frac{1}{k}, \cdots, \frac{1}{k}\right\}$ and let $k \to \infty$.

Asymptotic Achievability of $H(X)$

- $$H(X) \leq L_{\text{Huff}} < H(X) + 1$$
- Use a Huffman code to encode $X_1, X_2, \cdots, X_n, n$ i.i.d. copies of X. Then
$$nH(X) \leq L^n_{\text{Huff}} < nH(X) + 1$$
- Divide by n to obtain
$$H(X) \leq \frac{1}{n} L^n_{\text{Huff}} < H(X) + \frac{1}{n} \to H(X) \text{ as } n \to \infty$$

4.3 Redundancy of Prefix Codes

5 Weak Typicality

Theorem 5.1 (Weak AEP I)
$$-\frac{1}{n} \log p(\mathbf{X}) \to H(X)$$
in probability as $n \to \infty$, i.e., for any $\epsilon > 0$, for n sufficiently large,
$$\Pr\left\{\left|-\frac{1}{n} \log p(\mathbf{X}) - H(X)\right| \leq \epsilon\right\} > 1 - \epsilon$$
Note: $X_n \to X$ in probability means that
$$\lim_{n\to\infty} \Pr\{|X_n - X| \geq \epsilon\} = 0$$
for all $\epsilon > 0$.

Definition 5.2 The weakly typical set $W^n_{[X]\epsilon}$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ such that
$$\left|-\frac{1}{n} \log p(\mathbf{x}) - H(X)\right| \leq \epsilon$$
or equivalently,
$$H(X) - \epsilon \leq -\frac{1}{n} \log p(\mathbf{x}) \leq H(X) + \epsilon$$
where ϵ is an arbitrarily small positive real number. The sequences in $W^n_{[X]\epsilon}$ are called weakly ϵ -typical sequences.

Empirical Entropy
$$-\frac{1}{n} \log p(\mathbf{x}) = -\frac{1}{n} \sum_{k=1}^{n} \log p(x_k)$$
is called the empirical entropy of the sequence $\mathbf{x}$.
The empirical entropy of a weakly typical sequence is close to the true entropy $H(X)$.

Theorem 5.2 (Weak AEP II) The following hold for any $\epsilon > 0$:
1) If $\mathbf{x} \in W^n_{[X]\epsilon}$, then

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

2) For n sufficiently large,

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

3) For n sufficiently large,

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \le \left|W^n_{[X]\epsilon}\right| \le 2^{n(H(X)+\epsilon)}$$

WAEP says that for large n

- the probability of occurrence of the sequence drawn is close to $2^{-nH(X)}$ with very high probability;
- the total number of weakly typical sequences is approximately equal to $2^{nH(X)}$

WAEP DOES NOT say that

- most of the sequences in $\mathcal{X}^n$ are weakly typical;
- the most likely sequence is weakly typical.

When n is large, one can almost think of the sequence $\mathbf{X}$ as being obtained by choosing a sequence from the weakly typical set according to the uniform distribution.

The Source Coding Theorem A block code: $\mathcal{X}^n \to \mathcal{I}$

- $\mathcal{I} = \{1, 2, \cdots, M\}$
- blocklength $= n$
- coding rate $= n^{-1}\log M$
- $P_e = \Pr\{\mathbf{X} \notin \mathcal{A}\}$

Direct part: For arbitrarily small P_e, there exists a block code whose coding rate is arbitrarily close to $H(X)$ when n is sufficiently large.

- Fix $\epsilon > 0$ and take $\mathcal{A} = W^n_{[X]\epsilon}$ and hence $M = |\mathcal{A}|$.
- For sufficiently large n, by WAEP,

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \le M = |\mathcal{A}| = \left|W^n_{[X]\epsilon}\right| \le 2^{n(H(X)+\epsilon)}$$

- Coding rate $R = n^{-1}\log M$ satisfies

$$\frac{1}{n}\log(1-\epsilon) + H(X) - \epsilon \le \frac{1}{n}\log M \le H(X) + \epsilon$$

- By WAEP,

$$P_e = \Pr\{\mathbf{X} \notin \mathcal{A}\} = \Pr\left\{\mathbf{X} \notin W^n_{[X]\epsilon}\right\} < \epsilon$$

- Letting $\epsilon \to 0$, the coding rate tends to $H(X)$, while P_e tends to 0 .

