

1 Introduction

2 Information Measures

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, \dots, X_n are mutually independent if

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

for all x_1, x_2, \dots, x_n

Definition 2.3 (Pairwise Independence) For $n \geq 3$, random variables X_1, X_2, \dots, 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, \dots, X_n where $n \geq 3$, $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ forms a Markov chain if

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

for all x_1, x_2, \dots, x_n , or equivalently,

$$p(x_1, x_2, \dots, x_n) = \begin{cases} p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1}) & \text{if } p(x_1), p(x_2|x_1), \dots, p(x_{n-1}|x_{n-2}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

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

Proposition 2.8 $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ forms a Markov chain if and only if

$$\begin{aligned} &X_1 \rightarrow X_2 \rightarrow X_3 \\ &(X_1, X_2) \rightarrow X_3 \rightarrow X_4 \\ &\vdots \\ &(X_1, X_2, \dots, X_{n-2}) \rightarrow X_{n-1} \rightarrow X_n \end{aligned}$$

form Markov chains.

Proposition 2.9 $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 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 $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 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 $k_j \in \alpha_j, j = 1, 2, \cdots, m$

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

forms a Markov chain. That is, a subchain of $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ is also a Markov chain.

Example 2.11 Let $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_{10}$ form a Markov chain and $\alpha_1 = \{1, 2\}, \alpha_2 = \{4\}, \alpha_3 = \{6, 8\}$, and $\alpha_4 = \{10\}$ be subsets of \mathcal{N}_{10} . Then Proposition 2.10 says that

$$(X_1, X_2) \rightarrow X_4 \rightarrow (X_6, X_8) \rightarrow X_{10}$$

also forms 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 $X_1 \perp X_4 \mid (X_2, X_3), X_1 \perp X_3 \mid (X_2, X_4)$, 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 $x_4 \neq (x_1, x_2)$

Definition 2.13 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 $H(p)$
- A bit in information theory is different from a bit in computer science.
- Convention

$$Eg(X) = \sum_x p(x)g(x)$$

where summation is over \mathcal{S}_X

- linearity

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

- Can write

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

- In probability theory, when $Eg(X)$ is considered, usually $g(x)$ depends only on the value of x but not on $p(x)$.

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 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 For random variables X and Y , the conditional entropy of Y given X is defined as

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

- Write

$$H(Y | X) = \sum_x p(x) \left[- \sum_y p(y | x) \log p(y | x) \right]$$

- The inner sum is the entropy of Y conditioning on a fixed $x \in \mathcal{S}_X$, denoted by $H(Y | X = x)$
- Thus

$$H(Y | X) = \sum_x p(x) H(Y | X = x)$$

- Similarly,

$$H(Y | X, Z) = \sum_z p(z) H(Y | X, Z = z)$$

where

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

Proposition 2.16

$$H(X, Y) = H(X) + H(Y | X)$$

and

$$H(X, Y) = H(Y) + H(X | Y)$$

Proof.

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

Definition 2.17 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

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

and

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

provided that all the entropies and conditional entropies are finite.

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

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

Remark: $I(X; Y | Z)$ is symmetrical in X and Y .

Similar to entropy, we have

$$I(X; Y | Z) = \sum_z p(z) I(X; Y | Z = z)$$

where

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

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., $I(X; X | Z) = H(X | Z)$

Proposition 2.22

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

and

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

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.

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

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

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

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

Proposition 2.24 (Chain Rule for Entropy)

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

Proposition 2.25 (Chain Rule for Conditional Entropy)

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

Proof (1).

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

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

Proof (2).

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

Proposition 2.26 (Chain Rule for Mutual Information)

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

Proposition 2.27 (Chain Rule for Conditional Mutual Information)

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

Definition 2.28 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 2.31 (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 $p = q$

Theorem 2.32 (Log-Sum Inequality) For positive numbers a_1, a_2, \dots and nonnegative numbers b_1, b_2, \dots 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 $\frac{a_i}{b_i} = \text{constant}$ 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.
- The log-sum inequality also implies the divergence inequality.
- The two inequalities 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 $V(p, q)$.
- For a sequence of probability distributions q_k , as $k \rightarrow \infty$, if $D(p\|q_k) \rightarrow 0$ or $D(q_k\|p) \rightarrow 0$, then $V(p, q_k) \rightarrow 0$
- That is, "convergence in divergence" is a stronger notion than "convergence in variational distance."

