ICE2601 Cheat Sheet

Entropy and AEP

Definition 1. *Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x), x \in \mathcal{X}$. The entropy of X is*

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

Entropy is a measure of the uncertainty of a random variable, as well as the amount of information required on average to describe the random variable.

- $0\log 0 \to 0$
- $H(X)$ only depends on $p(x)$, so we can write $H(p)$ for $H(X)$
- When X is uniform over $\mathcal{X}$, then $H(X) = \log|\mathcal{X}|$

The entropy can be regard as a form of expectation

$$H(X) = E\left[\frac{1}{p(X)}\right]$$

The joint entropy $H(X,Y)$ of a pair of discrete random variable (X,Y) with joint distribution $p(x,y)$ is defined as

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

Similarly we can derive that $H(X,Y) = -E\log p(X,Y)$. Observe that $H(X,X) = H(X)$ and $H(X,Y) = H(Y,X)$.

Definition 2. *If $(X,Y) \sim p(x,y)$, the conditional entropy $H(Y|X)$ is defined as*

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \text{ (by total probability rule)} \\ &= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\ &= -E\log p(Y|X) \end{aligned}$$

The following are important properties

- $H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$
- If X,Y are independent, $H(X,Y) = H(X) + H(Y)$
- If X is a function of Y, then $H(X,Y) = H(Y)$
- Bayesian formula: $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
- If $H(Y|X) = 0$, then Y is a function of X. The property is referred to as zero entropy. If $H(X|Y) = 0$ also holds, then a bijection f exists $f := X \mapsto Y$. Note that it can be extended to multi-variable cases.

Definition 3. *Relative entropy: A measure of the distance between the two distributes. The relative entropy or KL distance between the two probability mass functions $p(x)$ and $q(x)$ over $\mathcal{X}$ is defined as*

$$D(p\parallel q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = E_p\left[\log\frac{p(X)}{q(X)}\right]$$

- The relative entropy $D(p\parallel q)$ is a measure of the **inefficiency** of assuming that the distribution is q when the true distribution is p.
- $p\log\frac{p}{0} = \infty$, so $D(p\parallel q)$ can go to infinity
- Important corollary: $D(p\parallel q) \geq 0$ (can use convexity or $\log x \leq x-1$ for proof), equality holds iff $p(x) = q(x)$ for all x
- $D(p\parallel q) = E_p[-\log q(x)] - H(p)$
- $D(p\parallel p) = 0$

Definition 4. *Mutual Information: consider two random variables X,Y with joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$.*

$$\begin{aligned} I(X;Y) &= \sum_X\sum_Y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D[p(x,y)\parallel p(x)p(y)] \\ &= E_{p(x,y)}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right] \geq 0 \end{aligned}$$

We can interpret mutual interpretation as the reduction in the uncertainty of X due to the knowledge of Y.

- $I(X;Y) = I(Y;X)$
- $I(X;X) = H(X)$
- If X,Y are independent, then $I(X;Y) = 0$

Theorem 1 (Chain rule of entropy)**.** *Suppose $X_1,\ldots,X_n$ are drawn according to $p(x_1,x_2,\ldots,x_n)$, then we have*

$$H(X_1,\ldots,X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1},\ldots,X_1)$$

Definition 5. *Conditional mutual information: shows the reduction in the uncertainty of X due to the knowledge of Y when Z is given.*

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E_{p(x,y,z)}\left[\log\frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right]$$

Theorem 2. *Chain rule for information.*

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

Definition 6. *Conditional Relative Entropy*

$$\begin{aligned} D[p(y|x)\parallel q(y|x)] &= \sum_x p(x)\sum_y p(y|x)\log\frac{p(y|x)}{q(y|x)} \\ &= \sum_x\sum_y p(x)p(y|x)\log\frac{p(y|x)}{q(y|x)} \\ &= E_{p(x,y)}\left[\frac{p(Y|X)}{q(Y|X)}\right] \end{aligned}$$

Theorem 3. *Chain rule for relative entropy*

$$D[p(x,y)\parallel q(x,y)] = D[p(x)\parallel q(x)] + D[p(y|x)\parallel q(y|x)]$$

Definition 7. *Independence Bound: Let $X_1,\ldots,X_n$ be drawn according to $p(x_1,\ldots,x_n)$. Then*

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

with equality iff X_i are independent.

Definition 8. *Markov Chain: Random variables X,Y,Z are said to form a Markov Chain in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends on Y and is conditionally independent of X, i.e.*

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

- $X \to Y \to Z$ iff X,Z are conditionally independent given Y, i.e.

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

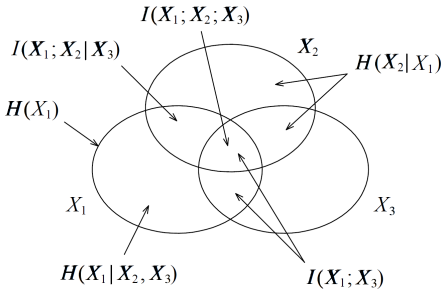
- $X \to Y \to Z$ implies that $Z \to Y \to X$.
- $Z = f(Y)$, then $X \to Y \to Z$
- If $X \to Y \to Z$, then $I(X;Z|Y) = 0$.

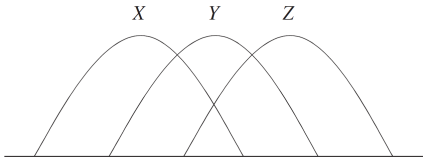
Theorem 4. *Data Processing Inequality: if $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.*

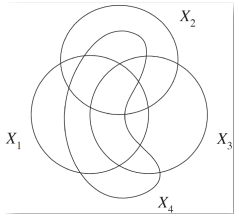
The following are important properties

- $I(X;Y,Z) = I(Y,Z;X) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y) = I(X;Y)$
- In particular, if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$
- If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$

The most important three figures for this part.







Suppose we want to estimate a random variable X with a distribution $p(x)$. We observe a random variable Y that is related to X by the conditional distribution $p(y|x)$. From Y, we calculate a function $g(Y) = \hat{X}$, where $\hat{X}$ is an estimate of X. To bound the probability $\hat{X} \neq X$, we observe that $X \to Y \to \hat{X}$ forms a Markov chain. Define the probability of error $P_e = Pr(\hat{X} \neq X)$.Fano's Inequality establishes the relationship between P_e and $H(X|Y)$.

Theorem 5. *Fano's Inequality: For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = Pr(\hat{X} \neq X)$, we have*

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

The second inequality is actually yielded by Data-processing Inequality. This inequality can be weakened to

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y)$$

From Fano's inequality, we can see that $P_e \to 0$ implies $H(X|Y) \to 0$.

