


國立台灣海洋大學
 National Taiwan Ocean University

通訊與導航工程系碩士班
數位通訊
Digital Communications
Fall 2004
 吳家琪 助理教授

Lecture 1: Information Sources and Source Coding



國立台灣海洋大學
 National Taiwan Ocean University

Announcement

- **上課時間:** 星期一第六七八節 (13:10-16:00)
- **OFFICE HOURS:** Email reservations are a preferable idea.
- **上課地點:** 延平技術大樓 TEC820
- **授課老師:** 吳家琪 助理教授 (TEC809) ext. 7211
- **成績評分方式:**
 - ◆ Midterm Exam (30%)
 - ◆ Final Exam (50%)
 - ◆ Quiz and Homework (20%)
- **Textbook:**
 - ◆ *Digital Communications, 4th ed.* John Proakis, McGraw Hill, 2000, (東華/新月書局代理, 02-23114027, 02-23317856)




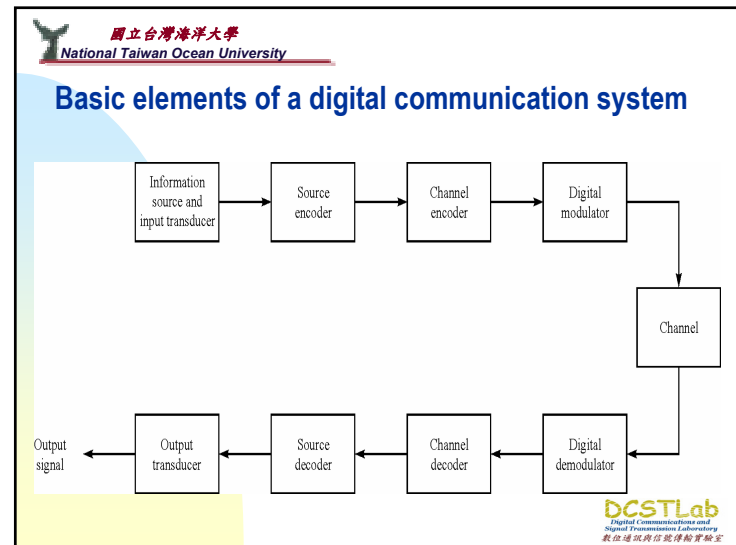

 DCSTLab
 Digital Communications and
 Signal Transmission Laboratory
 數位通訊與信號傳輸實驗室


國立台灣海洋大學
 National Taiwan Ocean University

Announcement

- **Course webpage:**
<http://www.gct.ntou.edu.tw/DCSTL/Web/dicomm.htm>
- **Textbook webpage:**
<http://www.mhhe.com/engcs/electrical/proakis/>
- **Reading Assignment:**
 - ◆ Chapter 1, 2, and 3
- **Homework No.1:** 2.3, 2.4, 2.9, 2.12, 2.20

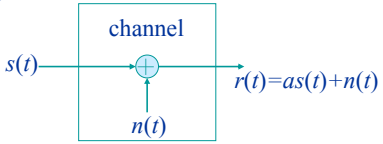

 DCSTLab
 Digital Communications and
 Signal Transmission Laboratory
 數位通訊與信號傳輸實驗室



國立台灣海洋大學
National Taiwan Ocean University

Communication channels and their characteristics

- Additive noise channel



$$r(t) = \alpha s(t) + n(t)$$

where α is the attenuation factor, $s(t)$ is the transmitted signal, and $n(t)$ is the additive random noise process.

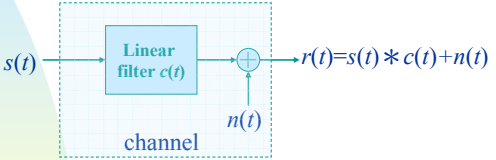
- Additive Gaussian noise channel
 - If $n(t)$ is a Gaussian noise process.

