Chapter 1 Information and Coding Theory

1.15 Coding Principle

1.16 Optimum Decoding

1.17 Maximum a posteriori (MAP) decoding

1.18 Maximum Likelihood Decoding (MLD)

1.19 Coding Gain

References

1. Massey, J.L.," Information Theory : The Copernican System of Communications," IEEE Communications Magazine , Vol.22 ,No.12 , pp. 26-28, Dec. 1984.
2. Costello, Jr. D.J. and Forney, G.D. , "Channel Coding : The Road to Channel Capacity ," Proc. IEEE , Vol.95 , No.6 , pp.1150-1177 , June 2007.
3. Verdu , S.," Fifty Years of Shannon Theory, " IEEE Trans. Inform. Theory , Vol.44 , No.6 ,pp. 2057-2078 ,Oct. 1998.
4. Shannon, C. ," A Mathematical Theory of Communications," Bell Syst. Tech. J. Vol.27 ,pp.379-423,July 1948 and pp. 623-656 , Oct. 1948.
5. .Lin, S. and Costello Jr. D.J. , Error Control Coding , Pearson Prentice Hall , 2004.
6. Castineira, J., and Farrel, P.G. , Essential of Error-Control Coding , Wiley, 2006
7. Proakis, J.G. , and Salehi ,M., Digital Communications , Fifth Edition, McGraw-Hill, 2008 .

1.1 Introduction

- **Claude Shannon's landmark paper published in 1948 is the foundation of information theory, which explores the theoretical performance limits of optimum communication systems.**
- **Knowledge of information-theoretic limits is useful for system designers because it indicates how far a real system could be improved .**
- **The theory provides answers to two fundamental questions (among others):**

  (i) What is the irreducible complexity below which a signal cannot be compressed ?

  (ii) What is the ultimate transmission rate for reliable communication over a noisy channel ?

- **The history of error-correction coding starts with Shannon's work published in 1948. He showed that it is possible to transmit data without errors as long as the bit rate is smaller than the channel capacity.**
- **The absence of errors is achieved by the use of " appropriate" codes. Shannon showed that (infinitely long) random codes achieve capacity. Unfortunately , such codes cannot be used in practice due to the enormous effort required for their decoding.**
- **For more than 50 years, the works of coding theorists mainly focused on finding practical codes that come close to the Shannon limits, i.e., allow communications with rates close to the channel capacity.**

Formal Architecture of Communication Systems

The following diagram illustrates the formal architecture Shannon offered as a schematic for a general communication system. Flip open to the beginning of any random textbook on communications, or even a paper or a monograph, and you will find this diagram.

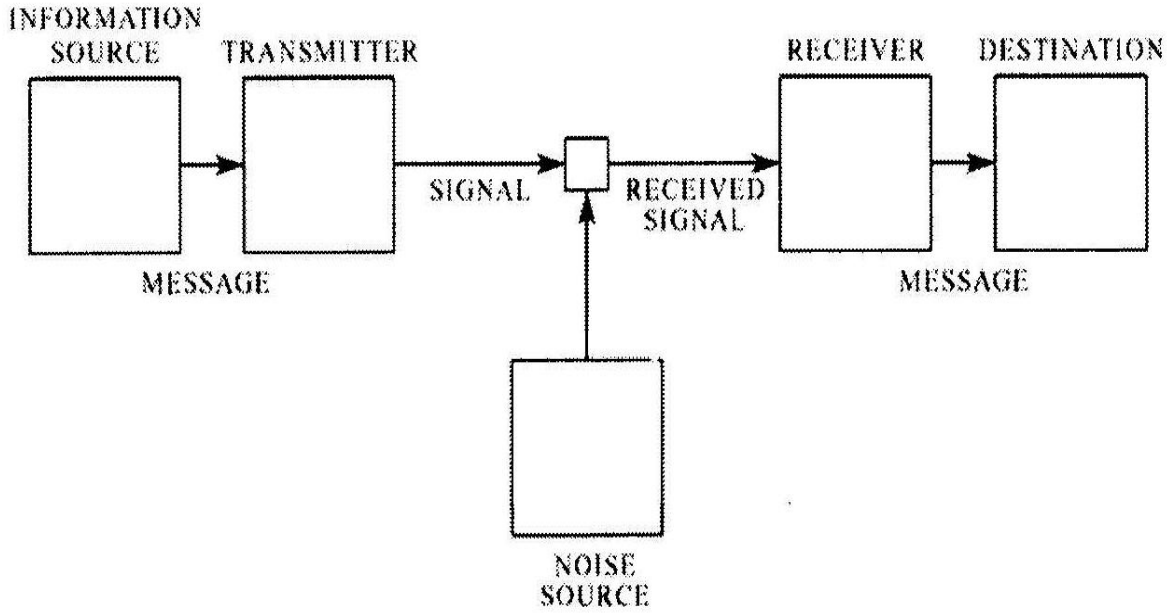


Figure 1. From Shannon's "A Mathematical Theory of Communication", page 3.

- **This figure represents one of the great contributions of " A Mathematical Theory of Communication" : the architecture and design of communication systems. It demonstrates that any communication system can be separated into components, which can be treated independently as distinct mathematical models. Thus, it is possible to completely separate the design of the source from the design of the channel. All of today's communication systems are essentially based on this model .**
- **Shannon's Information Theory**

  Source Coding Theorem

  Channel Coding Theorem

  Information Capacity Theorem

  Rate distortion Theorem

Note :

1. *R.V.L. Hartley (1928) had attempted to quantify of information as the logarithm of the number of possible message built from a pool of symbols. He introduced the concept of "information" as random variable and was the first to attempt to define " a measure of information" (1928 : " Transmission of Information " in Bell System Tech. Journal, vol.7, pp. 535-563") .*
2. *In 1924, H.Nyquist published "Certain Factors Affecting Telegraph Speed," gave an analysis of the relationship between the speed of a telegraph system and the number of signal values used by the system. His 1928 paper "Certain Topics in Telegraph Transmission Theory" refined his earlier results and established the principles of sampling continuous signals to convert them to digital signals. The Nyquist sampling theorem showed that the sampling rate must be at least twice the highest frequency present in the sample in order to reconstruct the original signal.*

These two papers by Nyquist, along with one by R.V.L. Hartley, are cited in the first paragraph of Claude Shannon's classic essay "A Mathematical Theory of Communication" (1948), where their seminal role in the development of information theory is acknowledged.

1.3 Nyquist–Shannon Sampling Theorem

- **The Nyquist–Shannon sampling theorem is a fundamental result in the field of information theory.**
  It is often referred to as simply *the sampling theorem*.
- **Sampling is the process of converting a signal (for example, a function of continuous time or space) into a numeric sequence (a function of discrete time or space).**
- **The theorem states that**
  "Exact reconstruction of a continuous-time baseband signal from its samples is possible if the signal is bandlimited and the sampling frequency is greater than twice the signal bandwidth.".

1.4 Discrete Memoryless Source

- **Memoryless source : The source is memoryless if successive symbols emitted by the source are statistically independent.**
- **Entropy of Discrete Memoryless Source :**

  Assume that the source output is modeled as a discrete random variable, which takes on symbols from a fixed finite alphabet

$$S = \{ s_0, s_1, \ldots, s_{K-1} \}$$

with probabilities

$$p(S = s_k) = p_k \quad , \quad k = 0,1,2,\ldots, K-1$$

and $\Sigma_{k=0}^{K-1} p_k = 1$

- **The amount of information gained after observing the event is defined as the logarithmic function**

$$I_k = \log_2 (1/p_k)$$

The entropy of the source is defined as the mean of information over source alphabet , and is given by

$$H(S) = E[I(s_k)]$$
$$= \Sigma_{k=0}^{K-1} p_k I(s_k)$$
$$= \Sigma_{k=0}^{K-1} p_k \log_2 (1/p_k)$$

The entropy is a measure of the average information content per source symbol.

1.5 Source Coding Theorem (Shannon's first theorem)

- **The theorem can be stated as follows:**

  Given a discrete memoryless source of entropy , the average code-word length L for any distortionless source coding is bounded as

  $$L \geqq H(s)$$

  where $H(s)$ is the entropy of the source

- **This theorem provides the mathematical tool for assessing data compaction, i.e. lossless data compression, of data generated by a discrete memoryless source.**
- **The entropy of a source is a function of the probabilities of the source symbols that constitute the alphabet of the source.**

1.6 Discrete Memoryless Channel

- A discrete memoryless channel (DMC) is a statistical model with an input X and an output Y that is a noisy version of X. Both X and Y are discrete random variables .

  $X = \{ x_1 , x_2 , \ldots , x_J \}$, $Y = \{ y_1 , y_2 , \ldots , y_K \}$

  It is memoryless if the current output symbol depends only the current input symbol and not any previous ones.