Definition 9. *The counterpart of Law of Large Numbers is Asymptotic Equipartition Property (AEP): if $X_1, \ldots, X_n$ are i.i.d $\sim p(x)$, then*

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \to H(X)$$

We can reinterpret the equation as follows

$$H(X) - \varepsilon \leq -\frac{1}{n} \log p(X_1, \ldots, X_n) \leq H(X) + \varepsilon$$

$$2^{-n(H(X)+\varepsilon)} \leq p(X_1, \ldots, X_n) \leq 2^{-n(H(X)-\varepsilon)}$$

Definition 10. *Typical set $A_\varepsilon^{(n)}$ is a set such that*

$$2^{-n(H(X)+\varepsilon)} \leq p(X_1, \ldots, X_n) \leq 2^{-n(H(X)-\varepsilon)}$$

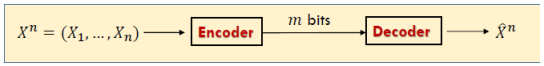
Important properties for typical sets

- All elements of the typical set are nearly equiprobable.
- The typical set has probability nearly 1. $Pr(A_\varepsilon^{(n)}) \geq 1 - \varepsilon$.
- The number of elements in the typical set is nearly 2^{nH} (much smaller than the scale of the alphabet). $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$ and $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$ for large n.

Entropy Rates and Data Compression

Intuition

The basic idea for data compression is to **find a shorter sequence to encode the source**. The basic process is shown as follows



- Suppose $X_1, \ldots, X_n$ are i.i.d $\sim p(x)$, $X^n = (X_1, \ldots, X_n)$ denotes the n-tuple that represents a sequence of n source symbols.
- The alphabet $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$ are the possible values that each X_i can take on.
- Encoder and decoder are a pair of functions f, g such that $f : \mathcal{X} \to \{0,1\}^*$ and $g : \{0,1\}^* \to \mathcal{X}$.
- Probability of error $P_e = P(X^n \neq \hat{X}^n)$
- The rate of a scheme: $R = \frac{m}{n}$. Intuitively, $R = \log |\mathcal{X}|$ is a trivial solution, as we can always represent $|\mathcal{X}|$ numbers using $\log |\mathcal{X}|$ bits. If so, then we naturally need $m = n \log |\mathcal{X}|$ bits for encoding n symbols.

Our task is to find an encoder and decoder pair such that $P_e \to 0$ as $n \to \infty$. Note that the typical set accounts for nearly all cases (with probability nearly 1). Hence, we know that we can represent $(A_\varepsilon^{(n)})^c$ with $n \log |\mathcal{X}| + 1 + 1$ bits. For $A_\varepsilon^{(n)}$, it has $2^{n(H+\varepsilon)}$ elements and there are $n(H+\varepsilon) + 1 + 1$ bits. The added two bits are used for reservation and flag (0 for being in the typical set and 1 for the other case) respectively.

Theorem 6. *We can represent sequences X^n using $nH(X)$ bits on average. Conversely, for any scheme with rate $r < H(X)$, $P_e \to 1$.*

$$E\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \varepsilon$$

证明.

$$\begin{aligned}
E(l(X^n)) &= \sum_{x^n} p(x^n) l(x^n) \\
&\leq \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)[n(H+\varepsilon)+2] + \sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n)(n \log |\mathcal{X}| + 2) \\
&= Pr\{A_\varepsilon^{(n)}\}[n(H+\varepsilon)+2] + Pr\{(A_{\varepsilon^{(n)}})^c\}(n \log |\mathcal{X}| + 2) \\
&\leq n(H+\varepsilon+2) + \varepsilon n(\log |\mathcal{X}|) + 2 \\
&= n(H+\varepsilon')
\end{aligned}$$

□

Stochastic Process and Entropy Rate

Definition 11 (Stochastic process)**.** *A stochastic process $\{X_i\}$ is an indexed sequence of random variables. It is said to be stationary if*

$$Pr(X_1 = x_1, \ldots, X_n = x_n) = Pr(X_{1+l} = x_1, \ldots, X_{n+l} = x_n)$$

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process, then we can prove that

$$H(X_0|X_{-1}, \ldots, X_{-n}) = H(X_0|X_1, \ldots, X_n)$$

Recall that $X_1, \ldots, X_n$ is said to be a Markov chain or a Markov process if for any n

$$P(X_{n+1} = x_{n+1}|X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

The Markov chain is said to be *time invariant* if the conditional probability $p(x_{n+1}|x_n)$ does not depend on n, i.e.

$$Pr(X_{n+1} = b|X_n = a) = Pr(X_2 = b|X_1 = a), \forall a, b \in \mathcal{X}$$

We will assume that the Markov chain is time invariant unless otherwise stated, and it is characterized by its initial state and a probability transition matrix $P = \{P_{ij}\}$, where

$$P_{ij} = Pr\{X_{n+1} = j|X_n = i\}$$

By the definition of stationary, a Markov Chain is stationary iff $p(X_{n+1}) = p(X_n)$. If the probability mass function at a time n is $p(x_n)$, then

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}} \Leftrightarrow x_n^T P = x_{n+1}^T$$

Definition 12. *Entropy rate: the entropy rate of a stochastic process $\{X_i\}$ is defined by*

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

Theorem 7. *Suppose $H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_1)$ and $\{X_n\}$ is stationary stochastic process, we have*

$$H(\mathcal{X}) = H'(\mathcal{X}) = H(X_2|X_1) = \sum_i p(x_i) H(X_2|X_1 = x_i)$$

Kraft Inequality and Optimal Codes

Without of loss of generality, we assume that the D-ary alphabet is $\mathcal{D} = \{0, 1 \ldots, D-1\}$. A **source code** C for a random variable X is a mapping from $\mathcal{X}$, the range of X, to D^*, the set of finite-length strings of symbols from a D-ary alphabet. Let $C(x)$ denote the **codeword** corresponding to x and let $l(x)$ denote the length of $C(x)$, then the expected length $L(C)$ of a source code $C(x)$ for a random variable X with probability mass function $p(x)$ is given by

$$L(C) = \sum_{x\in\mathcal{X}} p(x) l(x)$$

To minimize $L(C)$ and construct the code for achieving the minimum values, we have Kraft Inequality and Huffman Encoding. Some concepts are listed as follows.

- A code is *nonsingular* if for all $x \neq x' \Rightarrow C(x) \neq C(x')$
- The *extension* C^* of a code C is the mapping from finite length strings of X to finite length strings of D, defined by

  $$C(x_1 \ldots x_n) = C(x_1) \ldots C(x_n)$$

  where $C(x_1) \ldots C(x_n)$ indicates concatenation of the corresponding codewords
- A code is called *uniquely decodable* if its extension is nonsingular
- A code is called a prefix code or instantaneous code if no codeword is a prefix of any other codeword.

Theorem 8. *Kraft Inequality: for any instantaneous code over an alphabet of size D, the codeword lengths $l_1, \ldots, l_m$ must satisfy the inequality*

$$\sum_{i=1}^{m} D^{-l_1} \leq 1$$

Conversely, given a set of codeword lengths that satisfy this inequality, there exists an instantaneous code with these word lengths.

Note that the inequality can be extended to countable infinite set of codewords. With Kraft inequality, we can formulate the original problem into an optimization problem

$$\min L = \sum_i p_i l_i$$
$$\text{subject to } \sum_i D^{-l_i} \leq 1$$

The result of which is $l_i^* = -\log_D p_i$ (non-integer optimal value). The optimal expected length is thus

$$L^* = \sum_i p_i l_i^* = \sum_i -p_i \log p_i = H_D(X)$$

In general, $H_D(X)$ cannot be attained, and we have $L^* \geq H_D(X)$. We can further derive that $L^* < H_D(X) + 1$, as we might round l_i up to become an integer.