Converse: For any block code with block length n and coding rate less than $H(X) - \zeta$, where $\zeta > 0$ does not change with n, then $P_e \to 1$ as $n \to \infty$

- Consider any block code with rate less than $H(X) - \zeta$, where $\zeta > 0$ does not change with n. Then total number of codewords $\leq 2^{n(H(X)-\zeta)}$
- Use some indices to cover $\mathbf{x} \in W^n_{[X]\epsilon}$, and others to cover $\mathbf{x} \notin W^n_{[X]\epsilon}$
- Total probability of typical sequences covered is upper bounded by

$$2^{n(H(X)-\zeta)}2^{-n(H(X)-\epsilon)} = 2^{-n(\zeta-\epsilon)}$$

- Total probability covered is upper bounded by

$$2^{-n(\zeta-\epsilon)} + \Pr\left\{\mathbf{X} \notin W^n_{[X]\epsilon}\right\} < 2^{-n(\zeta-\epsilon)} + \epsilon$$

- Then $P_e > 1 - \left(2^{-n(\zeta-\epsilon)} + \epsilon\right)$ holds for any $\epsilon > 0$ and n sufficiently large.
- Take $\epsilon < \zeta$. Then $P_e > 1 - 2\epsilon$ for n sufficiently large.
- Finally, let $\epsilon \to 0$.

6 Strong Typicality

Setup

- $\{X_k, k \geq 1\}, X_k$ i.i.d. $\sim p(x)$
- X denotes generic r.v. with entropy $H(X) < \infty$.
- $|\mathcal{X}| < \infty$

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

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

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,

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

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

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

Proof. Chernoff bound.

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

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

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 $\delta \to 0$

Setup

- $\{(X_k, Y_k), k \geq 1\}, (X_k, Y_k)$ i.i.d. $\sim p(x, y)$
- (X, Y) denotes pair of generic r.v. with entropy $H(X, Y) < \infty$.
- $|\mathcal{X}|, |\mathcal{Y}| < \infty$

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 $\mathbf{y} \in T^n_{[Y]\delta}$

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 $1 \leq i \leq n$

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

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

2) For n sufficiently large,

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

3) For n sufficiently large,

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

Lemma 6.11 (simplified))(Stirling's Approximation)

$$\ln n! \sim n \ln n$$

Lemma For large n,

$$\binom{n}{np, n(1-p)} \approx 2^{nH_2(\{p,1-p\})}$$

Theorem 6.10 (Conditional Strong AEP) For any $x \in T^n_{[X]\delta}$, 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 $\left|T^n_{[Y|X]\delta}(\mathbf{x})\right| \geq 1$, then

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

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

Remark Weak Typicality guarantees that the number of $\mathbf{y}$ that are jointly typical with a typical $\mathbf{x}$ is approximately equal to $2^{n(H(Y|X)}$ on the average. Strong typicality guarantees that this is so for each typical $\mathbf{x}$ as long as there exists at least one $\mathbf{y}$ that is jointly typical with $\mathbf{x}$.

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

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

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

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

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

for n sufficiently large.

7 Discrete Memoryless Channels

Definition 7.1 (Discrete Channel I) 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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Definition 7.2 (Discrete Channel II) 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

$$Y = \alpha(X, Z)$$

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 7.4 (DMC I) A discrete memoryless channel (DMC) $p(y \mid x)$ is a sequence of replicates of a generic discrete channel $p(y \mid x)$. These discrete channels are indexed by a discrete-time index i, where $i \geq 1$, with the i th 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 equality

$$\Pr\{Y_i = y, X_i = x, T_{i-} = t\} = \Pr\{X_i = x, T_{i-} = t\}\, p(y \mid x)$$

holds for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{T}_{i-}$

Remark:

$$T_{i-} \to X_i \to Y_i,$$

or Given X_i, Y_i is independent of everything in the past.