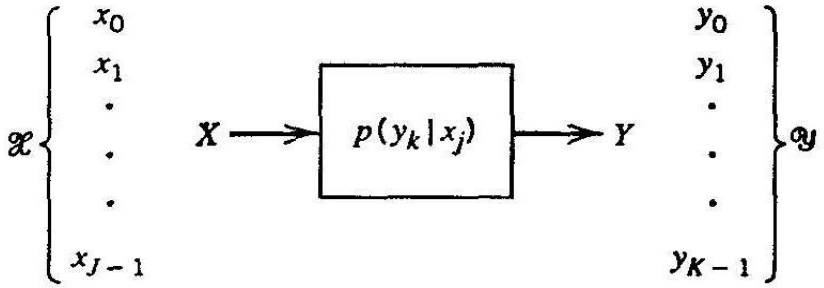
- The relation between the output and input can be defined by the transition matrix

$$P = \begin{bmatrix} p(y_1 \mid x_1) & p(y_2 \mid x_1) & \ldots & p(y_K \mid x_1) \\ p(y_1 \mid x_2) & p(y_2 \mid x_2) & \ldots & p(y_K \mid x_2) \\ & & & \\ p(y_1 \mid x_K) & p(y_2 \mid x_K) & \ldots & p(y_K \mid x_J) \end{bmatrix}$$

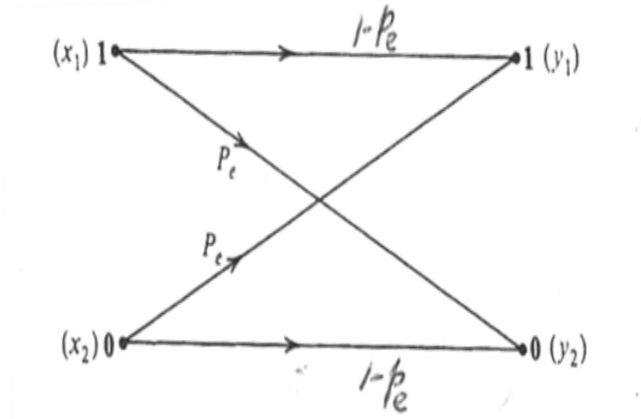
with the relation $\Sigma_{k=1}^{N} p(y_k \mid x_j) = 1$ for $j = 1, 2, \ldots, J$

Discrete memoryless channel Model

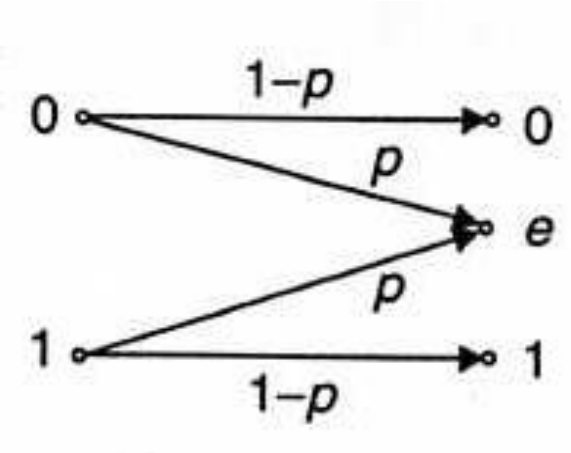
(a) General DMC



(b) Binary symmetric channel

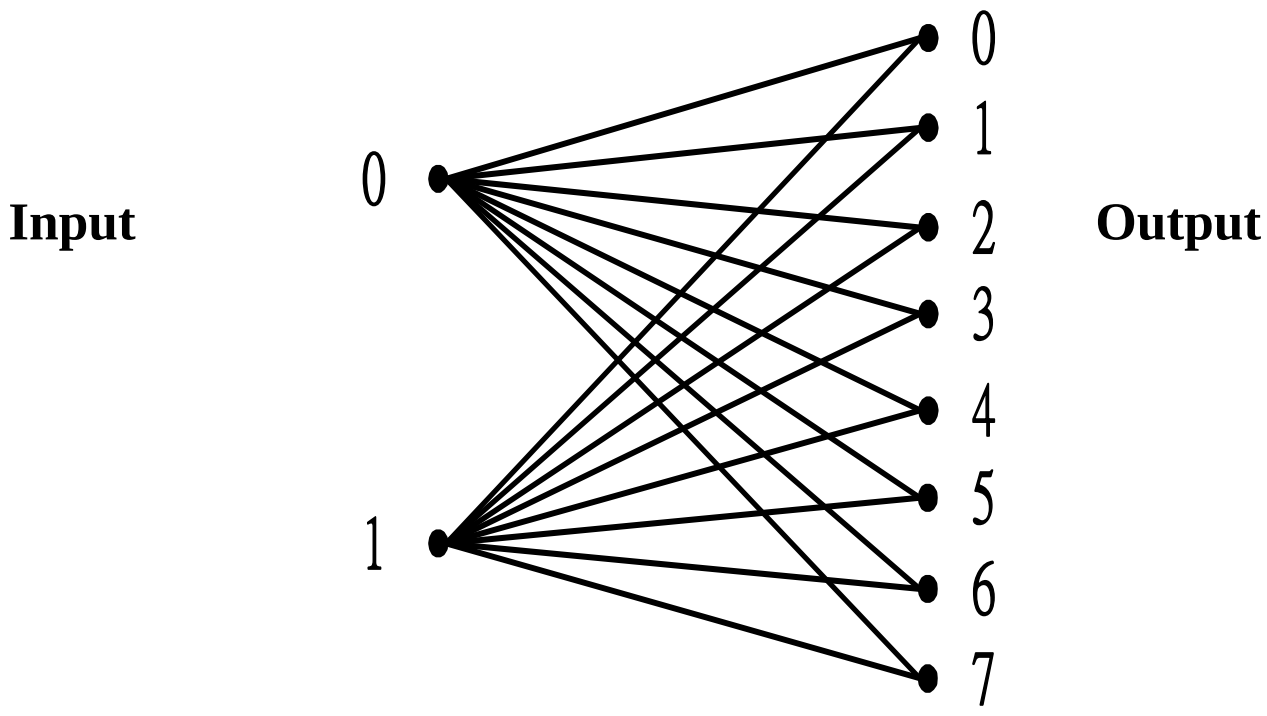


(c) Binary erasure channel



- **Binary-input, *M*-ary Output Discrete Channel**

Example : $M = 8$



- **Binary Symmetric Channel (BSC) :**

  A binary symmetric channel (BSC) is the most common channel model. In this model , both the input and output symbol alphabets are binary symbols (0 and 1) and the output symbols are dependent only on single input symbols .

  The transition probabilities are given by

  $p\,(y=0 \mid x=0) = 1-\varepsilon\,, \quad p(y=0 \mid x=1) = \varepsilon$

  $p\,(y=1 \mid x=0) = \varepsilon\,, \quad p(y=1 \mid x=1) = 1-\varepsilon$

- **Binary Erasure Channel**

Fig. (c) presents a binary symmetric erasure (BEC) channel. The output alphabet include 0, 1 and a third symbol denoted as " *e* ", called " erasure ".

The erasure symbol reflects the situation in which the receiver is not able to perform detection and decide if the received symbol is " 0 " or " 1 ".

Denoting the *a priori* probabilities of the input symbols as

$$p(X=0) = \alpha$$

$$p(X=1) = 1-\alpha$$

we obtain the probability of occurrence of an erasure

$$p(e) = p(Y=e)\,p(X=0) + p(Y=e)\,p(X=1)$$

$$= \varepsilon\alpha + \varepsilon(1-\alpha) = \varepsilon$$

Note : **The emergence of the Internet promoted the erasure channel into the class of " real world " channels . Indeed, erasure channels can be used to model data networks , where packets either arrive correctly or are lost .**

- **Binary-Input AWGN Channel**

  The binary-input additive white Gaussian noise (BI-AWGN) channel can be described by the equation

  $$y_i = x_i + z_i$$

  where x_i is the transmitted symbol , y_i is the received symbol and z_i is the additive noise .