Theorem 9. *Important bound on the optimal code length.*

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

This type of coding is called **Shannon codes**, where

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil$$

It can be shown that we can actually approach the limit and remove the extra 1 bit of Shannon codes. We have the following theorem: the minimum expected codeword length per symbol satisfies

$$\frac{H(X_1, \ldots, X_n)}{n} \leq L^* < \frac{H(X_1, \ldots, X_n)}{n} + \frac{1}{n}$$

If $X_1, \ldots, X_n$ is a stationary stochastic process, then $L^* \to H(\mathcal{X})$.

Theorem 10. *In some cases, we might design the code based on an estimated (possibly wrong) distribution $q(x)$, the expected length under $p(x)$ of the code assignment $l(x) = \log \frac{1}{q(x)}$ satisfies*

$$H(p) + D(p \parallel q) \leq E[l(x)] < H(p) + D(p \parallel q) + 1$$

证明. Both sides of the equation can be derived in a similar way, here we consider the right-hand side.

$$\begin{aligned}
El(x) &= \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil \\
&< \sum_x p(x) \left[\log \frac{1}{q(x)} + 1 \right] \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + \sum_x p(x) \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\
&= D(p \parallel q) + H(p) + 1
\end{aligned}$$

□

If we slack the conditions from instantaneous codes to uniquely decodable codes, then we have Kraft Inequality is still the tight bound (cannot be shorter).

Two Encoding Methods

D-ary *Huffman codes* for a given contribution: each time combine D symbols with the lowest possibilities into a single source symbol, until there is only symbol. Note that Huffman coding is optimal.

If $D \geq 3$, we may not have a sufficient number of symbols so that we can combine them D at a time. In such a case, we add *dummy variables* to the end of the set of symbols. Since at each stage of the reduction, the number of symbols is reduced by $D-1$, the total number of symbols should be $1 + k(D-1)$, where k is the number of merges.

Now we discuss some of the properties of Huffman coding

- Huffman code is not unique.
- If a probability distribution $P(X)$ is called *D-adic* if each of the probabilities $P(X = x_i) = D^{-n}$ for some n. As proved previously, for a *D-adic* distribution, the optimal solution in Lagrange is unique: $l_i = \log \frac{1}{p_i} = n_i$.
- Compare Huffman and Shannon codes, we can find that if the probability distribution is D-adic, Shannon codes are optimal. But Shannon codes may be much worse when $p_i \to 0$.

Finally, we consider Shannon-Fano-Elias coding as a specific way to put Shannon's coding in practice. Without loss of generality, we can take $\mathcal{X} = \{1, 2, \ldots, m\}$. Assume that $p(x) > 0$ for all x. The cumulative distribution function $F(x)$ is defined as $F(x) = \sum_{a\leq x} p(a)$.

Consider the modified cumulative distribution function [using an idea of taking the average]

$$\overline{F}(x) = \sum_{a<x} p(a) + \frac{1}{2} p(x) = F(x) - \frac{1}{2} p(x)$$

Since $\overline{F}(x)$ is a real number, we truncate $\overline{F}(x)$ to $l(x)$ bits and use the first $l(x)$ bit of $\overline{F}(x)$ as a code for x, denote it by $\lfloor \overline{F}(x) \rfloor_{l(x)}$. Apparently, we have $\overline{F}(x) - \lfloor \overline{F}(x) \rfloor_{l(x)} \leq \frac{1}{2^{l(x)}}$. If $l(x) = \lceil \log \frac{1}{p(x)} \rceil + 1$, we have

$$\frac{1}{2^{l(x)}} \leq \frac{p(x)}{2} = \overline{F}(x) - F(x-1)$$

We can see that $\lfloor \overline{F}(x) \rfloor$ lies within the step corresponding to x. Thus, $l(x)$ bits suffice to describe x, i.e.
$L = \sum p(x) l(x) < H(X) + 2$. The general idea stretches as follows

$$p(x) \Rightarrow F(x) = \sum_{a \le x} p(a) \Rightarrow \overline{F}(x) = F(x) - \frac{1}{2} p(x) \Rightarrow l(x) + 1 \text{ bits}$$

Optimality: Let $l(x)$ be the codeword lengths associated with the Shannon code, and let $l'(x)$ be the code word length associated with any other uniquely decodable code. Then

$$P[l(X) \ge l'(x) + c] \le \frac{1}{2^{c-1}}$$

Hence, no other code can do much better than the Shannon code most of the time.

Channel Capacity

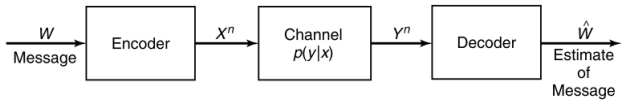
Intuition, Definition and Calculation

For each task of transmitting information, the message M with alphabet $\mathcal{M}$, the input is X with alphabet $\mathcal{X}$, the output is Y with alphabet $\mathcal{Y}$. $\mathcal{X}, \mathcal{Y}$ can be disjoint. The change from $X \to Y$ can be modeled as a transition matrix between X, Y, i.e. $p(Y|X)$.
The channel is just like a phone, each time a user can use it to make a call M. But the message may be too large to send in just one use of channel. Thus, we have

$$M \to X_1 \ldots, X_n$$

That is, the channel is used n times and we use a random process $\{X_i\}$ to denote it. We will now define the discrete memoryless channel (DMC).

Definition 13. *A discrete channel is a system consisting of an input alphabet $\mathcal{X}$ and an output alphabet $\mathcal{Y}$ and a probability transition matrix $p(y|x)$. The channel is said to be memoryless if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.*



Definition 14. *We define the information channel capacity of a discrete memoryless channel as*

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

where the maximum is taken over all input distributions of $p(x)$.

$I(X;Y) = H(Y) - H(Y|X)$ is the formula that's most frequently used. We can estimate $H(Y|X) = \sum_x H(Y|X = x)p(x)$ by the given transition probability transition matrix. In very few situations, we use $I(X;Y) = H(X) - H(X|Y)$.

- Binary Symmetric Channel. $C = 1 - H(p)$, where p is the crossover probability.
- Binary Erasure Channel. $C = 1 - \alpha$, where α is the erasure probability.
- BSC is a special case of symmetric channel. A channel is said to be symmetric if the rows of the channel transition matrix $p(y|x)$ are permutations of each other and the columns are permutations of each other. It can be shown that

  $$C = \log |\mathcal{Y}| - H(\mathbf{r})$$

  where $\mathbf{r}$ denotes a row of a transition matrix.

Channel Coding Theorem

The encoding and decoding process (i.e. $W \to X^n$ and $Y^n \to \hat{W}$) could be designed by us. $X^n \to Y^n$ is out of control and depends on $p(y|x)$. A good design should attempt to decrease n, in other word, we try to maximize $\frac{H(W)}{n}$.