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Communication channels and their characteristics

- The linear filter channel with additive noise
 - to ensure the specified bandwidth limitations.



$$r(t) = s(t) * c(t) + n(t)$$

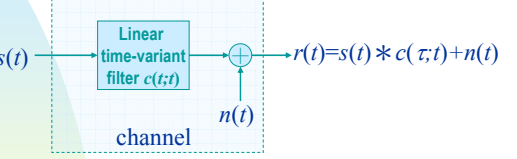
$$= \int_{-\infty}^{\infty} c(\tau) s(t - \tau) d\tau + n(t)$$

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Communication channels and their characteristics

- The linear time-variant filter channel with additive noise
 - Time-variant multipath propagation.



$$r(t) = s(t) * c(\tau; t) + n(t)$$

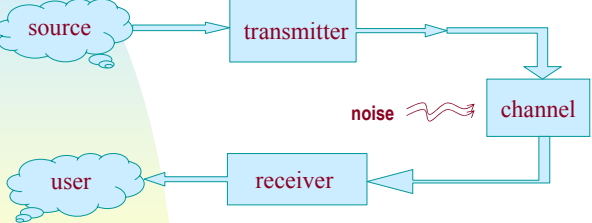
$$= \int_{-\infty}^{\infty} c(\tau; t) s(t - \tau) d\tau + n(t)$$

where $c(\tau; t)$ is the response of the channel time t due to an impulse applied at time $t - \tau$.

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Electrical Communication System

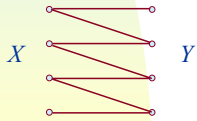


DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

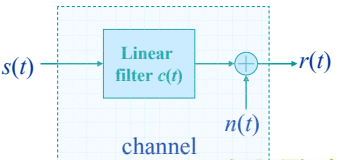
國立台灣海洋大學
National Taiwan Ocean University

Channels

- Physical channels: the atmosphere, wirelines, optical fibers, computer hard disks, compact disks, . . .
- Channel models: We need good models of the random phenomena introduced by the physical channel!
- Examples:
 - A discrete channel
 - A linear additive noise channel



A discrete channel

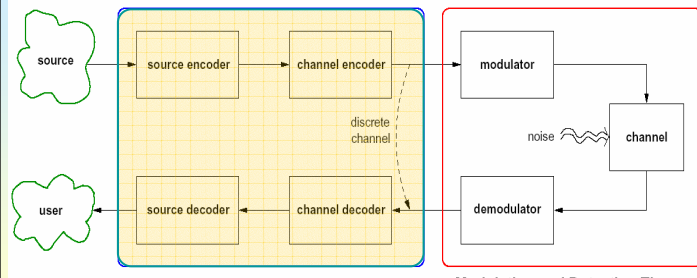


A linear additive noise channel

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Digital Communication System



Information and Coding Theory Modulation and Detection Theory

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Some Lectures on ...

- Information and Coding Theory
 - ◆ Information sources and source coding
 - Introduction to Information theory: Entropy, information measures, limits, . . .
 - Source coding: Use fewer bits to efficiently represent source data in digital form . . .
 - ◆ Channel capacity and coding
 - More on Information Theory: Channel capacity, limits, . . .
 - Channel coding: Use some clever redundant bits coding to detect and correct transmission errors. . .

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Some Lectures on ...

- Modulation and Detection
 - ◆ Modulation: Transform digital data into analog signals that can be transmitted or stored (the “real world” is analog, not digital). . .
 - ◆ Demodulation/Detection: The received signal contains information about the transmitted data but is corrupted by noise (etc.).
 - Estimate what data was sent, aiming at minimum possible probability of making mistakes. . .

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Why digital communications?

- Some sources are inherently digital (e.g. text) . . .
- Better “control” and more flexibility . . .
- Bits can be detected and regenerated. Noise does not propagate additively.
- Easier to remove redundancy and reduce bandwidth. . .
- Digital ICs are inexpensive to manufacture. A single chip can be mass produced at low cost, no matter how complex
- Digital communications allows integration of voice, video, and data on a single system (ISDN)
-

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University


Information Sources and Source Coding

- Sources, information and entropy
- Typical sequences and the source coding theorem
- Mutual information and differential entropy
- Lossless source coding
 - Definitions . . .
 - The source coding theorem . . .
 - Huffman codes
 - Lempel-Ziv codes

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Information Sources



- Source data: a speech signal, an image, a computer file, a fax, . . .
- Source data is usually time-varying and unpredictable.
- Bandlimited continuous-time signals (e.g. speech) can be sampled into discrete time and reproduced without loss.

A source is a discrete-time stochastic process $\{X_n\}$.