  The probability density function (pdf) for z is

  $$p(z) = \{1/\sqrt{(}\ 2\pi\ \sigma^2\ )\ \}\quad exp\ (-z^2 / 2\sigma^2\ )$$

1.7 Joint and Conditional Entropy

- Given the input probabilities $p(x_j)$, output probabilities $p(y_k)$, the transition probabilities $p(y_k \mid x_j)$, and the joint probabilities $p(x_j, y_k)$, we can define several different entropy functions for a channel with J inputs and K outputs as follows.

$$H(X) = \sum_{j=0}^{J-1} p(x_j) \log_2 \left(1/p(x_j) \right)$$

$$H(Y) = \sum_{k=0}^{K-1} p(y_k) \log_2 \left(1/p(y_k) \right)$$

$$H(X \mid Y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) \log_2 \left(1/p(x_j \mid y_k) \right)$$

$$H(X,Y) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) \log_2 \left(1/p(x_j, y_k) \right)$$

and $$H(Y \mid X) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) \log_2 \left(1/p(y_k \mid x_j) \right)$$

- **The above entropies can be easily interpreted. H(X) is the average uncertainty of the source , whereas H (Y) is the average uncertainty of the received symbol (channel output) . The function $H(Y \mid X)$ is the average uncertainty of the received symbol given that X was transmitted . The joint entropy $H(X, Y)$ is the average uncertainty of the communication system as a whole.**
- **The following two important and useful relationships can be obtained directly from the previously defined entropies.**

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

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

1.8 Mutual Information

- **As stated before, $H(X \mid Y)$ represents the uncertainty of the input that remains after the observation of the output symbol Y, while $H(X)$ represents the uncertainty of the input before the observation of Y.**

  The difference $H(X) - H(X \mid Y)$ represents the average amount of information obtained about the channel input symbol by the observer.

  The value $H(X) - H(X \mid Y)$ is called an average amount of mutual information, and is denoted by $I(X; Y)$.

  Thus we have

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

Using Bayes' rule , we can expressed $I(X;Y)$ as

$$I(X;Y) = \Sigma_{j=0}^{J-1} p(x_j) \log_2 (1/p(x_j))$$
$$- \Sigma_{j=0}^{J-1} \Sigma_{k=0}^{K-1} p(x_j , y_k) \log_2 (1/p(x_j \mid y_k)$$
$$= \Sigma_{j=0}^{J-1} \Sigma_{k=0}^{K-1} p(x_j , y_k) \log_2 [p(y_k \mid x_j) / p(y_k)]$$

or $$I(X;Y) = \Sigma_{j=0}^{J-1} \Sigma_{k=0}^{K-1} p(x_j) p(y_k \mid x_j) \cdot$$
$$\log_2 [p(y_k \mid x_j) / \Sigma_{k=1}^{N} p(x_j) p(y_k \mid x_j)]$$

- **It is easy to show mathematically the following properties of mutual information :**

(1) $I(X;Y) \geqq 0$

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

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

1.9 Channel Capacity of Discrete Memoryless Channel

- **In a communication channel with input symbol X and output symbol Y , $I(X; Y)$ represents the information transmitted over the channel . The value of $I(X; Y)$ is a function of the symbol probabilities $P(x_j)$.**

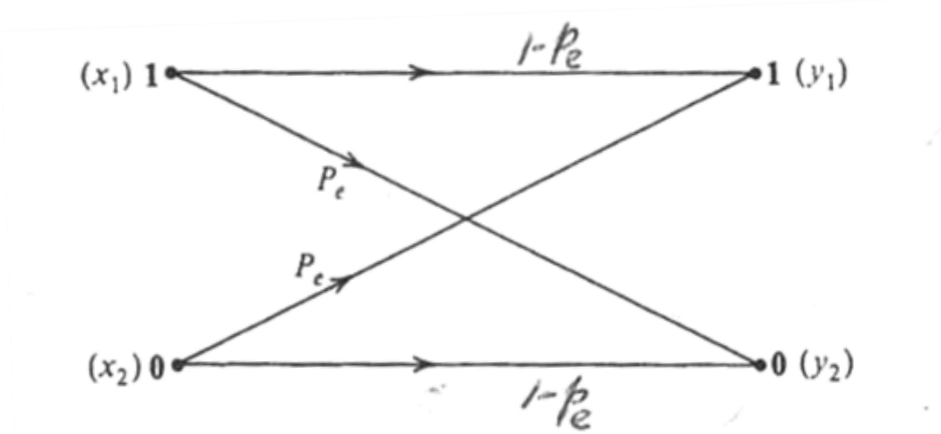
  For a given channel , $I(X; Y)$ will be maximum for some set of probabilities $p(x_j)$. The maximum value is the channel capacity C .

$$C = \max_{p(x_i)} I(X; Y) \quad \text{bits per symbol transmitted}$$

Thus, C represents the maximum information thatcan be transmitted per symbol over the channel.

- **Example :**

Channel capacity of a binary symmetric channel



$$p(x_1) = \alpha \quad , \quad p(x_2) = 1-\alpha$$

$$p(y_1 \mid x_1) = p(y_2 \mid x_2) = p_e$$

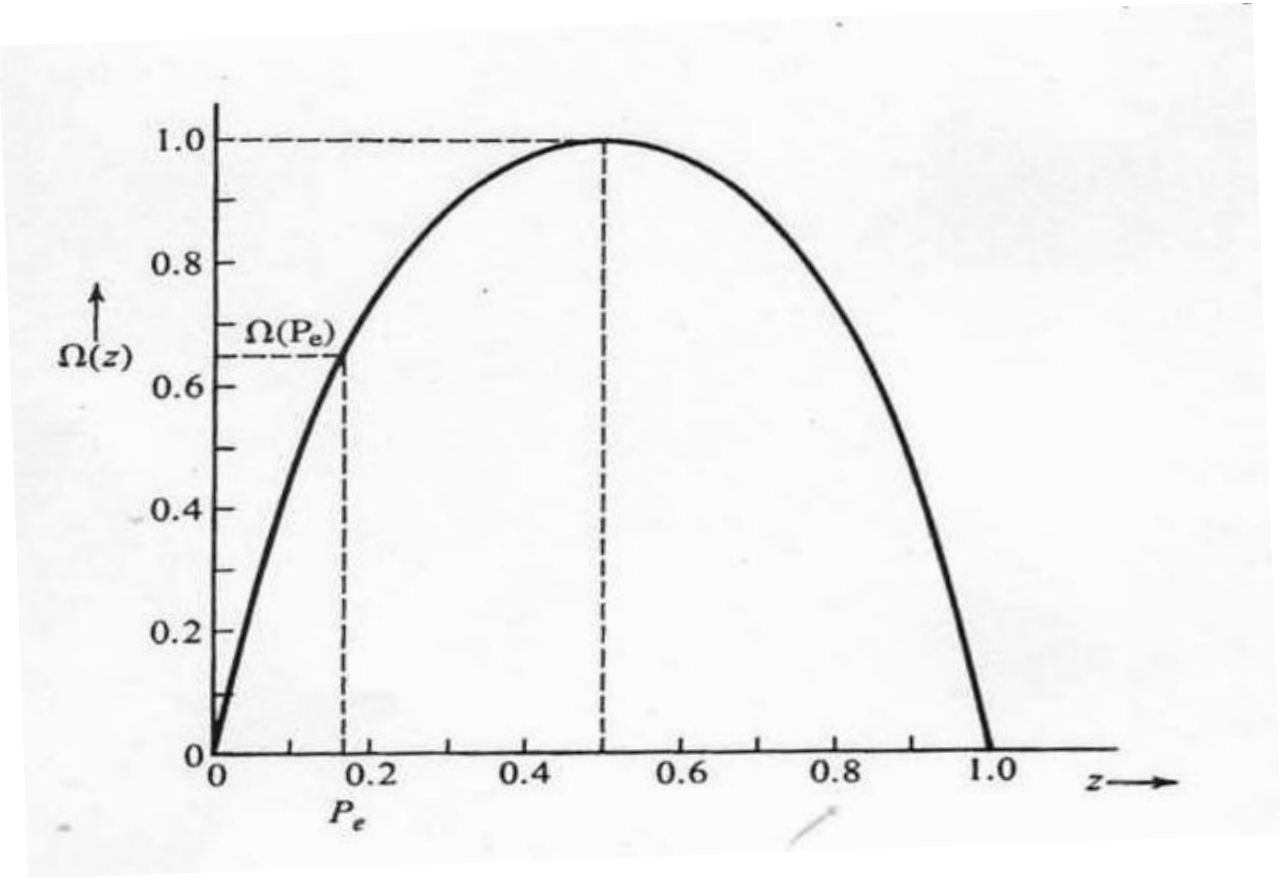
$$p(y_1 \mid x_2) = p(y_2 \mid x_2) = 1-p_e$$

$$\text{then } I(X;Y) = \alpha(1-p_e)\log_2[(1-p_e)/\alpha(1-p_e)+(1-\alpha)p_e]$$
$$+\alpha p_e \log_2[p_e)/\alpha p_e)+(1-\alpha)(1-p_e)]$$
$$+(1-\alpha)p_e) \log_2[p_e)/\alpha(1-p_e)+(1-\alpha)p_e)]$$
$$+(1-\alpha)(1-p_e)\log_2(1-p_e)/[\alpha p_e+(1-\alpha)(1-p_e)]$$