For each message w, we can denote it by their index set $\mathcal{M} = \{1, 2, \ldots, M\}$. Apparently, we need to use $\log M$ bits to represent a symbol in $\mathcal{M}$. Due to the length of the message, we need to use the channel n times on average to send an index. That is, for each $w \in \mathcal{M}$, $w = x_1 \ldots x_n \in \mathcal{X}^n \Rightarrow y_1 \ldots y_n \in \mathcal{Y}^n$. By the time of their generation, we can derive

$$w \to x_1 \to y_1 \to x_2 \to y_2 \to \cdots \to x_n \to y_n$$

In general, x_i depends on $w, x_1, \ldots, y_1, \ldots, y_{i-1}$.

The n-th extension of the discrete memoryless channel is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$$

Note that $x^k = x_1, x_2, \ldots, x_k$. When x_k is given, y_k is determined by $p(y|x)$ and is independent of all generated before time k. If the channel is used without feedback, then we have $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$. Combining memoryless and no feedback, we have the following property

$$\begin{aligned} p(y^n|x^n) &= p(y^{n-1}|x^n)p(y_n|y^{n-1}, x^n) \\ &= p(y^{n-1}|x^{n-1}, x_n)p(y_n|y^{n-1}, x^n) \\ &= p(y^{n-1}|x^{n-1})p(y_n|y^{n-1}, x^{n-1}, x_n) \\ &= p(y^{n-1}|x^{n-1})p(y_n|x_n) \\ &= \prod_{i=1}^{n} p(y_i|x_i) \end{aligned}$$

Thus, we can also derive that

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

So we can regard it as a Markov Chain $W \to X^n \to Y^n \to \hat{W}$.
Now, we focus our attention on the encoding and decoding process. An encoder is a function f such that $f(w) : \mathcal{M} \to \mathcal{X}^n$. f yields a distribution on $\mathcal{X}^n$ ($\mathcal{X}$ if the channel is a DMC). The encoding rule $f(w) = x^n \in \mathcal{X}^n$ generates a codebook. (e.g. $f(\text{hi}) = 01011$)
Decoder received $y^n \sim p(y^n|x^n) = \prod p(y_n|x_n)$. The decoder need to guess the possible x^n by y^n. By the codebook $f^{-1}(x^n) = w$, $\widehat{w}$ could be recovered by decoder. An error would occur if $\widehat{w} = w$.
Formally, an (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following

- An index set $\{1, 2, \ldots, M\}$. (M messages in total).
- An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), \ldots, x^n(M)$. The set of codewords is called the codebook.
- A decoding function

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

  which is a deterministic rule that assigns a guess to each possible received vector.

Definition 15. *We define the conditional probability of error given that the index i was sent*

$$\lambda_i = P(g(Y^n) \ne i | X^n = x^n(i)) := \sum_{y^n} p(y^n|x^n(i) I[g(y^n \ne i)])$$

The maximal probability of error and average probability of error can be easily defined.

Definition 16. *The rate R of (M, n) code is defined by*

$$R = \frac{\log M}{n}$$

A rate R is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes such that the maximal probability of error $\lambda(n)$ tends to zero as $n \to \infty$. The capacity of a channel is the supremum of all achievable rates.

Theorem 11. *Channel coding theorem: For a discrete memoryless channel, all rates below the capacity C are achievable. Specially, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$. Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.*

To minimize the chance of error, our goal is that no two X sequences produce the same Y output sequence. For each *typical* input n-sequence, there are *approximately* $2^{nH(Y|X)}$ possible Y sequences, all of them equally likely. The total *number* of possible *typical* Y sequences is $2^{nH(Y)}$.

This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences. The total number of disjoint sets is less than or equal to $2^{nI(X;Y)}$. Hence, we can send at most $2^{nI(X;Y)}$ distinguishable sequences of length n.

We decode a channel output Y^n as the i-th index if the codeword $X^n(i)$ is jointly typical with the received signal Y^n.

Definition 17. *Jointly typical. The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies*

$$\begin{aligned} A_\epsilon =&\{(x^n, y^n) \in X^n \times Y^n | \\ & H(X^n) \to H(X), H(Y^n) \to H(Y), H(X^n, Y^n) \to H(X, Y)\} \end{aligned}$$

Note that $X^n \in A_\epsilon^{(n)}$ and $Y^n \in A_\epsilon^{(n)}$ cannot imply $(X^n, Y^n) \in A_\epsilon^{(n)}$. The key properties for joint AEP are shown as follows.

- $P(X^n, Y^n) \in A_\epsilon^{(n)} \to 1$ as $n \to \infty$.
- $|A_\epsilon^{(n)}| \leq 2^{n(H(X;Y)+\epsilon)}$
- If $(\tilde{X^n}, \tilde{Y^n}) \sim p(x^n)p(y^n))$, then we have

  $$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq P((\tilde{X^n}, \tilde{Y^n}) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

  证明. Consider the RHS, we have

  $$\begin{aligned} P((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) &= \sum_{(x^n, y^n) \in A_\varepsilon^{(n)}} p(x^n)p(y^n) \\ &\leq 2^{n[H(X,Y)+\epsilon]} \cdot 2^{-n(H(X)-\epsilon)} e^{-n(H(Y)-\epsilon)} \\ &= 2^{-n(I(X;Y)-3\epsilon)} \end{aligned}$$

□

Proof of Channel Coding Theorem

Note that the Channel Coding Theorem has *two directions*, and we will prove the converse first.

Theorem 12. *(Converse Theorem) If $\lambda^{(n)} \to 0$ (maximum probability of error) for a $(2^{nR}, n)$ code, then $R \leq C$.*

If **no errors** are allowed, then we have

$$\begin{aligned} nR = \log M &= H(W) = H(W|Y^n) + I(X; Y^n) \\ &= I(W; Y^n) \\ &\leq I(X^n; Y^n) \quad \text{(Markov Chain)} \\ &\leq \sum_i I(X_i; Y_i) \quad \text{(To be proved soon)} \\ &\leq nC \end{aligned}$$

证明. $I(X^n; Y^n) \leq \sum_i I(X_i; Y_i)$

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

□

For the no-error case, we have proved that $R \leq C$. Generally, $H(W|Y^n) \neq 0$ and $H(W|\hat{W}) \neq 0$. By Fano's Inequality we have $1 + P_\epsilon^n H(W) \geq H(W|\hat{W})$. Hence the case becomes

$$\begin{aligned} nR = H(W) &= H(W|\hat{W}) + I(W; \hat{W}) \\ &\leq 1 + P_\epsilon^{(n)} nR + I(W; \hat{W}) \\ &\leq 1 + P_\epsilon^{(n)} nR + I(X^n; Y^n) \\ &\leq 1 + P_\epsilon^{(n)} nR + nC \text{ (proved previously)} \end{aligned}$$

Thus we have $R \leq P_\epsilon^{(n)} R + \frac{1}{n} + C \to C$.

Theorem 13. *(Forward Theorem). For a DMC, all rates below capacity C are achievable. For every rate $R < C$, there exists a $(2^{nR}, n)$ code such that $\lambda^{(n)} \to 0$.*

We can write the codebook in the form of matrix $\mathcal{C}$. 2^{nR} rows represent M messages in total, n columns denotes the D-nary code (e.g. $\{0, 1\}$ in binary case) for each use of the channel.