Theorem 2.34 (The Basic Inequalities) For random variables X, Y , and Z ,

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

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.

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

Proposition 2.36 $H(Y | X) = 0$ 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.

Theorem 2.38 (Conditioning Does Not Increase Entropy)

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

with equality if and only if X and Y are independent. • Similarly, $H(Y | X, Z) \leq H(Y | Z)$.

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

Theorem 2.39 (Independence Bound for Entropy)

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

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

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

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

and

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

Corollary

- If $X \rightarrow Y \rightarrow Z$, then

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

- 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 $U \rightarrow X \rightarrow Y \rightarrow V$ forms a Markov chain, then

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

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} .

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

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

$$H(X | \hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1)$$

where h_b is the binary entropy function.

Corollary 2.48

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

Interpretation

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

3 The I -Measure

Example 1.

$$\mu^*(\tilde{X}_1 - \tilde{X}_2) = H(X_1 | X_2)$$

$$\mu^*(\tilde{X}_2 - \tilde{X}_1) = H(X_2 | X_1)$$

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

2. Inclusion-Exclusion formulation in set-theory

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

corresponds to

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

in information theory.

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

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 \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

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

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

Construction of the I -Measure μ^*

- Let \tilde{X} be a set corresponding to a r.v. X .
- $\mathcal{N}_n = \{1, 2, \dots, n\}$
- Universal set

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

- 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 :

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

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

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

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

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

for all nonempty subsets G of \mathcal{N}_n .

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.

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

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 $0 \leq \lambda \leq 1$ and $\bar{\lambda} = 1 - \lambda$. Show that

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

Example 3.13 (Convexity of Mutual Information) Let

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

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

Example 3.14 (Concavity of Mutual Information) Let

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

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

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

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

4 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.

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

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

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.,

$$L \geq H_D(X)$$

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

Corollary 4.7

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

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, $R \geq 0$.

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, \dots, 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.

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 $p_1 \geq p_2 \geq \dots \geq p_m$
- 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

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

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 $L_{\text{Huff}} \leq L < H(X) + 1$
- For tightness, consider $P_k = \{1 - \frac{D-1}{k}, \frac{1}{k}, \dots, \frac{1}{k}\}$ and let $k \rightarrow \infty$.

Asymptotic Achievability of $H(X)$

•

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

- Use a Huffman code to encode X_1, X_2, \dots, X_n, n i.i.d. copies of X . Then

$$nH(X) \leq L_{\text{Huff}}^n < nH(X) + 1$$

- Divide by n to obtain

$$H(X) \leq \frac{1}{n} L_{\text{Huff}}^n < H(X) + \frac{1}{n} \rightarrow H(X) \text{ as } n \rightarrow \infty$$

5 Weak Typicality

Theorem 5.1 (Weak AEP I)

$$-\frac{1}{n} \log p(\mathbf{X}) \rightarrow H(X)$$

in probability as $n \rightarrow \infty$, 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: $X_n \rightarrow X$ in probability means that

$$\lim_{n \rightarrow \infty} \Pr \{ |X_n - X| \geq \epsilon \} = 0$$

for all $\epsilon > 0$.