If we define $\Omega(z) = z \log_2 (1/z) + (1-z) \log_2 (1/(1-z))$

then

$$I(X;Y) = \Omega[\alpha p_e + (1-\alpha)(1-p_e)] - \Omega(p_e)$$



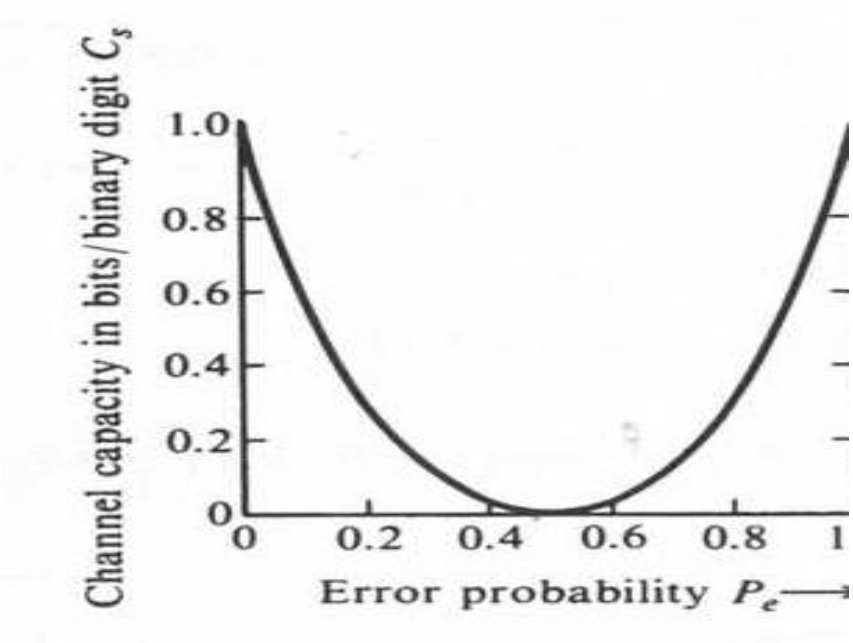
For a given p_e ,the maximum occurs when α= 0.5 .

For this value of α

$$\Omega\,[\alpha\, p_e + (1-\alpha)(1- p_e)] = 1$$

and $$C = 1 - \Omega\,(p_e)$$

$$= 1 - \{p_e \log(1/p_e) + (1-p_e)\log[1/(1-p_e)]\}$$



1.10 Channel Coding Theorem (Shannon's 2nd theorem)

The channel coding theorem for a discrete memoryless channel is stated in two parts as follows:

- **Let a discrete memoryless source with an alphabet have entropy $H(s)$ and produce symbols once every T_s seconds. Let a discrete memoryless channel have capacity C and be used once every T_c seconds.**

  If $H(S) / T_s < C / T_c$

  there exists a coding scheme for which the source output can be transmitted over the channel and be reconstructed with an arbitrarily small probability of error.

- **Conversely,**

  if $H(S) / T_s \geqq C / T_c$

  it is not possible to transmit information over the channel and reconstruct it with an arbitrarily small probability of error.

- **The theorem specifies the channel capacity as a fundamental limit on the rate at which the transmission of reliable error-free message can take place over a discrete memoryless channel.**

  Note : ***C is defined as the maximum rate of reliable transmission over the channel.***

1.11 Channel Capacity of AWGN Channel

1.11.1 Discrete-Time Channel

- Let X and Y denote the channel input and output, respectively. The additive white Gaussian is denoted as n.

$$Y = X + n$$

The samples are expressed by

$$Y_k = X_k + n_k \qquad k = 1,2,\ldots, K$$

where n_k's are iid zero-mean Gaussian random variables with variance σ^2.

The input X is subject to the power constraint

$$E[X^2] \leqq P$$

- If the number of samples, K, is very large, the noise power can be calculated as the average of the noise samples:

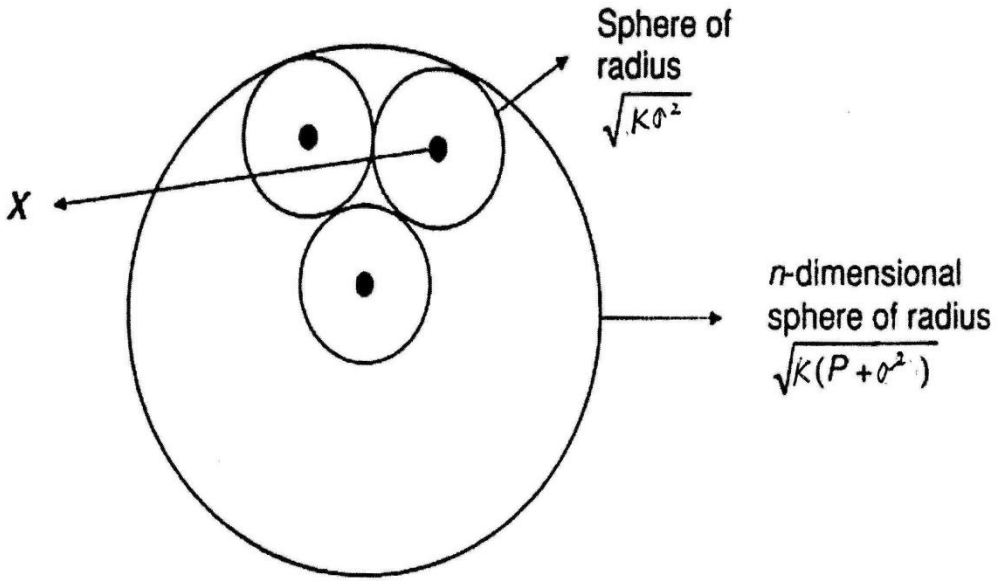
$$(1/K)\,\Sigma_{k=1}^{K}\, n_k^2 = (1/K)\,\Sigma_{k=1}^{K} \left| Y-X \right|^2 \leqq \sigma^2$$

which means that $\left| Y-X \right|^2 \leqq K\sigma^2$

- **Thus, if X is transmitted, with high probability the vector**

  Y will be inside an K-dimensional sphere of radius $\sqrt{(K\sigma^2)}$ and centered at X .

  The maximum number of spheres of radius $\sqrt{(K\sigma^2)}$ that can be packed in a sphere of radius $\sqrt{K(P+\sigma^2)}$ is the ratio of the volumes of the spheres .



- For a K-dimensional hypersphere of radius R_e the volume is equal to

$$V_K = B_K R_e^{\,K}$$

where B_K is a constant .

Therefore, the maximum number of message symbols that can be transmitted and still resolvable at the receiver is

$$M = \{B_K \ (\sqrt{K(P+\sigma^2)}\ )^K\} / \{B_K \ (\sqrt{K(\sigma^2)}\ )^K\}$$
$$= (\ 1 + P/\sigma^2\ )^{K/2}$$

- The channel capacity , the number of possible signal that can be transmitted reliably , is given by

$$C_s = (1/K) \log_2 M$$
$$= (1/2\ ) \log_2 (\ 1 + P/\sigma^2\ ) \quad \text{bits per transmission}$$

1.11.2 Band-Limited Continuous-Time Channel

- **In the band-limited continuous –time channel model , the channel bandwidth is B Hz , the power spectral density of the noise is $N_0/2$.**

  The noise power is equal to

  $$P_n = (N_0/2)\,(2B) = N_0B$$

  Then

  $$C_s = (1/2)\log_2(1 + P/N_0B) \quad \text{bits per transmission}$$

- **The channel capacity ,number of bits transmitted per second) is calculated by multiplying C_s by the number of samples per second of the signal :**

  $$C = 2B\,(1/2)\log_2(1 + P/N_0B)$$

  $$= B\log_2(1 + P/N_0B)$$

1.12 Information Capacity Theorem

- **Information Capacity Theorem , also known as Shannon-Hartley law , can be stated as follows :**

  The information capacity of a continuous channel of bandwidth B Hz , perturbed by additive white Gaussian noise of power spectral density $N_0/2$ and limited in bandwidth to B , is given by

$$C = B \log_2 (1 + P_{av} / N_0 B) \quad \text{bits/ second}$$

  where P_{av} is the average transmitted power.