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix}$$

The general process of the transmission is summarized as follows.

- Fix $p(x)$, generate a $(2^{nR}, n)$ code at random according to $p(x)$

  $$p(x^n) = \prod_{i=1}^n p(x_i)$$

- The probability we generate a particular code C is

  $$P(C) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^n p(x_i(w))$$

- The code C is revealed to both the sender and the receiver. Both of them know $p(y|x)$. A message W is chosen according to a **uniform distribution**.

  $$P(W = w) = 2^{-nR}, \quad w = 1, 2, \dots, 2^{nR}$$

- The w-th codeword $X^n(w)$ is sent over the channel.
- The receiver receives a sequence Y^n according to the distribution

  $$p(y^n|x^n(w)) = \prod_{i=1}^N p(y_i|x_i(w))$$

The receiver guess which message was sent. In jointly typical decoding, the receiver declares that the index $\widehat{W}$ was sent if the following conditions are satisfied:

- $(X^n(\hat{W}), Y^n)$ is jointly typical
- There is no other index $W' \neq W$, such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

If no such $\widehat{W}$ exists or if there is more than one such, an error is declared. Let ε be the event $\{\hat{W} \neq W\}$, we need to show that

$$P(\varepsilon) \to 0$$

Main idea: If we could prove that for all codebooks (all the possible C), the average $P(\varepsilon) \leq \epsilon$, then the error probability of the best code $\leq \epsilon$. We let W be drawn according to a uniform distribution over $\{1, 2, \dots 2^{nR}\}$ and use jointly typical decoding $\hat{W}(y^n)$. The following the average of *all codewords in all codebooks*.

$$\begin{aligned} P(\varepsilon) &= \sum_C P(C) P_e^{(n)}(C) = \sum_C P(C) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(C) \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_C P(C) \lambda_w(C) \end{aligned}$$

Consider a specific codeword

$$P(\varepsilon) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} P(\varepsilon | W = w)$$

Take $P(\varepsilon|W = 1)$ for example, we have

$$\sum_C P(C)\lambda_1(C) = P(\varepsilon|W = 1)$$

$$E_i = \{(X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)}, \quad i \in \{1, 2, \dots, 2^{nR}\}\}$$

Note that an error occurs in decoding scheme if either E_1^c occurs or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$. occurs. Then we have

$$\begin{aligned} P(\varepsilon|W=1) &= P(E_1^c \cup E_2 \cdots \cup E_{2^{nR}}|W=1) \\ &\leq P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1) \end{aligned}$$

By joint AEP, $P(\varepsilon|W=1) \to 0$, and hence $P(E_1^c|W=1) \leq \epsilon$ for n sufficiently large. For $i \geq 2$, $(E_i|W=1)$, since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, so are Y^n and $X^n(i)$. Hence the probability that $X^n(i)$ and Y^n are jointly typical is smaller or equal to $2^{-n(I(X;Y)-3\epsilon)}$ by joint AEP. Here we have used the property that if $(\tilde{X}, \tilde{Y}) \sim p(x^n)p(y^n)$, then $P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}$

$$\begin{aligned} P(\varepsilon|W=1) &\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &= \epsilon + 2^{-n(I(X;Y)-R-3\epsilon)} \end{aligned}$$

If n is sufficiently large and $R < I(X;Y) - 3\epsilon$, $P(\varepsilon|W=1) \leq 2\epsilon$ and $P(\varepsilon) \leq 2\epsilon$. Hence, there exists a best codebook C^* such that

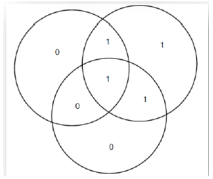
$$P(\varepsilon|C^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(C^*) \leq 2\epsilon$$

Without loss of generality, assume that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{2^{nR}}$, we can prove by contradiction that **at least half of them must all be smaller than** 4ϵ.

We can then further refine the codebook C^* by throwing away the worst half of the codewords in the best codebook C^*. The best half have a maximal probability of error less than 4ϵ. If we reindex these codewords, we have 2^{nR-1} codewords. Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n.

Hamming Code

Hamming Codes are described by (n, k, d), where the first k bits in each codeword represent the message and the last $n-k$ bits are parity check bits. A common example is $(7, 4, 3)$ Hamming Code. We can use information graph for the purpose, as we plug in the 4 numbers in the middle. Values for parity check bits are derived using the fact that every *big circle* must have sum 0 mod 2.



Differential Entropy

Concepts and Basic Properties

Definition 18. *The differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ is defined as*

$$h(X) = -\int_S f(x) \log f(x) dx$$

where $S = \{x|f(x) > 0\}$ is the support set of the random variable.

- Translation does not change the differential entropy, i.e. $h(X+c) = h(X)$
- $H(X)$ is always non-negative, but $h(X)$ may be negative.
- If $X \sim N(\mu, \sigma^2)$, then we have
$$h(f) = \frac{1}{2} \log 2\pi\sigma^2 e$$
- $h(aX) = h(X) + \log|a|$

Suppose we divide the range of X into bins of length Δ, by the mean value theorem, there exits a value x_i within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$$

Consider the quantized random variable X^Δ, which is defined by

$$X^\Delta = x_i \quad i\Delta \leq x < (i+1)\Delta$$

Then the probability that $X^\Delta = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

Yielding $H(X^\Delta) = -\sum \Delta f(x_i) \log f(x_i) - \log \Delta$, which can be simplified to

$$H(X^\Delta) + \log \Delta = -\sum \Delta f(x_i) \log f(x_i) \to h(f)$$

Theorem 14. *AEP for Differential Entropy.Let $X_1, \ldots, X_n$ be a sequence of random variables drawn i.i.d according to the density $f(x)$, then we have*

$$-\frac{1}{n} \log f(X_1, \ldots, X_n) \to E(-\log f(X)) = h(f)$$

For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows

$$A_\epsilon^{(n)} = \left\{(x_1, \ldots, x_n) \in S^n : \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \right\}$$

We can see that $\mathbb{P}(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large. Note that for continuous cases, we cannot use the number of elements to illustrate how big a set is. Instead, we use *volume*. The results are that $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$ and $\text{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$.

Definition 19. *The differential entropy of a set $X_1, \ldots, X_n$ with density $f(x_1, \ldots, x_n)$ is defined as*

$$h(X_1, \ldots, X_n) = -\int f(x^n) \log f(x^n) dx^n$$

Definition 20. *If X, Y have a joint density function $f(x, y)$, we can define the conditional differential entropy $h(X|Y)$ as*

$$h(X|Y) = -\int f(x, y) \log f(x|y) dxdy$$

Note that $h(X|Y) \leq h(X)$ with equality iff X, Y are independent.