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Information Sources

- If $X_n \in \mathcal{X}, \forall n$, the set \mathcal{X} is the source alphabet.
- The source is
 - stationary** if $\{X_n\}$ is stationary.
 - memoryless** if X_n and X_m are independent for $n \neq m$.
 - i.i.d.** if $\{X_n\}$ is i.i.d. (independent and identically distributed).
 - continuous** if \mathcal{X} is a continuous set (e.g. the real numbers).
 - discrete** if \mathcal{X} is a discrete set (e.g. the integers $\{0, 1, 2, \dots, 9\}$).
 - binary** if $\mathcal{X} = \{0, 1\}$.

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Entropy and Information

- Consider a binary random variable $X \in \{0, 1\}$ and let $p = P_r(X=1)$.
- Before we observe the value of X there is a certain amount of uncertainty about its value. After getting to know the value of X , we gain information.
Uncertainty \leftrightarrow Information
- The average amount of uncertainty lost = information gained, over a large number of observations, should behave like

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Entropy and Information

- Define the entropy $H(X)$ of the binary variable X as

$$H(X) = \Pr(X = 1) \cdot \log \frac{1}{\Pr(X = 1)} + \Pr(X = 0) \cdot \log \frac{1}{\Pr(X = 0)}$$

$$= -p \cdot \log p - (1-p) \cdot \log (1-p) \equiv H_b(p)$$
 where $H_b(p)$ is the binary entropy function.
- $\log = \log_2$; unit = bits, $\log = \log_e = \ln$; unit = nats

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Entropy and Information

- Entropy for a general discrete variable X with alphabet \mathcal{X} and pmf $p(x) \equiv P_r(X=x); \forall x \in \mathcal{X}$

$$H(X) \stackrel{\Delta}{=} - \sum_{x \in \mathcal{X}} p(x) \log p(x) = -E[\log p(X)]$$
- $H(X)$ = the average amount of uncertainty removed when observing the value of X = the average information obtained when observing \mathcal{X}
- It holds that $0 \leq H(X) \leq \log |\mathcal{X}|$ (where $|\mathcal{X}|$ = “size of \mathcal{X} ”)
- Entropy for an N-tuple $X_1^N = (X_1, \dots, X_N)$

$$H(X_1^N) = - \sum_{x_1^N} p(x_1^N) \log p(x_1^N)$$

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Conditional Entropy

- Conditional entropy of $Y \in \mathcal{Y}$ given $X = x$

$$H(Y | X = x) \stackrel{\Delta}{=} - \sum_{y \in \mathcal{Y}} p(y | x) \log p(y | x)$$
 - $H(Y | X = x)$ = the average information obtained when observing Y when it is already known that $X = x$
- Conditional entropy of Y given X (on the average)

$$H(Y | X) \stackrel{\Delta}{=} \sum_{x \in \mathcal{X}} p(x) H(Y | X = x)$$
- Chain rule: $H(Y, X) = H(Y|X) + H(X)$
 (c.f., $p(y, x) = p(y|x)p(x)$)

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Entropy

- For a given (stationary) source $\{X_n\}$
 - ◆ $H(X_n)$ = entropy of a source sample
 - ◆ $H(X_1, \dots, X_N)$ = entropy of a set of source samples
- The entropy H of “the whole source” is called the entropy rate of the source

$$H = \lim_{N \rightarrow \infty} \frac{1}{N} H(X_1, \dots, X_N)$$
- For a memoryless (i.i.d.) source

$$H = H(X_1) = H(X_N)$$

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Typical Sequences and Compression

- A discrete memoryless source (DMS) $\{X_n\}$, with $X_n \in \mathcal{X}$, $M = |\mathcal{X}|$, entropy (rate) $H = H(X_n)$, and with

$$p(x_1, x_2, \dots, x_N) = P_r(X_1 = x_1, X_2 = x_2, \dots, X_N = x_n)$$
- Definition: A typical sequence of length N (a large number) is a sequence $x_1^N = (x_1, \dots, x_N)$ such that

$$p(x_1^N) \approx 2^{-N \cdot H}$$

(for a rigorous definition, “ \approx ” involves epsilons and deltas. . .)