- **This theorem implies that, for given average transmitted power and channel bandwidth , we can transmit information at the rate C bits per second, with arbitrarily small probability of error by employing sufficiently complex encoding systems..**

- **Shannon's Information Capacity Theorem can also be expressed as , for digital communications) as**

  $C = B \log_2 (1 + E_b R / N_0 B)$

  where E_b **= bit energy of the transmitted signal in *joules***

  R **= data rate in *bits/s***

  $N_0/2$ **= single-sided noise power spectral density**

- **In digital communications , we more often use** E_b / N_0 **, *a normalized version of SNR* , as a figure of merit .**

  $E_b / N_0 = S T_b / (N/B) = (S/N) (B/ R)$

■ Shannon limit :

- For an ideal system that transmits data at a rate

$$R_b = C$$

Then $P = R_b E_b = C E_b$

$$C/B = \log_2 \{1 + (E_b / (N_0)) (C/B)\}$$

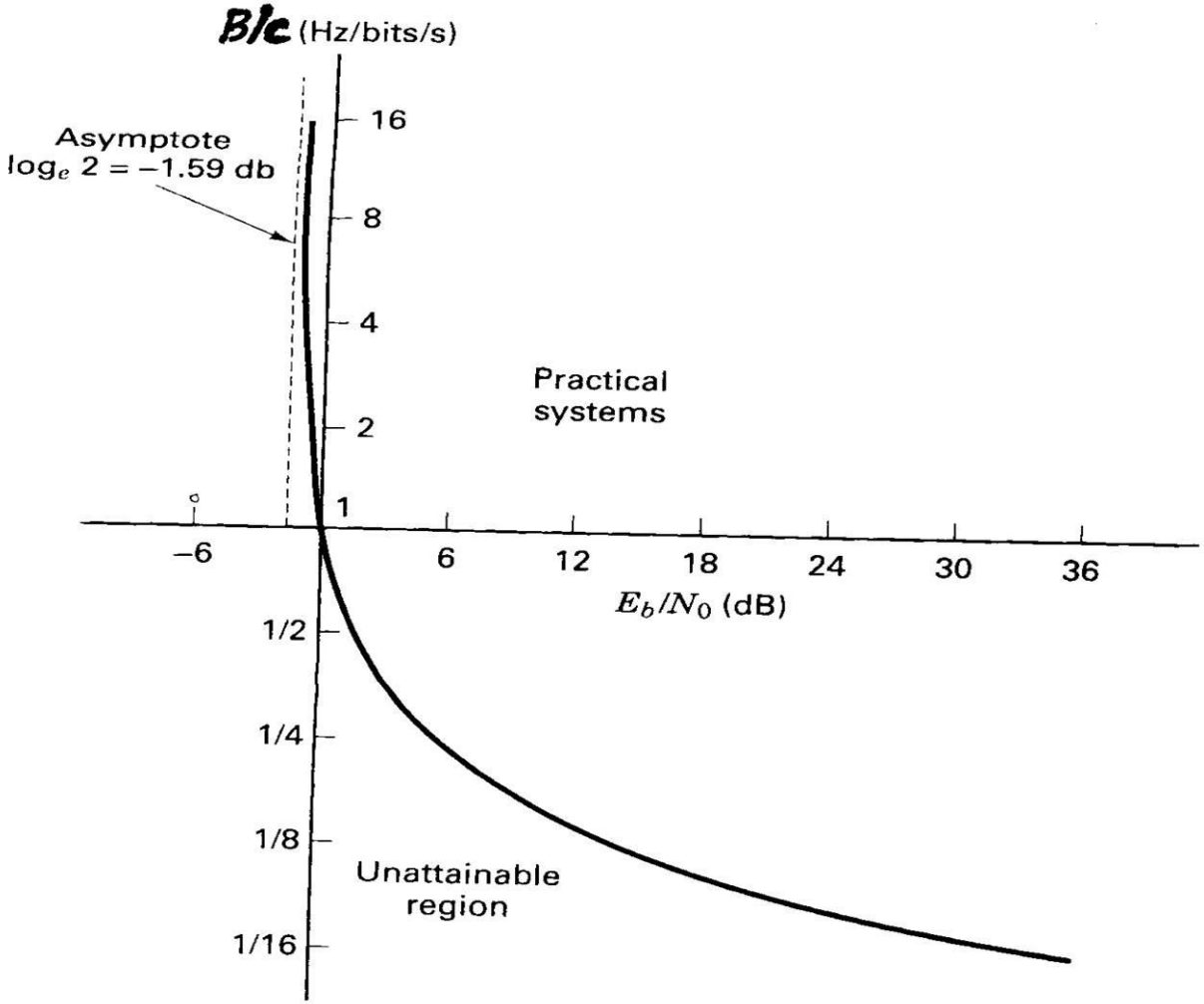
Therefore

$$E_b / N_0 = (2^{C/B} - 1) / (C/B)$$

- For infinite bandwidth, the approaches the limiting value

$$(E_b / N_0)_{\infty} = \lim_{B \to \infty} (E_b / N_0) = \ln 2 = 0.693 = -1.6 \text{ dB}$$

- This value is called Shannon limit.
  There exists a limiting value of E_b/N_0 below which there can be no error-free communication at any information rate.



Normalized channel bandwidth versus channel E_b/N_0.

Example

A communication system is to transmit data at a rate of 38.4 k bits/second over a telephone line with a bandwidth of 4 kHz.

What E_b / N_0 is required to achieve a communication reliability of one error or less in 10^3 transmitted bits ?

Solution :

$R/B = 38.4/4 = 9.6$

By information capacity theorem

$$38.4 = 4 \log_2 \{1 + (E_b / (N_0)) (38.4/4) \}$$

Thus $E_b / N_0 = 80.73 = 19$ dB

If we use enough transmitter power to obtain $E_b / N_0 = 19$ dB

or more , error correction can be used to obtain arbitrarily low error rate.

1.13 Channel Capacity of a Nonideal Linear Filter Channel

- **Recall that the capacity of an ideal , band-limited, AWGN channel is**

  $C = B \log_2 (1 + P_{av}/(N_0 B))$ **bits/ second**

  where P_{av} **is the average transmitted power .**

- **In a multicarrier system , with** Δf **sufficiently small the subchannel has capacity**

  $C_i = \Delta f \log_2 \{1 + [\Delta f P(f_i) \mid H(f_i) \mid^2 / \Delta f S_{nn}(f_i)] \}$

  where $H(f)$ **is the frequency of the channel .**

  Hence the total capacity of the channel is

  $C = \sum_{n=1}^{N} C_i$

  $= \Delta f \sum_{n=1}^{N} \log_2 \{1 + [P(f_i) \mid H(f_i) \mid^2 / S_{nn}(f_i)] \}$

- **In the limit as Δf --> 0 , we obtain the capacity of the overall channel in bits/second as**

$$C = \int_B \log_2 \{1 + [P(f_i) \mid H(f_i) \mid^2 / S_{nn}(f_i)]\} df$$

Under the constraint on $P(f)$ *given*

$$\int_B P(f)\, df \leqq P_{av}$$

the choice of $P(f)$ that maximizes C may be determined by maximizing the intrgral

$$\int_B \log_2 \{1 + [P(f_i) \mid H(f_i) \mid^2 / S_{nn}(f_i)] + \lambda P(f)\}\, df$$

where λis a Lagrange multiplier, which is chosen to satisfy the constraint, P(f) distribution as a function of frequency.

The optimum distribution of $P(f)$ is obtained from the solution to the equation

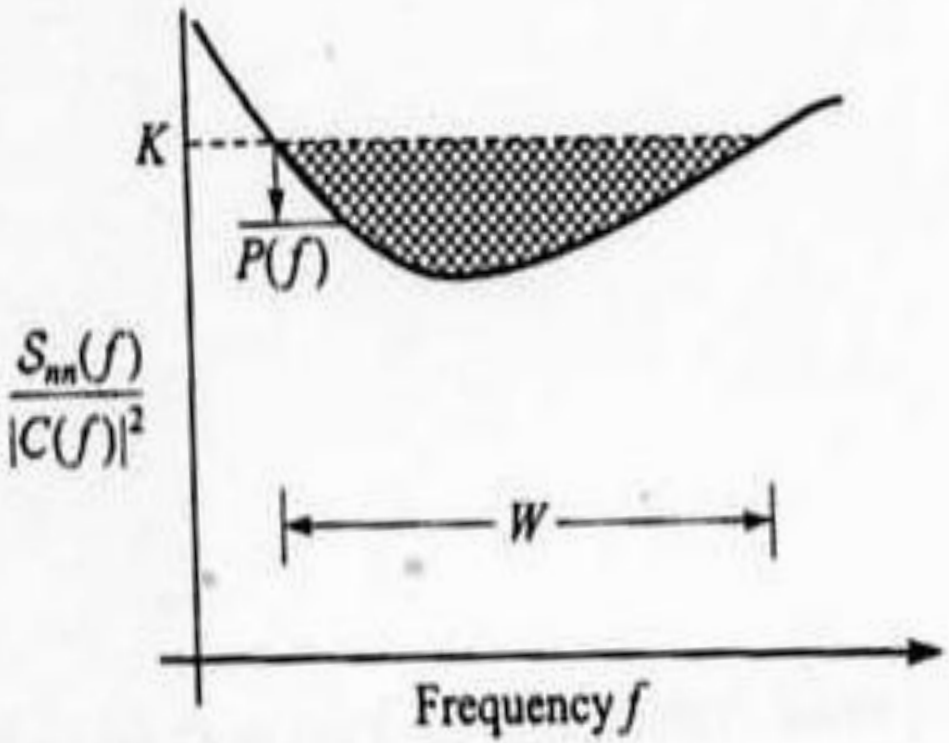
$$1 + \lambda\{P(f) + S_{nn}(f) / \mid H(f_i) \mid^2\} = 0$$