Definition 7.5 (DMC II) 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 i th 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

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

Remark: The equivalence of Definitions 7.4 and 7.5 can be shown. See textbook.

$$\begin{aligned}
&\Pr\{Y_i = y, X_i = x, T_{i-} = t\} \\
&\overset{a)}{=} \Pr\{X_i = x, T_{i-} = t\} \Pr\{Y_i = y \mid X_i = x\} \\
&\overset{b)}{=} \Pr\{X_i = x, T_{i-} = t\} \Pr\{\alpha(X_i, Z_i) = y \mid X_i = x\} \\
&= \Pr\{X_i = x, T_{i-} = t\} \Pr\{\alpha(x, Z_i) = y \mid X_i = x\} \\
&\overset{c)}{=} \Pr\{X_i = x, T_{i-} = t\} \Pr\{\alpha(x, Z_i) = y\} \\
&\overset{d)}{=} \Pr\{X_i = x, T_{i-} = t\} \Pr\{\alpha(x, Z) = y\} \\
&\overset{e}{=} \Pr\{X_i = x, T_{i-} = t\}\, p(y \mid x)
\end{aligned}$$

where

- (a) follows from the Markov chain $T_{i-} \to X_i \to Y_i$
- (b) follows from (7.39)
- (c) follows from Definition 7.5 that Z_i is independent of X_i
- (d) follows from Definition 7.5 that Z_i and the generic noise variable Z have the same distribution;
- (e) follows from (7.34)

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

$$C = \max_{p(x)} I(X; Y) \geq 0$$

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 $p(x)$.

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

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

Remarks:

- since $I(X;Y)$ is a continuous functional of $p(x)$ and the set of all $p(x)$ is a compact set (i.e., closed and bounded) in $\Re^{|\mathcal{X}|}$, the maximum value of $I(X;Y)$ can be attained.
- Will see that C is in fact the maximum rate at which information can be communicated reliably through a DMC.
- Can communicate through a channel at a positive rate while $P_e \to 0$!

Example 7.7 (Binary Symmetric Channel)

$$Y = X + Z \bmod 2$$

This representation for a BSC is in the form prescribed by Definition 7.2 . In order to determine the capacity of the BSC, we first bound $I(X;Y)$ as follows:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= H(Y) - \sum_x p(x) H(Y \mid X = x) \\ &= H(Y) - \sum_x p(x) h_b(\epsilon) \\ &= H(Y) - h_b(\epsilon) \\ &\le 1 - h_b(\epsilon) \end{aligned}$$

where we have used h_b to denote the binary entropy function in the base 2 . In order to achieve this upper bound, we have to make $H(Y) = 1$, i.e., the output distribution of the BSC is uniform. This can be done by letting $p(x)$ be the uniform distribution on {0,1} . Therefore, the upper bound on $I(X;Y)$ can be achieved, and we conclude that

$$C = 1 - h_b(\epsilon) \text{ bit per use}$$

When $\epsilon = 0.5$, the channel output is independent of the channel input. Therefore, no information can possibly be communicated through the channel.

Example 7.8 (Binary Erasure Channel)

7.1 The Channel Coding Theorem

Direct Part Information can be communicated through a DMC with an arbitrarily small probability of error at any rate less than the channel capacity.

Converse If information is communicated through a DMC at a rate higher than the capacity, then the probability of error is bounded away from zero.

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

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

and a decoding function

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

- **Message Set** $\mathcal{W} = \{1, 2, \cdots, M\}$
- **Codewords** $f(1), f(2), \cdots, f(M)$
- **Codebook** The set of all codewords.
- W is randomly chosen from the message set $\mathcal{W}$, so $H(W) = \log M$.
- $\mathbf{X} = (X_1, X_2, \cdots, X_n)\,; \mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$
- Thus $\mathbf{X} = f(W)$
- Let $\hat{W} = g(\mathbf{Y})$ be the estimate on the message W by the decoder.

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 **vs** $\lambda_{\max}$

$$\begin{aligned} P_e &= \Pr\{\hat{W} \neq W\} \\ &= \sum_w \Pr\{W = w\} \Pr\{\hat{W} \neq W \mid W = w\} \\ &= \sum_w \frac{1}{M} \Pr\{\hat{W} \neq w \mid W = w\} \\ &= \frac{1}{M} \sum_w \lambda_w \end{aligned}$$

Therefore, $P_e \leq \lambda_{\max}$

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

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

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.