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

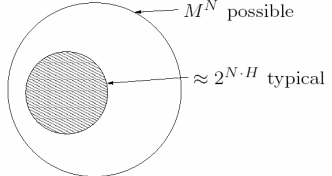
國立台灣海洋大學
National Taiwan Ocean University

Typical Sequences

- Example: Binary source $\{X_n\}$ with $p = P_r(X_n=1) = 0.1$.
 - ◆ $H = -p \cdot \log p - (1-p) \cdot \log (1-p) \approx 0.47$ [bits]
- Sequences of length $N = 20$, $2^{-NH} \approx 0.0015$
 - ◆ $x_1^N = (1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0) \Rightarrow$
 $p(x_1^N) = p^{13}(1-p)^7 \approx 5 \cdot 10^{-14} \Rightarrow$ Not typical!
 - ◆ $x_1^N = (1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0) \Rightarrow$
 $p(x_1^N) = p^2(1-p)^{18} \approx 0.0015 \Rightarrow$ Typical!
 - ◆ $x_1^N = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \Rightarrow$
 $p(x_1^N) = (1-p)^{20} \approx 0.012 \Rightarrow$ Not typical!
- A (long) sequence of length N typically contains $\approx N \cdot p$ ones and $(1-p) \cdot N$ zeros!

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University



Possible sequences and typical sequences

- There are M^N possible sequences of length N .
- There are $\approx 2^{NH} \leq M^N$ typical sequences of length N .
- For any random sequence X_1^N produced by a DMS

$$P_r(X_1^N \text{ is a typical sequence}) \approx 1$$

with equality as $N \rightarrow \infty$. That is, the probability that the source produces a non-typical sequence goes to zero!

DCSTLab
Digital Communications and
Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Typical Sequences and Compression

- We need $\approx NH$ bits to enumerate all different typical sequences. That is, H bits per source symbol.
- A source code: (1) Observe a sequence; (2a) If typical produce and store/transmit its NH bit index; (2b) If non-typical declare an error; (3) Reproduce the source sequence from the stored/transmitted index.
 - ◆ This code has rate $R = H$ bits per source symbol. As $N \rightarrow \infty$ the code works without errors.

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Typical Sequences and Compression

- The (lossless) Source Coding Theorem: For a source with entropy rate H , a lossless source code of rate R exists as long as $R > H$. For $R < H$ no lossless source code can be found.
- H measures the information content in the source, in the sense that H bits per symbol are required to describe its output!

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Mutual Information

- Consider a pair (X, Y) of discrete variables.
 - ◆ $H(X)$ = average information in observing X
 - ◆ $H(Y)$ = average information in observing Y
 - ◆ $H(X, Y)$ = average information in observing (X, Y)
 - ◆ $H(Y|X)$ = average info. in observing Y when X is known
 - ◆ $H(X|Y)$ = average info. in observing X when Y is known
- Measure of the information about X obtained when observing Y

$$I(X; Y) = H(X) - H(X|Y)$$

The mutual information between X and Y

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

$$I(X; Y) = I(Y; X)$$

$$I(X; Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

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

$$I(X; X) = H(X)$$

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

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Continuous Random Variables

- A continuous random variable, X , with pdf $f(x)$.
- Define the differential entropy $h(X)$ of X as

$$h(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx = -E[\log f(X)]$$
- $h(X)$ is not an “entropy” in the sense of a measure of uncertainty/information
- The amount of uncertainty removed/information gained when observing $X = x$ is always infinite for a continuous X .
- Mutual information for continuous variables X and Y

$$I(Y; X) = h(X) - h(X|Y) = h(Y) - h(Y|X) = I(X; Y)$$
- Still valid as a measure of information!

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Digital Communication System

Information and Coding Theory Modulation and Detection Theory

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Source Coding

- Represent source data efficiently in digital form for transmission or storage
- We concentrate on the source code and assume there are no transmission errors; the channel is noiseless.
- The source code is
 - lossless if $\hat{X} = X$
 - compression, Human, Lempel-Ziv, . . .
 - lossy if $\hat{X} \neq X$
 - rate-distortion, quantization, waveform coding, . . .

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Lossless Source Coding

- Definition: A d -ary (lossless) source code for a discrete random variable X maps a realization x of X into a finite-length string $c(x)$ of symbols from a discrete alphabet D of size $|D| = d$.
- The expected length L of the source code is

$$L = \sum_{x \in \mathcal{X}} p(x) l(x)$$
- where $l(x)$ is the length of the string $c(x)$ in symbols.
- The (average) rate R of the source code is

$$R = L \cdot \log d$$

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Lossless Source Coding

- Desired properties of a source code. A code is
 - non-singular if $c(x_1) \neq c(x_2)$ whenever $x_1 \neq x_2$.
 - uniquely decodable if any encoded string has only one possible source string producing it.
 - instantaneous if $c(x)$ can be decoded into x without looking at future codewords \iff no codeword is a prefix of any other codeword