- **Therefore , $\{P(f) + S_{nn}(f) / |H(f_i)|^2\}$ must be a constant , whose value is adjusted to satisfy the average [power constraint,. That is ,**

$$P(f) = \begin{cases} K - S_{nn}(f) / |H(f)|^2 & f \varepsilon W \\ 0 & f \text{ is out of } W \end{cases}$$

Note : This expression is due to Holsinger (1964)

- **The result shows that the signal power should be high when the channel's SNR , $|H(f)|^2 / S_{nn}(f)$, is high .**
- **The transmitted power distribution is illustrated in Fig. 1.XX. If $S_{nn}(f) / |H(f)|^2$ is interpreted as the bottom of a bowl of unit depth , and we pour an amount of water equal to P_{av} into the bowl , the water will distributed itself in the bowl so as to achieve capacity. This is called *the water-filling interpretation* of the optimum power distribution as a function of frequency.**



The optimum power distribution based on water-filling interpretation.

!.14 Rate –Distortion Theorem (Lossy source coding)

1.14.1 Rate Distortion Function

- By applying the sampling theorem , the output of an analog source can be converted to an equivalent discrete-time sequence of samples. If the samples are quantized in amplitude and encoded as a sequence of binary digits , some distortion may be introduced , A loss of signal fidelity will be found during waveform reconstruction.
- Here we are to study the fundamental limits on lossy source coding given by the rate distortion function.

Now , consider the message samples $\{ x_k \}$ from the analog source . The quantized samples are denoted as $\{ \hat{x}_k \}$.

By the term *distortion* , we mean some measure of the difference between $\{ \hat{x}_k \}$ *and* $\{ x_k \}$. For example , a commonly used distortion measure is the squared-error distortion , defined as

$$d(x_k, \hat{x}_k) = (x_k, \hat{x}_k)^2$$

- If $d(x_k, \hat{x_k})$ is the distortion measure per letter , the distortion between a sequence of n samples x_k, $k=1,..,n$ and the corresponding quantized values $\hat{x_k}$, $k=1,\ldots,n$, is the average over the n source samples, i.e.,

$$d(X_n, \hat{X_n}) = (1/n) \sum_{k=1}^{n} d(x_k, \hat{x_k})$$

Since the source output is a random process , and hence the n samples in X_n are random variables . Its ensemble average is defined as the distortion D , i.e.,

$$D = E[d(X_n, \hat{X_n})] = (1/n) \sum_{k=1}^{n} d(X_k, \hat{X_k})$$
$$= E[d(X, \hat{X})]$$

where the last step follows from the assumption that the source output process is stationary .

- **Now suppose we have a memoryless source with a continuous-amplitude output X that has a PDF $p(x)$, a quantized amplitude alphabet $X\hat{}$ and a per letter distortion measure $d(x, x\hat{})$. Then the minimum rate in bits per source output that is required to represent the output x of the memoryless source with a distortion less than or equal to D is called the rate distortion function $R(D)$ and is defined as**

$$R(D) = \min_{p(x \mid x\hat{})} I(X, X\hat{})$$

subject to $E[d(X, X\hat{})] \leqq D$, where $I(X, X\hat{})$ is the mutual information between X and $X\hat{}$.

In general, the rate $R(D)$ decreases as D increases.

- **The rate distortion $R(D)$ is associated with the following fundamental source coding theorem .**

Shannon's Third Theorem

(Source Coding with a Fidelity Criterion , 1959)

A memoryless source X can be encoded at rate R for a distortion not exceeding D if $R > R(D)$.

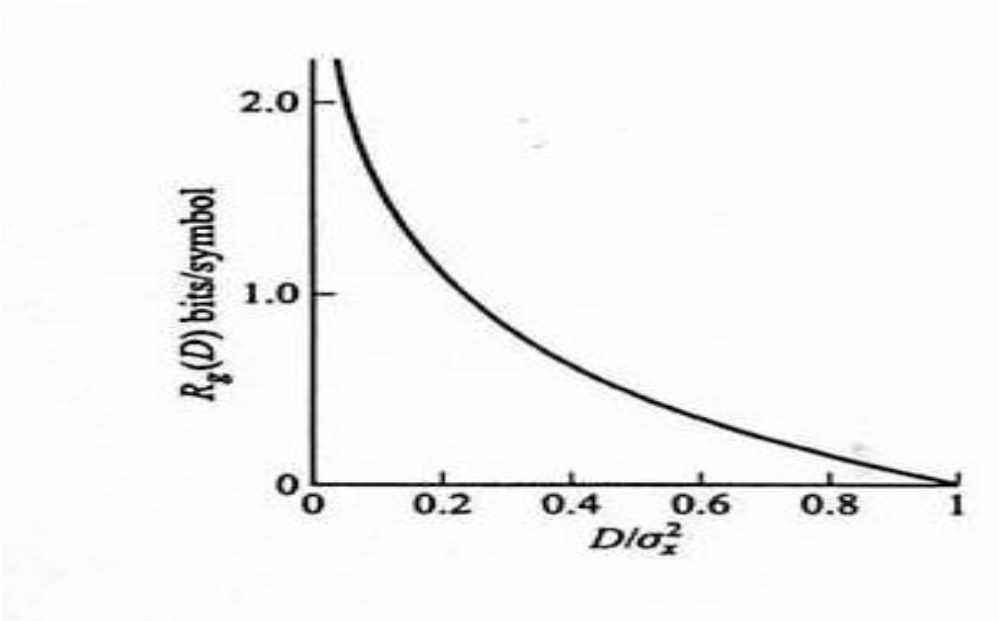
- Example : The rate distortion function of a Gaussian source with square-error distortion .

$$R_g(D) = \begin{cases} \frac{1}{2}\log_2(\sigma^2/D) & 0 \leqq D \leqq \sigma^2 \\ 0 & D > \sigma^2 \end{cases}$$

where σ^2 is the variance of the source.

Rate distortion function for a continuous amplitude , memoryless Gaussian source

.



Appendix A: Differential Entropy

- For a discrete random variable X taking the values $x_1, x_2, \ldots, x_J$ with probabilities $P(x_1), P(x_2), \ldots, P(x_J)$, the entropy $H(X)$ was defined as

$$H(X) = \Sigma_{j=0}^{J-1} p_j \log_2 (1/p_j)$$

- For analog data, we have to deal with continuous random variables or vectors. Therefore we extend the definition of entropy to continuous random variables.

  The probability distribution function (*PDF*) of the random variable X is denoted as $P_X(x)$, which will be abbreviated as $P(x)$ in the following discussions.

  The we introduce the following definition for differential entropy:

$$h(X) = \int_{-\infty}^{\infty} p(x) \log_2(1/p(x)) \, dx$$

$$= E[I(x)]$$

- ***Conditional Entropy***

$$h(X \mid Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) \log_2\{1/p(x,y)\}\, dx\, dy$$

Mutual information

$$I(X:Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) \log_2\{p(x,y)/\, p(x)p(y)\}\, dx\, dy$$

and

$$I(X:Y) = h(X) - h(X \mid Y)$$

$$= h(Y) - h(Y \mid X)$$

We also define

$$C_s = \max_{p(x,y)} I(X;Y)$$

- **We can derived the information capacity equation**

$$C_s = (1/2) \log_2 (1 + S/N) \quad \text{bits per transmission}$$

$$C = B \log_2 (1 + S/N) \quad \text{bits /sec}$$

Example :

