
Chapter 1

Introduction

In this chapter, fundamental knowledge for TCM is presented. And the idea of conventional channel coding is also briefly introduced, including the problems that have been encountered and the concept that TCM presents to remedy the problems.

1.1 Background

1.1.1 Channel Capacity and Channel Coding

The concept of channel capacity and its associated theorems are the most important concepts in communications. The capacity of a channel is the number of information bits per seconds that can theoretically be transmitted over the channel. The information theory tells us that, if the output of the source encoder is a binary sequence with a rate less than channel capacity, the transmission error rate can be reduced by using forward error correction (FEC) techniques without increasing transmitter power. The price paid for such techniques is the increased complexity and bandwidth.

1.1.2 Conventional Sense

In conventional sense, channel coding and modulation techniques are separated as follows :

Coding occurs before modulation at the digital level and generally adds redundancy into information bits, hence requires additional bandwidth or in another way of speech, lowers the bandwidth efficiency or the effective information rate. To compensate for the rate loss, two methods are generally applied : **1)** increasing the modulation rate if the channel permits bandwidth expansion, or **2)** enlarging the signal set of the modulation system if the channel is band-limited.

Modulation occurs after coding and acts as an algorithm of mapping information bits into signal points of the selected constellation. At the receiving end, hard decoding occurs after demodulation. The theoretical performance loss due to hard decoding amounts to about 2 dB compared to soft decoding. To remedy this performance loss, very powerful codes are required. This implies consequently that either a convolutional code that has a very large constraint length or a block code that has a very large block size is needed, hence induces even greater complexity. Despite all the efforts that can be made, disappointing results are still obtained if the coding and modulation are performed independently.

1.1.3 Driving Force for New Code Design

As previously mentioned, coding occurs at digital level if it is treated separately from modulation. Thus, hard-decision decoding must be applied at the receiving end. Since only discrete code symbols are of concern, Hamming distance (the number of symbols in which two code sequences differ) seems to be the appropriate measure of

distance for decoding. However, hard-decision decoding causes an irreversible loss of information in the receiver. A soft-decision decoder, which operates directly on unquantized “soft” output samples of the channel, is therefore an ideal choice for the new code design.

Let the samples be $r_n = a_n + w_n$ (either real for one-dimensional modulation or complex for two-dimensional modulation), where a_n stands for the discrete signals sent into the channel by the modulator and w_n stands for the samples of an additive white Gaussian noise (AWGN) process. The rule of making decisions is to determine, among the set C of all signal sequences produced by the cascaded encoder and modulator, the sequence $\{\hat{a}_n\}$ with minimum squared Euclidean distance from $\{r_n\}$, that is, the sequence $\{\hat{a}_n\}$ which satisfies the equation :

$$\left| r_n - \hat{a}_n \right|^2 = \underset{\{a_n\} \in C}{\text{Min}} \sum \left| r_n - a_n \right|^2 . \quad (1)$$

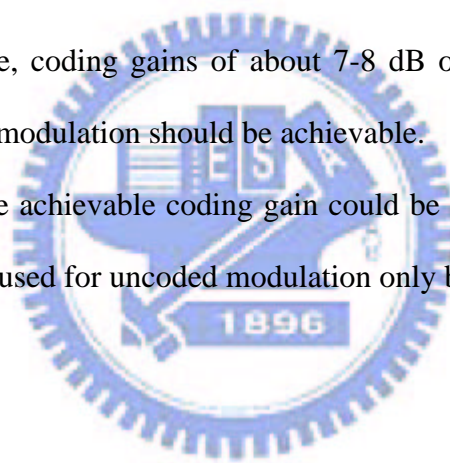
The motivation of the developer of TCM came from work on multilevel systems that employ Viterbi algorithm to improve signal detection in the presence of inter-symbol interference. The work not only provided ample evidence of the importance of Euclidean distance between signal sequences but also conveyed the idea that codes should be designed for maximum free Euclidean distance rather than Hamming distance.

1.1.4 Trellis-Coded Modulation

Trellis-coded modulation, which is also known as TCM, has been in existence

for the last couple of decades. It is regarded as a combination of coding and modulation technique for digital transmission over band-limited channels. At the transmitting end, TCM schemes employ redundant non-binary modulation in conjunction with a finite state encoder that selects the modulation signals to generate coded signal sequences. In the receiver, a soft-decision maximum-likelihood sequence decoder (MLSD) is used to decode the noisy signals. Ungerboeck, the developer of TCM, computed the capacity of channels with additive Gaussian noise for the case of discrete multilevel modulation at the channel output. As will be seen in the next section, the results allowed making two observations [2] :

1. In principle, coding gains of about 7-8 dB over conventional uncoded multilevel modulation should be achievable.
2. Most of the achievable coding gain could be obtained by expanding the signal sets used for uncoded modulation only by a factor of two.