Definition 5.2 The weakly typical set $W_{[X]\epsilon}^n$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ 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 $W_{[X]\epsilon}^n$ 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 $\mathbf{x} \in W_{[X]\epsilon}^n$, then

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

2) For n sufficiently large,

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

3) For n sufficiently large,

$$(1 - \epsilon) 2^{n(H(X)-\epsilon)} \leq |W_{[X]\epsilon}^n| \leq 2^{n(H(X)+\epsilon)}$$

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: $\mathcal{X}^n \rightarrow \mathcal{I}$

- $\mathcal{I} = \{1, 2, \dots, M\}$
- 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 $\mathcal{A} = W_{[X]\epsilon}^n$ and hence $M = |\mathcal{A}|$.
- For sufficiently large n , by WAEP,

$$(1 - \epsilon)2^{n(H(X) - \epsilon)} \leq M = |\mathcal{A}| = |W_{[X]\epsilon}^n| \leq 2^{n(H(X) + \epsilon)}$$

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

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

- By WAEP,

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

- Letting $\epsilon \rightarrow 0$, 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 $P_e \rightarrow 1$ as $n \rightarrow \infty$

- 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 $\mathbf{x} \in W_{[X]\epsilon}^n$, and others to cover $\mathbf{x} \notin W_{[X]\epsilon}^n$
- 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

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

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

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 $T_{[X]\delta}^n$ with respect to $p(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ 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 $T_{[X]\delta}^n$ are called strongly δ -typical sequences.

Theorem 6.2 (Strong AEP) There exists $\eta > 0$ such that $\eta \rightarrow 0$ as $\delta \rightarrow 0$ and the following hold:

1) If $\mathbf{x} \in T_{[X]\delta}^n$, then

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

2) For n sufficiently large,

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

3) For n sufficiently large,

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

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

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

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 ,

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

and

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

Proposition 6.5 For any $\mathbf{x} \in \mathcal{X}^n$, if $\mathbf{x} \in T_{[X]\delta}^n$, then $\mathbf{x} \in W_{[X]\eta}^n$, where $\eta \rightarrow 0$ as $\delta \rightarrow 0$

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 $T_{[XY]\delta}^n$ 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 $T_{[XY]\delta}^n$

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

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

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \in T_{[X]\delta}^n$$

then

$$f(\mathbf{x}) = (y_1, y_2, \dots, y_n) \in T_{[Y]\delta}^n$$

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

Theorem 6.9 (Strong JAEP) Let

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

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

1) If $(\mathbf{x}, \mathbf{y}) \in T_{[XY]\delta}^n$, 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,

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

3) For n sufficiently large,

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

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 $x \in T_{[X]\delta}^n$, define

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

If $|T_{[Y|X]\delta}^n(\mathbf{x})| \geq 1$, then

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

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

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 $S_{[X]\delta}^n$ be the set of all sequences $\mathbf{x} \in T_{[X]\delta}^n$ such that $T_{[Y|X]\delta}^n(\mathbf{x})$ is nonempty. Then

$$|S_{[X]\delta}^n| \geq (1 - \delta)2^{n(H(X)-\psi)}$$

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

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

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

for n sufficiently large.

7 Discrete Memoryless Channels

Definition 7.1 (Discrete Channel I) Let \mathcal{X} and \mathcal{Y} be discrete alphabets, and $p(y | x)$ be a transition matrix from \mathcal{X} to \mathcal{Y} . A discrete channel $p(y | x)$ 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

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

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 $\alpha : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{Y}$, 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 $p(y | x)$ and (α, Z) defined on the same input alphabet \mathcal{X} and output alphabet \mathcal{Y} are equivalent if

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

for all x and y

Definition 7.4 (DMC I) A discrete memoryless channel (DMC) $p(y | x)$ is a sequence of replicates of a generic discrete channel $p(y | x)$. 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

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

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

Remark: $T_{i-} \rightarrow X_i \rightarrow Y_i$, or Given X_i, Y_i is independent of everything in the past.

Definition 7.4 (DMC I) A discrete memoryless channel (DMC) $p(y | x)$ is a sequence of replicates of a generic discrete channel $p(y | x)$. 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

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

holds for all $(x, y, t) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{T}_{i-}$ Remark: $T_{i-} \rightarrow X_i \rightarrow Y_i$, 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.

Definition 7.6 The capacity of a discrete memoryless channel $p(y | x)$ 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 $p(x)$.

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 $\mathbb{R}^{|\mathcal{X}|}$, 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 $P_e \rightarrow 0$!