For a random variable X with Gaussian distribution

pdf of X is given by

$$p(x) = 1/\sqrt{(2\pi\sigma^2)} \quad \exp\{ - x^2 / 2\sigma^2 \}$$

$$h(X) = \int_{-\infty}^{\infty} p(x) \log_2(1/p(x)) \, dx$$

$$= \ldots.$$

$$= (\tfrac{1}{2}) \log_2(17.1\sigma^2)$$

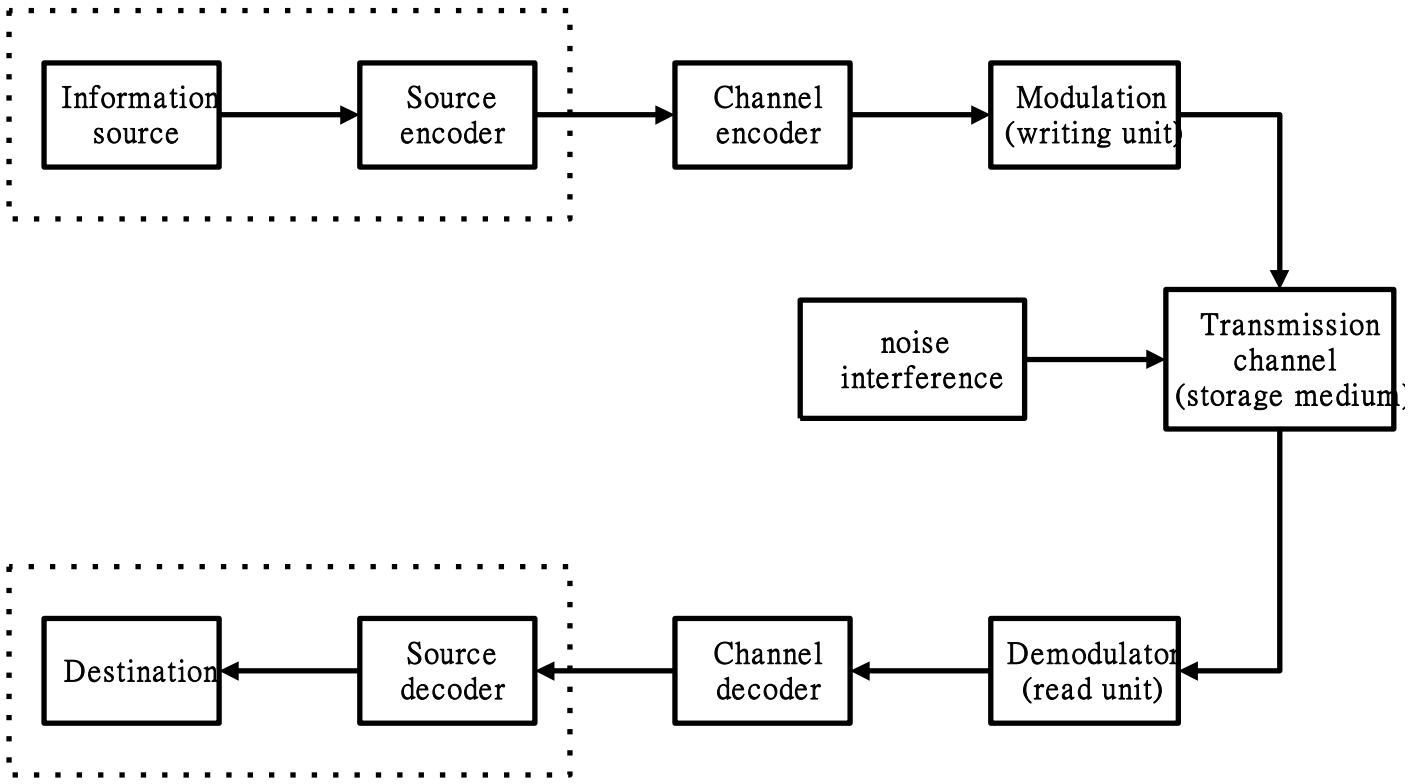
1.15 Error-Correction Coding Principle

Error- correction coding is achieved by adding properly designed redundant digits (bits) to each message. These redundant digits (bits) are used for detecting and/or correcting transmission (or storage) errors.

The redundant digits are selected in such a way that the transmitted message could be easily distinguished from other messages that could potentially be transmitted

Digital Communication System Block Diagram

communications, coding is used for controlling transmission errors induced by channel noise or other impairments, such as fading and interference, so that error-free communication can be achieved.

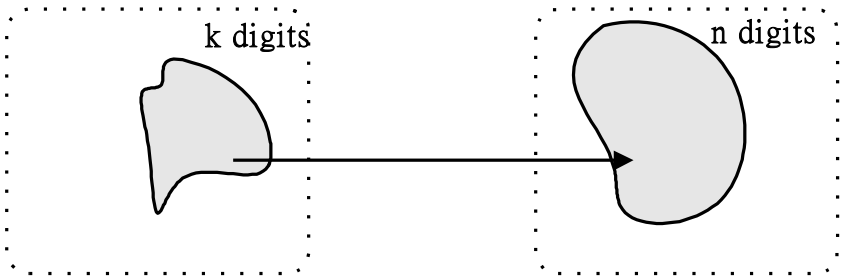


15.1.1 Types of Codes

In structure, error-correcting codes can be classified into two types : Block codes and Convolutional codes.

- **Block Coding:**

  A message of k digits is mapped into a structured sequence of n digits, called a codeword.



The mapping operation is called encoding. Each encoding operation is independent of the past encoding, i.e. memoryless. The collection of all codeword is called a "block code".

Examples : Hamming codes, Golay codes , CRC codes, BCH codes, Reed-Solomon codes.

- **Convolutional Coding:**

  An information sequence is divided into (short) blocks of k digits each. Each k-digit message is encoded into an n-digit coded block. The n-digit coded block depends not only the corresponding k-digit message block but also on m previous message blocks. That is, the encoder has memory of order m. The encoder has k inputs and n outputs.

  An information is encoded into a coded sequence. The collection of all possible code sequences is called an (n, k, m) convolutional code.

  Normally,

  $1 \leqq k \leqq 8 \quad , \quad 2 \leqq n \leqq 9$

  k/n **= code rate**

1.15.2 Type of Errors and Channels

- **Types of errors :**

Random errors and burst errors.

- **Types of Channel :**

(1) Random error channels:

Deep space channel, satellite channels,

line of sight transmission channel, etc.

(2) Burst error channels:

Radio links, terrestrial microwave links, wire and cable transmission channels, etc.

1.15.3 Decoding

- **Suppose a codeword corresponding to a message is transmitted over a noisy channel.**

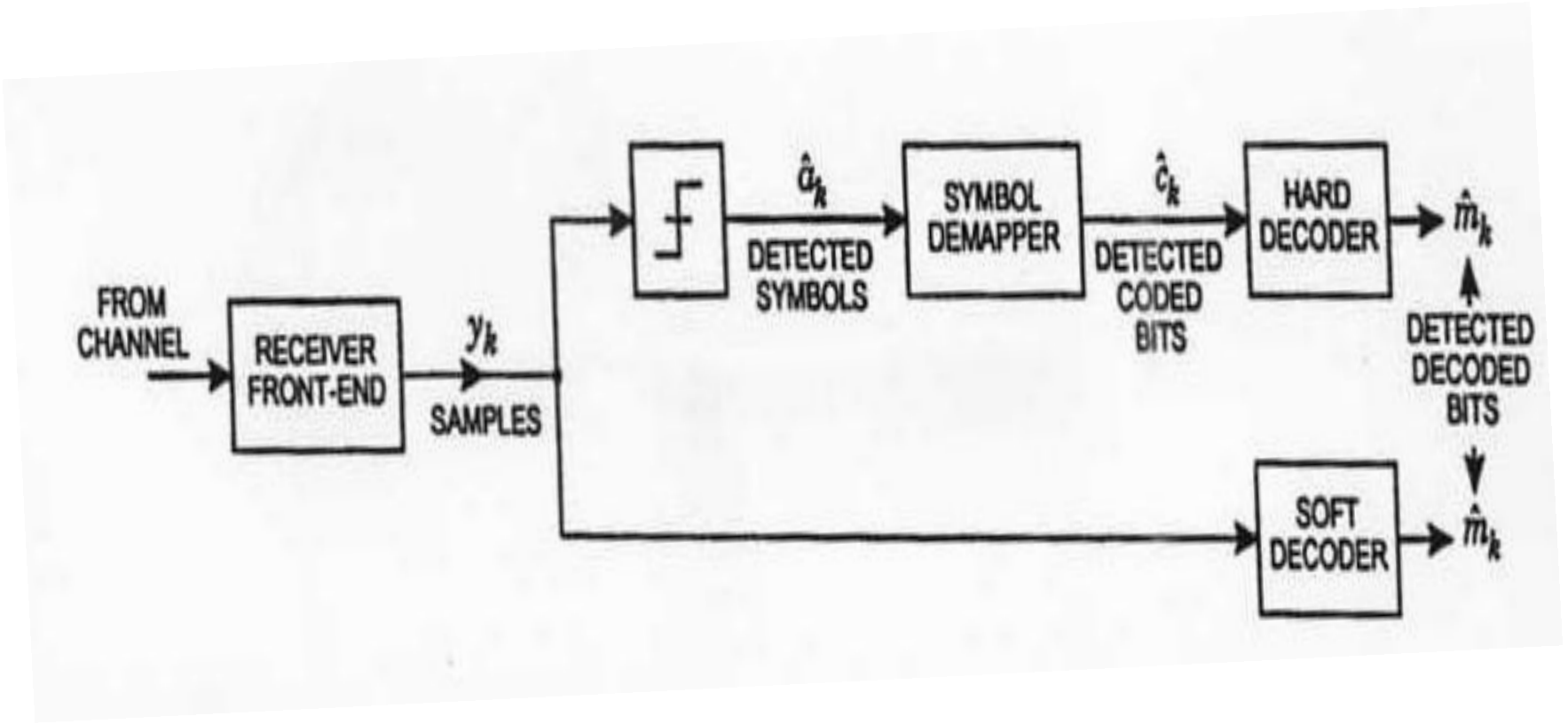
  Let $\bar{r}$ be the received sequence. Based on $\bar{r}$, the encoding rules and the noise characteristics of the channel, the receiver (or decoder) makes a decision which message was actually transmitted.