X	Singular	Non-singular, but not uniquely decodable	Uniquely decodable, but not instantaneous	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Instantaneous \iff The Prefix Condition

Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Other classifications

- fixed-length to fixed-length
- fixed-length to variable-length
- variable-length to fixed-length
- variable-length to variable-length

Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Shannon's Source Coding Theorem

- A discrete source with entropy rate H [bits per source symbol], and a lossless source code of rate R [bits per source symbol].
- In 1948 Claude Shannon showed, based on typical sequences, that a lossless (error-free) code exists as long as $R > H$. A lossless code does not exist for any $R < H$.
- Shannon's result is an "existence/non-existence result" (as are many results in information theory). The theorem does not say how to design practical coding schemes.
- Now we first look at a more concrete coding theorem (for lossless uniquely decodable codes) and then we study two code design algorithms (Human and Lempel-Ziv).

Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Inequalities

- The codeword lengths of a uniquely decodable d -ary code for a discrete variable X must satisfy the *Kraft inequality*

$$\sum_{x \in \mathcal{X}} d^{-l(x)} \leq 1$$
- Conversely, given a set of codeword lengths that satisfy the *Kraft inequality*, it is possible to construct a uniquely decodable code with these codeword lengths.
- Any two sets $\{a_1, \dots, a_N\}$ and $\{b_1, \dots, b_N\}$ of N positive numbers must satisfy the *log-sum inequality*

$$\sum_{n=1}^N a_n \log \frac{a_n}{b_n} \geq \left(\sum_{n=1}^N a_n \right) \log \frac{\sum_{n=1}^N a_n}{\sum_{n=1}^N b_n}$$

with equality iff $a_n/b_n = \text{constant}$.

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Source Coding Theorem

- The source coding theorem for uniquely decodable codes
 - The rate R of a uniquely decodable code for a random variable X satisfies $R \geq H(X)$ [bits].
 - A uniquely decodable code for a random variable X exists that satisfies $R < H(X) + \log d$ [bits].
 - A stationary discrete source $\{X_n\}$: Code a block $X_1^N = (X_1, \dots, X_N)$ of source symbols \Rightarrow

$$\frac{1}{N} H(X_1^N) \leq R < \frac{1}{N} H(X_1^N) + \frac{\log d}{N}$$
 where R is the rate in bits per source symbol.
- A uniquely decodable code exists such that $R \rightarrow H$ as $N \rightarrow \infty$. No such code with $R < H$ can be found.

DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Huffman Codes

- A simple way of constructing a fixed-length to variable-length instantaneous code.
- Merge the two least probable nodes at each level


DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室

國立台灣海洋大學
National Taiwan Ocean University

Huffman Codes


- + Optimal in the sense that no other (fixed-length to variable-length) code can provide a lower average rate!
- + Works well also for symbol-by-symbol encoding.
- + In principle it is straightforward to obtain better compression by grouping blocks of source symbols together and letting these define new “symbols.”
 - The source statistics have to be known \Rightarrow in practice a coding in two passes needed.
 - Coding blocks significantly increases the complexity \Rightarrow the complexity of a code that achieves the entropy rate of the source is generally very high.


DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室


國立台灣海洋大學
 National Taiwan Ocean University

The Lempel-Ziv Algorithm

- A universal variable-length to fixed-length scheme.
 - ◆ A given string of source symbols is parsed into phrases of varying length.
 - ◆ new phrase = one of minimum length that has not appeared before
 - ◆ A “dictionary” is built.
 - ◆ Indices of dictionary entries and “new bits” are transmitted/stored.
- Does not work well for short strings (even expands the size of the source string), but is asymptotically optimal in the sense that the entropy rate of the source is achieved for very long strings!
- Often used in practice (‘compress’, ‘.ZIP’, ‘.ARJ’, ...)


DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室



國立台灣海洋大學
 National Taiwan Ocean University

The Lempel-Ziv Algorithm

- source string
010000110000101
- parse
0, 1, 00, 001, 10, 000, 101

Dictionary		Codeword
Index	Contents	
000	empty	
001	0	000 0
010	1	000 1
011	00	001 0
100	001	011 1
101	10	010 0
110	000	011 0
111	101	101 1

- Inefficient since the whole string must be visited to determine number of bits spent on each index. Several modifications exist.


DCSTLab
Digital Communications and Signal Transmission Laboratory
數位通訊與信號傳輸實驗室