1.2 Channel Capacity of Multilevel/Phase Modulation Channels

1.2.1 In Terms of Channel Capacity [1]

Before we go on and deal with the code-design problem with expanded modulation signal sets, it is appropriate to examine in terms of channel capacity the limits to performance gains which may be achieved. For the time being, we consider only one- and two-dimensional modulation and inter-symbol interference-free signaling over band-limited channels with only additive white Gaussian noise

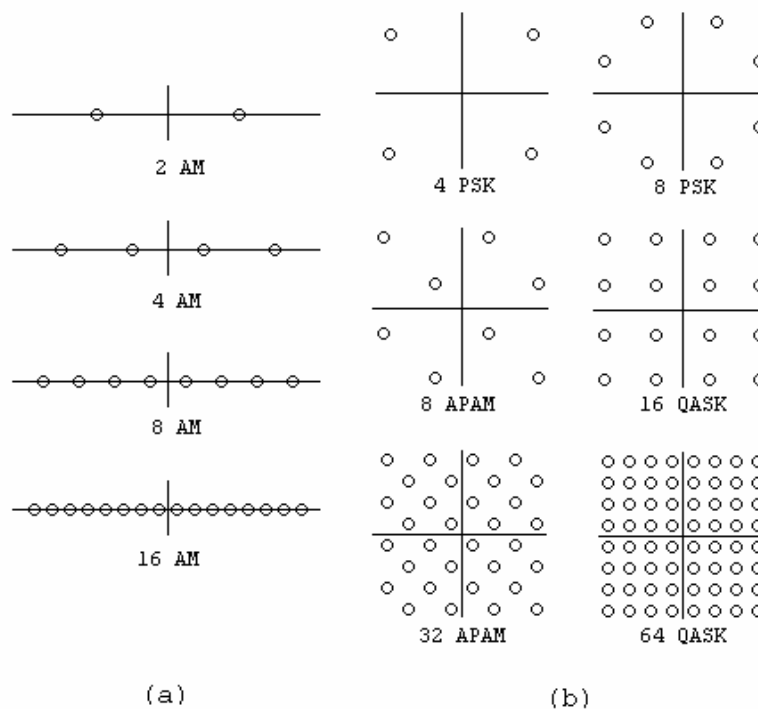
(AWGN) for simplicity. Perfect timing and carrier-phase synchronization are also assumed. If $r_n = a_n + w_n$, and w_n is an independent, normally distributed noise sample with zero mean and variance σ^2 along each dimension. The average signal to noise ratio (SNR) is defined as :

$$\begin{aligned} \text{SNR} &= \frac{E\{|a_n|^2\}}{E\{|w_n|^2\}} \\ &= \text{(i) } \frac{E\{|a_n|^2\}}{\sigma^2} \text{ for 1-D modulation , or (ii) } \frac{E\{|a_n|^2\}}{2\sigma^2} \text{ for 2-D modulation} \end{aligned} \quad (2)$$

Figure 1-1 illustrates the constellations of (a) one-dimensional channel signal sets, and (b) two-dimensional channel signal sets. We have assumed that the normalized average signal power is unity, that is :



$$E\{|a_n|^2\} = 1 \quad (3)$$



【Figure 1-1. Channel signal sets of (a) one- (b) two-dimensional constellation】

We extend a well-known formula for the capacity of a discrete memoryless channel [4] to the case of continuous-valued output and the result yields :

$$C = \max_{Q(0)\dots Q(N-1)} \sum_{k=0}^{N-1} Q(k) \int_{-\infty}^{+\infty} p(r/a^k) \cdot \log_2 \left\{ \frac{p(r/a^k)}{\sum_{i=0}^{N-1} Q(i)p(r/a^i)} \right\} dr \quad \text{in bit/T} \quad (4)$$

N denotes the number of discrete channel input signals $\{a^0 \dots a^{N-1}\}$ and $Q(k)$ is the *a priori* probability associated with a^k . Note that AWGN has been assumed, we can therefore substitute $p(r/a^k)$ in (4) :

$$p(r/a^k) = \begin{cases} (2ps^2)^{-1/2} \cdot \exp\left[-|r-a^k|^2/2s^2\right] \dots\dots(a) \\ (2ps^2)^{-1} \cdot \exp\left[-|r-a^k|^2/2s^2\right] \dots\dots(b) \end{cases} \quad (5)$$

Next we further assume that all possible channel input signals are equally probable, that is, each of the N elements in the set $\{a^0 \dots a^{N-1}\}$ has the probability $1/N$. With this assumption, we can therefore omit the maximization operation over the $Q(k)$ in (4) and replace it with $1/N$. We can also replace the integration operation in (4) with the expectation value operation $E\{\}$ over the normally distributed noise variable w . Now we can rewrite (4) in the form :