- The communication system consists of the r.v.'s

  $$W, X_1, Y_1, X_2, Y_2, \cdots, X_n, Y_n, \hat{W}$$

  generated in this order.
- The memorylessness of the DMC imposes the following Markov constraint for each i :

  $$(W, X_1, Y_1, \cdots, X_{i-1}, Y_{i-1}) \to X_i \to Y_i$$
- The dependency graph can be composed accordingly.
- Use q to denote the joint distribution and marginal distributions of all r.v.'s.
- 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 $q(\mathbf{y}) > 0$

  $$q(w, \mathbf{x}, \mathbf{y}\hat{w}) = q(w) \left(\prod_{i=1}^n q\left(x_i \mid w\right) \right) \left(\prod_{i=1}^n p\left(y_i \mid x_i\right) \right) q(\hat{w} \mid \mathbf{y})$$
- $q(w) > 0$ for all w so that $q\left(x_i \mid w\right)$ are well-defined. $q\left(x_i \mid w\right)$ and $q(\hat{w} \mid \mathbf{y})$ are deterministic.
- The dependency graph suggests the Markov chain $W \to \mathbf{X} \to \mathbf{Y} \to \hat{W}$.
- This can be formally justified by invoking Proposition 2.9 .

Why C is related to $I(X;Y)$?

- $H(\mathbf{X} \mid W) = 0$
- $H(\hat{W} \mid \mathbf{Y}) = 0$
- since W and $\hat{W}$ are essentially identical for reliable communication, as- sume

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

- Then from the information diagram for $W \to \mathbf{X} \to \mathbf{Y} \to \hat{W}$, we see that

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

- This suggests that the channel capacity is obtained by maximizing $I(X;Y)$.

Lemma 7.16 $I(\mathbf{X}; \mathbf{Y}) \le \sum_{i=1}^{n} I(X_i; Y_i)$
Proof.
1. Establish

$$H(\mathbf{Y} \mid \mathbf{X}) = \sum_{i=1}^{n} H(Y_i \mid X_i)$$

2 .

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y} \mid \mathbf{X}) \\ &\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i \mid X_i) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \end{aligned}$$

Building Blocks of the Converse

- For all $1 \le i \le n$

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

- Then

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

- To be established in Lemma 7.16 ,

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

- Therefore,

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

Converse (Formal Proof)

1. Let R be an achievable rate, i.e., for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code such that

$$\frac{1}{n} \log M > R - \epsilon \quad \text{and} \quad \lambda_{\max} < \epsilon$$

2. Consider

$$\begin{aligned} \log M &\overset{a)}{=} H(W) \\ &= H(W \mid \hat{W}) + I(W; \hat{W}) \\ &\leq H(W \mid \hat{W}) + I(\mathbf{X}; \mathbf{Y}) \\ &\leq H(W \mid \hat{W}) + \sum_{i=1}^{n} I(X_i; Y_i) \\ &\leq H(W \mid \hat{W}) + nC \end{aligned}$$

3. By Fano's inequality,

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

4. Then,

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

Therefore,

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

5. Letting $n \to \infty$ and then $\epsilon \to 0$ to conclude that $R \leq C$.

- For large n,

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

- $\frac{1}{n} \log M$ is the actual rate of the channel code.
- If $\frac{1}{n} \log M > C$, then $P_e > 0$ for large n.
- This implies that if $\frac{1}{n} \log M > C$, then $P_e > 0$ for all n.
- If there exists an $\epsilon > 0$ such that $\frac{1}{n} \log M \geq C + \epsilon$ for all n, then $P_e \to 1$ as $n \to \infty$.

Achievability

- Consider a DMC $p(y \mid x)$.
- For every input distribution $p(x)$, prove that the rate $I(X; Y)$ is achievable by showing for large n the existence of a channel code such that
  1. the rate of the code is arbitrarily close to $I(X; Y)$;
  2. the maximal probability of error $\lambda_{\max}$ is arbitrarily small.
- Choose the input distribution $p(x)$ to be one that achieves the channel capacity, i.e., $I(X; Y) = C$

Lemma 7.17 Let $(\mathbf{X}', \mathbf{Y}')$ be n i.i.d. copies of a pair of generic random variables (X', Y'), where X' and Y' are independent and have the same marginal distributions as X and Y, respectively. Then

$$\Pr\left\{(\mathbf{X}', \mathbf{Y}') \in T^n_{[XY]\delta}\right\} \le 2^{-n(I(X;Y)-\tau)}$$

where $\tau \to 0$ as $\delta \to 0$.

Proof.