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

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

and a decoding function

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

- **Message Set** $\mathcal{W} = \{1, 2, \dots, M\}$
- **Codewords** $f(1), f(2), \dots, f(M)$
- **Codebook** The set of all codewords.
- W is randomly chosen from the message set \mathcal{W} , so $H(W) = \log M$.
- $\mathbf{X} = (X_1, X_2, \dots, X_n); \mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$
- 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 λ_{\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

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

generated in this order.

- The memorylessness of the DMC imposes the following Markov constraint for each i :

$$(W, X_1, Y_1, \dots, X_{i-1}, Y_{i-1}) \rightarrow X_i \rightarrow Y_i$$

- 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 $q(\mathbf{y}) > 0$

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

- $q(w) > 0$ for all w so that $q(x_i \mid w)$ are well-defined. $q(x_i \mid w)$ and $q(\hat{w} \mid \mathbf{y})$ are deterministic.
- The dependency graph suggests the Markov chain $W \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \hat{W}$.
- 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, assume

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

- Then from the information diagram for $W \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \hat{W}$, 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 $I(\mathbf{X}; \mathbf{Y}) \leq \sum_{i=1}^n I(X_i; Y_i)$

Proof.

1. Establish

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

2 .

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

Building Blocks of the Converse

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

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

- Then

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

- To be established in Lemma 7.16 ,

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

- Therefore,

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

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

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

3. By Fano's inequality,

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

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 $n \rightarrow \infty$ and then $\epsilon \rightarrow 0$ 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 $P_e \rightarrow 1$ as $n \rightarrow \infty$.

Achievability

- Consider a DMC $p(y | x)$.
- 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 λ_{\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

$$\Pr \left\{ (\mathbf{X}', \mathbf{Y}') \in T_{[XY]\delta}^n \right\} \leq 2^{-n(I(X; Y) - \tau)}$$

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

Proof.

- Consider

$$\Pr \left\{ (\mathbf{X}', \mathbf{Y}') \in T_{[XY]\delta}^n \right\} = \sum_{(\mathbf{x}, \mathbf{y}) \in T_{[XY]\delta}^n} p(\mathbf{x}) p(\mathbf{y})$$

- Consistency of strong typicality: $\mathbf{x} \in T_{[X]\delta}^n$ and $\mathbf{y} \in T_{[Y]\delta}^n$ • Strong AEP: $p(\mathbf{x}) \leq 2^{-n(H(X) - \eta)}$ and $p(\mathbf{y}) \leq 2^{-n(H(Y) - \zeta)}$
- Strong JAEP: $\left| T_{[XY]\delta}^n \right| \leq 2^{n(H(X, Y) + \xi)}$
- Then

$$\begin{aligned} \Pr \left\{ (\mathbf{X}', \mathbf{Y}') \in T_{[XY]\delta}^n \right\} &\leq 2^{n(H(X, Y) + \xi)} \cdot 2^{-n(H(X) - \eta)} \cdot 2^{-n(H(Y) - \zeta)} \\ &= 2^{-n(H(X) + H(Y) - H(X, Y) - \xi - \eta - \zeta)} \\ &= 2^{-n(I(X; Y) - \xi - \eta - \zeta)} \\ &= 2^{-n(I(X; Y) - \tau)} \end{aligned}$$

Interpretation of Lemma 7.17

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

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 $\tilde{\mathbf{X}}(1), \tilde{\mathbf{X}}(2), \dots, \tilde{\mathbf{X}}(M)$
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

$$\Pr\{\mathbf{Y} = \mathbf{y} \mid \tilde{\mathbf{X}}(W) = \mathbf{x}\} = \prod_{i=1}^n p(y_i \mid x_i)$$

6. The sequence \mathbf{Y} is decoded to the message w if

- $(\tilde{\mathbf{X}}(w), \mathbf{Y}) \in T_{[XY]\delta}^n$, and
- there does not exist $w' \neq w$ such that $(\tilde{\mathbf{X}}(w'), \mathbf{Y}) \in T_{[XY]\delta}^n$. 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.