Theorem 15. *The Chain rule for differential entropy*

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

Theorem 16. *Let $N(\mu, K)$ denote the multivariate Gaussian distribution with mean μ and covariance matrix K, then we have*

$$f(x) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)}$$

Its entropy and property are similar to the previous case

$$h(X_1, \ldots, X_n) = \frac{1}{2} \log(2\pi e)^n |K|$$

$$h(AX) = h(X) + \log|\det(A)|$$

Definition 21. *The relative entropy between two densities f and g is defined by*

$$D(f||g) = \int f \log \frac{f}{g}$$

Definition 22. *The mutual information $I(X;Y)$ between two random variables with joint density $f(x, y)$ is defined as*

$$I(X;Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dxdy$$

We still have the following properties

- $I(X;Y) = h(X) + h(Y) - h(X, Y) = D(f(x, y)||f(x)f(y))$
- $D(f||g) \geq 0$
- $I(X;Y) \geq 0$ (mutual information is still non-negative, but the entropy can be negative)

Now we introduce the *master definition* between two random variables. First, we can prove that the mutual information between two random variables is the limit of mutual information between their quantized versions [two mutual information are essentially equal]

$$\begin{aligned} I(X^\Delta;Y^\Delta) &= H(X^\Delta) - H(X^\Delta|Y^\Delta) \\ &\approx h(X) - \log\Delta - h(X|Y) + \log\Delta \text{ (shown previously)} \\ &= I(X;Y) \end{aligned}$$

Now we generalize and formalize the statement. The mutual information between two random variables X and Y (whether it is discrete or continuous) is given by

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_\mathcal{P};[Y]_\mathcal{Q})$$

where the supremum is over all finite partitions $\mathcal{P},\mathcal{Q}$. $[X]_\mathcal{P}$ denotes the quantization of X by $\mathcal{P}$, where

$$P([X]_P = i) = P(X \in P_i) = \int_{P_i} dF(x)$$

Maximum Entropy Principle

We consider the case with constraints. Let the random variable $X \in \mathbb{R}$ have mean μ and variance σ^2, then

$$h(X) \le \frac{1}{2}\log 2\pi e\sigma^2$$

with equality iff $X \sim N(\mu,\sigma^2)$

Let the random variable $X \in \mathbb{R}$ satisfy $E(X^2) \le \sigma^2$, then

$$h(X) \le \frac{1}{2}\log 2\pi e\sigma^2$$

with equality iff $X \sim N(0,\sigma^2)$. Other common properties are listed as follows for reference.

- Let $S = [a,b]$ with no other constraints, then the maximum entropy distribution is the uniform distribution over this range.
- $S = [0,+\infty)$ and $E(X) = \mu$, the the entropy-maximizing distribution is
$$f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}, \quad x \ge 0$$
- $S = \mathbb{R}$, $E(X) = \alpha_1$ and $E(X^2) = \alpha_2$, then the maximum entropy distribution is $N(\alpha_1, \alpha_2 - \alpha_1^2)$.

Some Inequalities

Definition 23. *K is a nonnegative definite symmetric $n \times n$ matrix. Let $|K|$ denote the determinant of K. Hadamard's Inequality states that $|K| \le \prod K_{ii}$ with equality iff $K_{ij} = 0, i \ne j$.*

证明. Suppose $X \sim N(0,K)$, then

$$\frac{1}{2}\log(2\pi e)^n|K| = h(X_1,\dots,X_n) \le \sum h(X_i) = \sum_{i=1}^n \frac{1}{2}\log 2\pi e|K_{ii}|$$

with equality iff $X_1,\dots,X_n$ are independent, i.e. $K_{ij} = 0, i \ne j$. □

To determine whether we can generalize a discrete entropy inequality to differential cases, we introduce balanced information entropy.

Let $[n] := \{1,2,\dots,n\}$. For any $\alpha \subseteq [n]$, denote $\{X_i, i \in \alpha\}$ by X_α. For example, $\alpha = \{1,3,4\}$, we denote X_1, X_3, X_4 by $X_{\{1,3,4\}}$ for simplicity. We could write any information inequality in the form $\sum_\alpha w_\alpha H(X_\alpha) \ge 0$ or $\sum_\alpha w_\alpha h(X_\alpha) \ge 0$.

Any information inequality is called balanced if for any $i \in [n]$, the net weight of X_i is zero. The linear continuous inequality $\sum_\alpha w_\alpha h(X_\alpha) \ge 0$ is valid iff its corresponding discrete counterpart $\sum_\alpha w_\alpha H(X_\alpha) \ge 0$ is valid and balanced.

Definition 24 (Han's Inequality)**.** *Let $(X_1,\dots,X_n)$ have a density, and for every $S \subseteq \{1,2,\dots,n\}$. We denote the subset $\{X_i := i \in S\}$ by $X(S)$. Han's Inequality states that*

$$h_1^{(n)} \ge h_2^{(n)} \ge \cdots \ge h_n^{(n)} = \frac{H(X_1,\dots,X_n)}{n} = g_n^{(n)} \ge \cdots \ge g_2^{(n)} \ge g_1^{(n)}$$

where

$$h_k^{(n)} = \frac{1}{\binom{n}{k}}\sum_{S:|S|=k}\frac{h(X(S))}{k}, \quad g_k^{(n)} = \frac{1}{\binom{n}{k}}\sum_{S:|S|=k}\frac{h(X(S)|X(S^c))}{k}$$

Definition 25. *Energy Power inequality. If X,Y are independent random n-vectors with densities, then we have*

$$e^{\frac{2}{n}h(X+Y)} \ge e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}$$

Energy Constraints

Gaussian Channel

The most important continuous channel is the Gaussian channel. The noise Z_i is drawn i.i.d. from a Gaussian distribution with variance N, and it is assumed to be independent of the signal X_i. It is a time-discrete channel with output Y_i at time i such that

$$Y_i = X_i + Z_i, \quad Z_i \sim N(0,N)$$

Without further conditions, the capacity of the channel may be ∞. Assume the variance of N is neglected compared to the distances of the values of X, then $Y = X + Z \approx X$, yielding $H(X;Y) \approx H(X)$, which might be ∞.

Usually, the most common limitation on the input is an energy or power constraint. For any codeword $(x_1,\dots,x_n)$ transmitted over the channel, we require that

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \le P$$

The information capacity of the Gaussian channel with power constraint P is

$$C = \max_{f(x):E(X^2)\le P} I(X;Y)$$

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

Note that $h(Z) = \frac{1}{2}\log 2\pi eN$, and we also have

$$E(Y^2) = E[(X+Z)^2] = E(X^2) + 2E(X)E(Z) + E(Z^2) \le P + N$$

Therefore, we have the following inequality (by Maximum Entropy Principle), equality attained when $X \sim N(0,P)$.

$$h(Y) \le \frac{1}{2}\log 2\pi e(P+N)$$

$$I(X;Y) = h(Y) - h(Z) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

Hence, we can conclude that

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

Worst Additive Noise

Under the energy constraint P, the channel capacity of additive channel $Y = X + Z$ is

$$C(Z) = \max_{X:E(X)^2\le P} I(X;Y) = \max_{X:E(X)^2\le P} = h(X+Z) - h(Z)$$

We now consider what is the minimum of $C(Z)$, if we could choose $Z := E(Z^2) \le N$. This is intended to give a lower bound of channel capacity with respect to all Z. The problem can be formalized as follows

$$\max_{Z:E(Z^2)\le N} C(Z) := \min_{E(Z^2)\le N}\max_{E(X^2)\le P} I(X;X+Z)$$

We need to find a Z^*. When $C(Z^*)$ is attained by X^*, we have

$$I(X^*;X^*+Z^*) \le \max_{X:E(X^2)\le P} I(X;X+Z)$$

Shannon proposed that $\min_{Z:E(Z^2)\le N} C(Z)$ is attained iff $Z = Z_G \sim N(0,\sigma^2)$, which proved to be true (Gaussian noise is the worst additive noise).