- Consider
$$\Pr\left\{(\mathbf{X}', \mathbf{Y}') \in T^n_{[XY]\delta}\right\} = \sum_{(\mathbf{x},\mathbf{y}) \in T^n_{[XY]\delta}} p(\mathbf{x})p(\mathbf{y})$$
- Consistency of strong typicality: $\mathbf{x} \in T^n_{[X]\delta}$ and $\mathbf{y} \in T^n_{[Y]\delta}$ • Strong AEP: $p(\mathbf{x}) \le 2^{-n(H(X)-\eta)}$ and $p(\mathbf{y}) \le 2^{-n(H(Y)-\zeta)}$
- Strong JAEP: $\left|T^n_{[XY]\delta}\right| \le 2^{n(H(X,Y)+\xi)}$
- Then
$$\begin{aligned}\Pr\left\{(\mathbf{X}', \mathbf{Y}') \in T^n_{[XY]\delta}\right\} & \\ \le & 2^{n(H(X,Y)+\xi)} \cdot 2^{-n(H(X)-\eta)} \cdot 2^{-n(H(Y)-\zeta)} \\ = & 2^{-n(H(X)+H(Y)-H(X,Y)-\xi-\eta-\zeta)} \\ = & 2^{-n(I(X;Y)-\xi-\eta-\zeta)} \\ = & 2^{-n(I(X;Y)-\tau)}\end{aligned}$$

Interpretation of Lemma 7.17

- Randomly choose a row with uniform distribution and randomly choose a column with uniform distribution.
- $Pr\{$ Obtaining a jointly typical pair $\} \approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI((X;Y)}$

Random Coding Scheme

1. Construct the codebook $\mathcal{C}$ of an (n, M) code by generating M codewords in $\mathcal{X}^n$ independently and identically according to $p(x)^n$. Denote these codewords by $\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \cdots, \tilde{\mathbf{X}}(M)$
2. Reveal the codebook $\mathcal{C}$ to both the encoder and the decoder.
3. A message W is chosen from $\mathcal{W}$ according to the uniform distribution.
4. Transmit $\mathbf{X} = \tilde{\mathbf{X}}(W)$ through the channel.
5. The channel outputs a sequence $\mathbf{Y}$ according to
$$\Pr\{\mathbf{Y} = \mathbf{y} \mid \tilde{\mathbf{X}}(W) = \mathbf{x}\} = \prod_{i=1}^{n} p\left(y_i \mid x_i\right)$$
6. The sequence $\mathbf{Y}$ is decoded to the message w if
   - $(\tilde{\mathbf{X}}(w), \mathbf{Y}) \in T^n_{[XY]\delta}$, and
   - there does not exists $w' \ne w$ such that $\left(\tilde{\mathbf{X}}\left(w'\right), \mathbf{Y}\right) \in T^n_{[XY]\delta}$ Otherwise, $\mathbf{Y}$ is decoded to a constant message in $\mathcal{W}$. Denote by $\hat{W}$ the message to which $\mathbf{Y}$ is decoded.

   Otherwise, $\mathbf{Y}$ is decoded to a constant message in $\mathcal{W}$. Denote by $\hat{W}$ the message to which Y is decoded.

Definition 7.18 An (n, M) code with complete feedback for a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ is defined by encoding functions

$$f_i : \{1, 2, \cdots, M\} \times \mathcal{Y}^{i-1} \to \mathcal{X}$$

for $1 \leq i \leq n$ and a decoding function

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

Notations: $\mathbf{Y}^i = (Y_1, Y_2, \cdots, Y_i), X_i = f_i\left(W, \mathbf{Y}^{i-1}\right)$

Definition 7.19 A rate R is achievable with complete feedback for a discrete memoryless channel $p(y \mid x)$ if for any $\epsilon > 0$, there exists for sufficiently large n an (n, M) code with complete feedback such that

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

and

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

Definition 7.20 The feedback capacity, C_{FB}, of a discrete memoryless channel is the supremum of all the rates achievable by codes with complete feedback.

Proposition 7.21 The supremum in the definition of C_{FB} in Definition 7.20 is the maximum.

Lemma 7.22 For all $1 \leq i \leq n$

$$\left(W, \mathbf{Y}^{i-1}\right) \to X_i \to Y_i$$

forms a Markov chain.