  This decision making operation is called "decoding". The device which performing the decoding operation is called a "decoder".

- **Two fundamentally different types of decoding are used : hard and soft .**

  With hard decoding , the receiver first make hard decisions about the transmitted symbols using a memoryless slicer . The hard decoder operates on these hard decisions .

- **A soft decoder , by contrast , operates directly on the incoming continuous-valued signal without making intermediate decisions about the transmitted symbols , as illustrated in Fig. 1.xx**



- **Decoding techniques can also be classified as algebraic decoding and probabilistic decoding .**

--The algebraic coding and decoding dominated the first several decades of the field of channel coding.

Over a BSC , the optimal decoding rule for a (n,k) block code is to decode to the codeword closest in Hamming distance to the received n-tuple. With this rule , a code with minimum distance d can correct (d-1)/2 or fewer channel errors (assuming that d is odd), but can not correct some patterns containing a greater number of errors. This is also known as bounded- distance decoding.

-- Probabilistic coding and decoding was more directing inspired by Shannon's probabilistic approach to coding.

Probabilistic decoder typically use soft-decision (reliability) information, both as input 9from the channel outputs), and at intermediate stages of the decoding process.

1.16 MAP Decoding and ML Decoding

- Suppose the codeword c corresponding to a certain message m is transmitted.

  Let r be the corresponding output of the demodulator.

- The decoder produces an estimate of the message $\hat{m}$ based on r.
- An optimum decoding rule is that minimize the probability of a decoding error. That is, $p(\hat{c} \neq c \mid r)$ is minimized.

  Or, equivalently, maximizing $p(\hat{c} = c \mid r)$.
- The decoding error is minimized for a given r by choosing $\hat{c}$ to be a codeword c that maximizes

  $$p(c \mid r) = p(r \mid c)\, p(c) / p(r)$$
- That is, $\hat{c}$ is chosen to be the most likely codeword, given that r is received.

  If knowledge or an estimation of $P(c)$ is used for decoding, the technique is called **MAP decoding**.

- **Suppose all the messages are equally likely. An optimum decoding can be done as follows:**

  For every codeword c_j , compute the conditional probability $p(r \mid c_j)$

  The codeword c_j with the largest conditional probability is chosen as the estimate $\hat{c}$ for the transmitted codeword c .

  This decoding rule is called the Maximum Likelihood decoding (MLD).

Remark :

$\bar{r}$

Bounded Distance Decoding

- **Given a received word r , a *t*-error correcting, bounded distance decoder selects that codeword which minimizes d_H(r,c) if and only if there exists c ^ such that d_H(r, c ^) $\leqq$ t .**

  If no such c exists, then a decoding failure is declared.

- **The bounded distance decoding is usually an "incomplete decoding" since it decodes only those received words lying in a radius-*t* sphere about a codeword.**

$\bar{c}$ $\bar{c}$

$$d_H(\bar{r}, \bar{c}) \leq \mathrm{t}$$

History of error-correction codes

- The field of channel coding started with Shannon's 1948 landmark paper.
- The first nontrivial code to appear in the literature was (7, 4 ,3) Hamming code , mentioned by Shannon in his 1948 paper. Richard Hamming was a colleague of Shannon at Bell Labs.

  Hamming developed a class of single-error correcting binary linear block codes.

  R.W. Hamming, " Error detecting and error correcting codes ," Bell Sys. Tech. vol.29 ,pp.147-160, 1950 .
- Shortly after the publication of Shannon's paper , the Swiss mathematician Marcel Golay published a half-page paper with a " perfect" binary linear (23 ,12 ,7) triple-error

  correcting code.

- Another early class of error-correcting code was the Reed-Muller (RM) codes, which were introduced in 1954 by David Muller and then reintroduced shortly thereafter with an efficient decoding algorithm by Irving Reed.
- In the 1960s , research in channel coding was dominated by the development of algebraic block codes, particularly cyclic codes.

  Cyclic codes are codes that are invariant under cyclic shifts of a *n*-tuple codewords. They were first investigated by Eugene Prange in 1957.

  E.Prange, " Cyclic error-correcting codes in two symbols ," Air Force Cambridge Res. Center, Cambridge, MA, Tech. Note AFCRC-TN-57-103, Sep.1957.

  Cyclic have a nice algebraic theory and attractively simple encoding and decoding procedures based on cyclic shift-register implementation

- **The BCH codes are, first discovered by A. Hocquenghem in 1959 and independently by R.C. Bose and D. K. Ray-Chaudhuri in 1960.**

  The first decoding algorithm for binary BCH codes was devised by W.w. Peterson in 1960.

  Since then Peterson's algorithm has been refined by Elwyn Berlekamp, J.L.Massey , R.Chien , G.D.Forney and many others.

  BCH codes are a large class of multiple random error-correcting codes.
- **The RS codes were discovered in 1960 by I. Reed and G. Solomon at MIT . They are nonbinary cyclic codes with code symbols from a Galois field.**
- **An important property of RS and BCH codes is that they can be efficiently decoded by algebraic decoding .**

Convolutional Codes

- **Elias' invention of convolutional codes :**

  In 1955 , Peter Elias invented the convolutional codes . These codes are simpler to encode than general linear block codes. Elias showed that convolutional have the same average performance as randomly chosen codes.

  The Fano's sequential decoding algorithm for convolutional codes was introduced by R.M. Fano in 1963.

  Subsequently ,J.L. Massey proposed a threshold decoding method for convolutional codes in 1963.

- **In 1967, Andy Viterbi introduced what became known as the Viterbi algorithm (VA) as an "asymptotically optimal " decoding algorithm for convolutional codes. It was quickly recognized that VA was actually an optimal decoding algorithm.**

Product Codes and Concatenated Codes

- Before inventing convolutional codes, Elias had invented another class of codes known as product codes(1954).

  P.Elias, " Error-free coding , " IRE Trans. Inform. Theory , vol.IT-4 ,pp.29-37, Sep.1954 .

- In 1966, G.D. Forney introduced concatenated codes which involves a serial cascade of two linear block codes.

  G.D. Forney, Jr. , Concatenated codes, Cambridge, MA : MIT Press,1988

Trellis –Coded Modulation

- In 1982, G.Ungerboeck introduced trellis-code modulation for bandwidth-limited channel .

  Ungerboeck relized that in the bandwidth-limited regime , the redundancy needed for coding should be obtained by expanding the signal constellation while keeping the bandwidth fixed. From capacity calculation , he showed that doubling the signal constellation should suffice.

Iterative Error Correction --- The Turbo Revolution

- In 1972, L.R. Bahl, J.Cocke, F. Jelinek , and J. Raviv introduced BCJR algorithm .
- In 1989 , J. Hagenauer and P. Hoeher introduce soft- output Viterbi Algorithm (SOVA) for decoding convolutional codes .
- In 1993 , at ICC of IEEE in Geneva , C. Berrou , A.Glavieux , and P.Thitimajshima stunned the coding research community by introducing a new class of turbo codes which can achieve near-Shannon-limit performance with modest decoding complexity.

Comments to the effect of " It can't be true ; they must have made a 3 dB error" were widespread. However , within the next year various laboratories confirmed these astonishing results, and the " turbo revolution " was launched.

- **1n 1995, D.J.C. MacKay and R.M. Neal re-discovered the low-density parity-check (LDPC) codes.**

  MacKay showed that in practice moderate-length LDPC codes (10^3 – 10^4 bits) could attain near-Shannon-limits performance .

  The results kicked off a similar explosion of research on LDPC codes, which are currently seen as competitors to turbo codes in practice.