Appendix

- (2.8) 分析本题需要注意的是考虑两种情况下unconditional entropy $H(X_i)$相同，这样就可以很轻松地得到需要的结论。本题需要概统中的结论，不放回取球的情况下第i次取到对应颜色球的概率就是开始时的概率。

- (2.10) 本题涉及到一种非常常见且重要的处理方法，如果$X = X_1 \cup X_2$且$p(x \in X_1), p(x \in X_2)$已知，那么可以构造辅助函数

$$\theta = f(X) := \begin{cases} 1, & X \in X_1 \\ 0, & X \in X_2 \end{cases}$$

此时有$H(X) = H(X, \theta) = H(\theta) + H(X|\theta)$，正常展开即可得到对应的结论。注意本题的一种错误解法为

$$\begin{aligned} H(X) &= \sum_{x \in X} p(x) \log \frac{1}{p(x)} \\ &= \sum_{x \in X_1} p(x) \log \frac{1}{p(x)} + \sum_{x \in X_2} \log \frac{1}{p(x)} \\ &= \alpha H(X_1) + (1-\alpha) H(X_2) \end{aligned}$$

注意上式中的$p(x)$为整体的概率分布，并非是在X_1, X_2内部的概率分布，也即$-\sum_{x \in X_1} p(x) \log p(x)$中的$p(x)$的和并不为1，因而正确的代数解法需要做如下变化，此处以X_1为例

$$\begin{aligned} \sum_{x \in X_1} p(x) \log \frac{1}{p(x)} &= \alpha \sum_{x \in X_1} \left[\frac{p(x)}{\alpha} \log \frac{\alpha}{p(x)} - \frac{p(x)}{\alpha} \log \alpha \right] \\ &= \alpha H(X_1) - \sum_{x \in X_1} p(x) \log \alpha = \alpha H(X_1) - \alpha \log \alpha \end{aligned}$$

- (2.14) 这里给出一个很常用结论$H(X+Y|X) = H(Y|X)$的证明，设$Z = X + Y$

$$\begin{aligned} H(Z|X) &= \sum p(x) H(Z|X=x) \\ &= \sum_x p(x) \sum_z p(Z=z|X=x) \log p(Z=z|X=x) \\ &= \sum_x p(x) \sum_y p(Y=z-x|X=x) \log p(Y=z-x|X=x) \\ &= \sum_x p(x) H(Y|X=x) \\ &= H(Y|X) \end{aligned}$$

- (2.18) 本质上是考概统，对于七局四胜的比赛，可能的比赛场次为Y，可能的比赛结果（如AAABA）为X，我们采用如下思考方式：一共有k种组合是打m局，每种组合的概率都是0.5^m，而$k = 2C_{k-1}^3$【前$k-1$场要赢三场，乘以2表示两个人都可以赢】

- (2-additional) 在利用代数方法处理MC问题时，处理$p(x,y,z)$要灵活一些，注意考虑$p(x,y)$这种二维形式，有时可能需要进一步展开

- (2.20) 容易得到$H(X_1, \ldots, X_n) > H(R)$，在考虑定量关系的时候需要严格代入公式计算：

$$H(X_n, R) = H(R) + H(X_n|R) = H(R) + H(X_n)$$

- (2.40) $H(X+Y, X-Y) = H(X,Y)$，这是由于$(X,Y) \mapsto (X+Y, X-Y)$为双射，因而两者的熵相同（这是在两个变量情况下，函数关系所带来的熵的关系）。又因为X, Y独立，所以自然有$H(X,Y) = H(X) + H(Y)$

- (4.6) 本题可以采用归纳法来处理，如何应用MC和Stationary的条件是需要积累的

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1, X_2, \ldots, X_{n-1}) + H(X_n|X_1, \ldots, X_{n-1}) \\ &= H(X_1, \ldots, X_{n-1}) + H(X_n|X_{n-1}) \\ &= H(X_1, \ldots, X_{n-1}) + H(X_2|X_1) \end{aligned}$$

- (5.3) If you consider only finite codes (prefix or not) or even regular codes, then equality in the Kraft inequality implies maximality. Kraft不等式在一般情况下的取等条件为最优编码，事实上在Lagrange乘子法构造最优编码时我们已经使用了这个取等条件。如果没有取等，那么说明编码事实上还可以进一步压缩。

  在补充了这一点后，还需要理解[5.3]本身的题目意义。[5.3]要求构造任意长度的在$\mathcal{D}^*$中无法被解码的序列。在这里举一个简单的例子，比如对于两个码字定义$1 \mapsto 000$, $2 \mapsto 001$，我们可以构造一个子序列001001，它无法解码的；对于本题而言我们可以重复无限次001，自然也就构造出了任意长度在$\mathcal{D}^*$中无法解码的一个例子。

  更加通用的情况下，我们假设$l_1 \le l_2 \le \cdots \le l_m$，也即码字最多为$l_m$位，那么所有涉及到的编码总数为$\sum_i D^{l_m - l_i}$（依旧按照子树的思路），因而我们有

$$\sum_i D^{l_m - l_i} = D^{l_m} \sum_i D^{-l_i} < D^{l_m}$$

  这样也就说明了必然存在某个序列它不对应任何codeword，将其重复叠加便可以得到任意长度的无法被解码的$\mathcal{D}^*$中的序列。对应树的角度，可以理解为某一个非叶节点并没有填满D个孩子，此时自然一定能找到一个未被包含进这棵编码树的码字。

- (5.7) 是“提问”类型题目的一个变种，本质上一共有2^{-n}种不同的组合。从直觉上而言，如果设1为Yes而0为No，那么首先想到的是1111...的提问方式。自然，我们可以求出每一个组合对应的概率分布，之后自然可以对这n个概率分布作Huffman Coding，此时根据最优编码原理有$H(X_1, \ldots, X_n) \le L^* < H(X_1, \ldots, X_n) + 1$. 相比之下，[5.13]的解答便来得直接得多。

- (5.8) 注意MC问题考虑熵率一般都是Stationary，要求$\mathbf{xT} = \mathbf{x}$去解出初始的概率分布，根据得到的这个概率分布再去做进一步操作，有点类似求Channel Capacity中根据初始概率分布求$I(X;Y)$最大值的情况。最终得到的结论为无条件下每个字符的平均编码长度和MC的熵率相同，这是因为每个码字的平均期望长度在最优编码的条件下恰好为条件熵，即$H(X_2|X_1 = S_i)$

- (5.16) 关于处理这种问题的基本思路

  - 每一个quaternary编码都可以转化为由两个数字构成的binary编码，

  - 如果更加精细化处理，可以分奇偶来讨论Huffman Coding的码字，如果为奇数就补0使其变成偶数，这样就可以用quaternary的情况来讨论了，此时我们有

$$L_{BQ} = \frac{1}{2} \left(L_H + 1 \cdot \sum_{i:=l_i \text{ is odd}} p_i \right) \le \frac{L_H + 1}{2}$$

- (7.5/7.28) 两道题都是考虑将两个不同的信道整合成一个信道，但区别在于[7.5]中是将对应$\mathcal{X}_1, \mathcal{X}_2$的信息合在一起发送，并且没有说明$\mathcal{Y}_1 \cap \mathcal{Y}_2 \neq \emptyset$，$X_1 \perp X_2$；而[7.28]中是利用两条信道中的某一条来进行发送，并且在$\mathcal{Y}$上两组输出没有重叠部分。

  为了求得$I(X_1, X_2; Y_1, Y_2)$，都需要建立这四个变量之间的关系。对[7.5]而言，我们有$X_1 \to Y_1$, $X_2 \to Y_2$的基本关系，但X_1, X_2并不一定独立（[7.14]便是一个相关的例子），考虑到不存在反馈可以写成$X_1 \to X_2$，这样也就构成了完整的MC：$Y_1 \to X_1 \to X_2 \to Y_2$（由于我们要求能正确解码，因而$Y_1 \to X_1$是合理的）。对[7.28]而言，参考了[2.10]中处理选择类问题的解法，引入Z变量表示信道的选择，并通过$I(X; Y, Z)$的两种展开得到了合理的函数关系式。事实上，我们也可以从MC的角度来考虑本问题，本题事实上满足$X \to Y \to Z$（这运用了Y_1, Y_2无相交部分的条件）

- (7.10 (b)) Zero-error capacity的定义中'the number of bits per channel use that...'的表述与信道容量是高度相似的，只是这里要求错误为零而非趋近于零。同样注意这里的bits是一个单位，因而不一定是一个整数。我们可以写成

$$\text{zero\_err\_cap} = \frac{1}{\text{num\_chan\_uses}} \log_2 \text{num\_codewrds}$$

  本题的另一个难点在于构造出合理的编码模式，对于$y = x \pm 1 \mod 5$这种等可能对应关系，我们用长度为2的编码可以构造出形如$(i+1, 2i+1) \mod 5$的二元组【是否存在一种方式将这种构造进行推广？更加一般地，如果$y = ax \pm b \mod p$，又该如何处理？】

- (7.16) 本题相比于[7.5 & 7.28]则是将编码和解码过程加入原来的信道中，整合成为一个新的信道。处理该问题需要熟悉编码与解码的基本关系：$W \to X^n \to Y^n \to \hat{W}$，结合Data processing inequality容易发现彼此之间的关系

- (7.23) 本题涉及到一个处理$C = I(X;Y)$的思路，如果接收方可观测到的变量不止Y（以Z为例），则我们有

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

- **一些常见计算$\max H(Y)$的结论**

  - $\arg_\pi \max H(\frac{1}{2}, \frac{1}{2}(1-\pi), \frac{1}{2}\pi) = \frac{1}{2}$

  - $\arg_\pi \max H(\alpha, (1-\alpha-\epsilon)(1-\pi), \epsilon(1-\pi)+\pi(1-\alpha-\epsilon)) = \frac{1}{2}$

- (7.32) 本题是【提问题】类的内容，(a)问确定X所需问的问题数$n \geq H(X) = \log|\mathcal{X}|$，如果用Huffman Tree来建模的话必然满足$H(X) \leq L^* \leq H(X)+1$；(b)问只需要计算$X=1$与$X=2$同时包含或不包含的子集数即可（不要忘了前一种情况）；(c)问需要用概统中的一种常见思路：考虑object i和object 1面对某个问题回答相同的概率，定义

$$Y_i = \begin{cases} 1, & \text{i and 1 have same answers} \\ 0, & \text{otherwise} \end{cases}$$

  这样我们也就得到了$E(Y) = \sum_i E(Y_i)$，利用前文结论易得。

- (7.34) 利用了非常重要的结论，在这里复述一遍

  平行信道：如果某人可以从channel 1或channel 2中发送一条信息，且这两个信号输出Y对应的字母表没有重合，那么我们有$2^C = 2^{C_1} + 2^{C_2}$**；如果是n条平行信道，则有$2^C = \sum_i 2^{C_i}$

- (8.1) 在这里给出常见分布的微分熵计算公式

  - Cauchy分布

$$f(x) = \frac{\lambda}{\pi}\frac{1}{\lambda^2+x^2} \Leftrightarrow h(X) = \ln(4\pi\lambda)$$

  - 指数分布

$$f(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}} \Leftrightarrow h(X) = 1+\ln\lambda$$

  - Laplace分布

$$f(x) = \frac{1}{2\lambda}e^{-\frac{|x-\theta|}{\lambda}} \Leftrightarrow h(X) = 1+\ln 2\lambda$$

  - Lognormal分布

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}}e^{-\frac{\ln(x-m)^2}{2\sigma^2}} \Leftrightarrow h(X) = m+\frac{1}{2}\ln(2\pi e\sigma^2)$$

  - Maxwell-Boltzmann分布

$$f(x) = 4\pi^{-\frac{1}{2}}\beta^{\frac{3}{2}}x^2e^{-\beta x^2} \Leftrightarrow h(X) = \frac{1}{2}\ln\frac{\pi}{\beta}+\gamma-\frac{1}{2}$$

  - 均匀分布

$$f(x) = \frac{1}{\beta-\alpha} \Leftrightarrow h(X) = \ln(\beta-\alpha)$$

  - 三角分布

$$f(x) = \begin{cases} \frac{2x}{a}, & 0 \leq x \leq a \\ \frac{2(1-x)}{a}, & a \leq x \leq 1 \end{cases} \Leftrightarrow h(X) = \frac{1}{2}-\ln 2$$

  此外，我们也给出一些关于指数积分的基本结论，方便后续使用

$$\int_0^\infty xe^{-x}dx = \int_0^\infty e^{-x}dx = 1$$

$$\int_{-\infty}^\infty e^{-\frac{1}{2}ax^2+bx}dx = \sqrt{\frac{2\pi}{a}}e^{-\frac{b^2}{2a}}$$

- (9.1) 注意一点即可，如果$Y_1 \perp Y_2|X$，那么$I(Y_1;X) = I(Y_2;X)$，同时$I(X;Y_2|Y_1) = I(X;Y_1|Y_2)$（与$X$相关的一些计算都具有一定的对称性）

- (9.2) 处理这种问题首先根据前文的一般思路得到$I(X;Y_1,Y_2) = h(Y_1,Y_2) - h(Z_1,Z_2)$，$h(Z_1,Z_2)$固定。在Energy Constraint下$X \sim N(0,P)$时最大化熵，此时$Y_1,Y_2$的依然为二元正态分布，注意$K' = K+P$

- (9.4) 本题与高斯信道基本相同，需要注意的是此时的bound为$E(X)$，因而最大熵分布为指数分布而非正态分布。$Y = X+Z$为两个指数分布的和，其分布不要求。