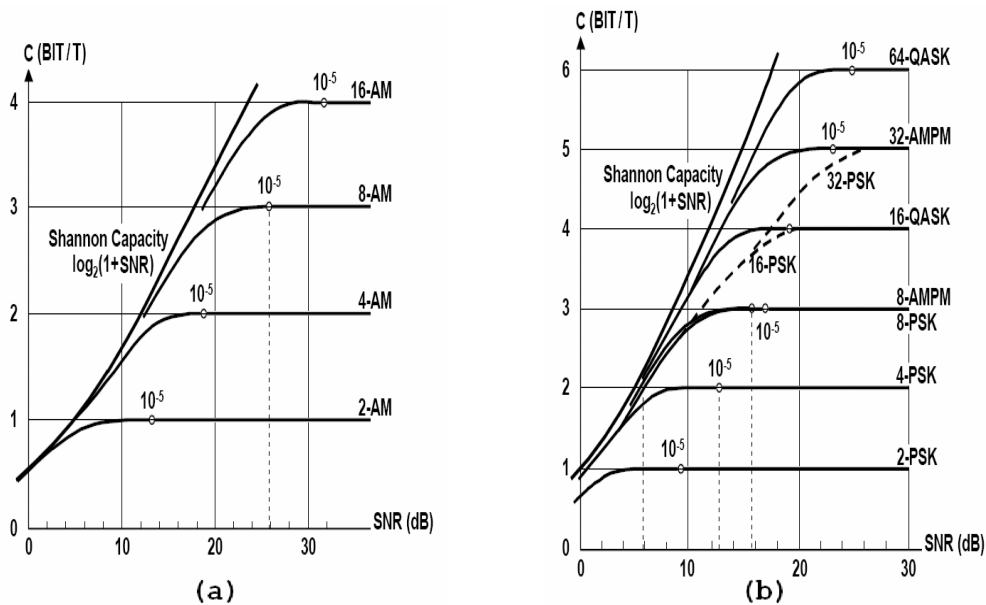
$$\begin{aligned}
C^*_{Q(k)=1/N} &= \frac{1}{N} \sum_{k=0}^{N-1} E \left\{ \log_2 \left[\frac{p(r/a^k)}{\frac{1}{N} \sum_{i=0}^{N-1} p(r/a^i)} \right] \right\} \\
&= \log_2 N + \frac{1}{N} \sum_{k=0}^{N-1} E \left\{ \log_2 \left[\frac{\exp[-|r-a^k|^2/2\mathbf{s}^2]}{\sum_{i=0}^{N-1} \exp[-|r-a^i|^2/2\mathbf{s}^2]} \right] \right\} \\
&= \log_2 N - \frac{1}{N} \sum_{k=0}^{N-1} E \left\{ \log_2 \sum_{i=0}^{N-1} \exp \left[\frac{-|r-a^i|^2 - |r-a^k|^2}{2\mathbf{s}^2} \right] \right\} \quad (6)
\end{aligned}$$

From above, if we define the normally distributed noise variable $w = r - a^k$ and $r - a^i = r - a^k + a^k - a^i = w + a^k - a^i$, (6) can be further rewritten as :

$$C^*_{Q(k)=1/N} = \log_2 N - \frac{1}{N} \sum_{k=0}^{N-1} E \left\{ \log_2 \sum_{i=0}^{N-1} \exp \left[\frac{-|w + a^k - a^i|^2 - |w|^2}{2\mathbf{s}^2} \right] \right\} \quad (7)$$

Note that the normally distributed noise variable w in (7) is real with variance \mathbf{s}^2 for (a) one-dimensional condition, and complex with variable $2\mathbf{s}^2$ for (b) two-dimensional condition.

We can evaluate the channel capacity C^* in (7) by Monte Carlo averaging which requires a Gaussian random number generator. As a result, C^* is plotted as a function of SNR in Figure 1-2 for the signal sets depicted in Figure 1-1. Also in Figure 1-2, a special case of the value of SNR at which the symbol-error probability $P_r(e) = 10^{-5}$ can be achieved in uncoded transmission is particularly indicated.



【Figure 1-2. Channel capacity of band-limited AWGN channels with discrete-valued input and continuous-valued output. (a) One-dimensional modulation. (b) Two-dimensional modulation.】

1.2.2 Interpretation of The Channel Capacity

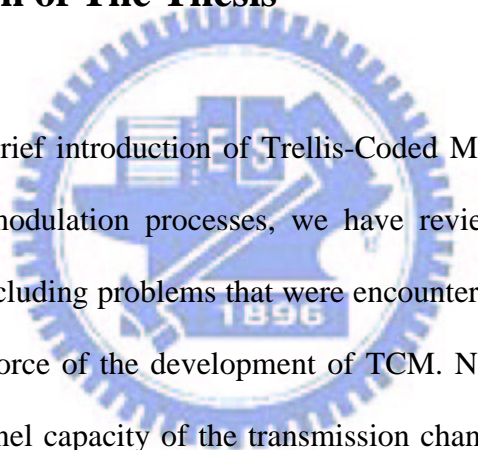
To interpret Figure 1-2, we consider as an example the transmission of 2bit/T using uncoded 4-PSK modulation where $P_r(e) = 10^{-5}$ occurs at SNR = 12.9 dB. However, if we decide to double the number of channel signals, that is, to choose 8-PSK modulation instead of 4-PSK modulation, 2bit/T transmission of approximately the same symbol-error probability $P_r(e) = 10^{-5}$ can be achieved at a lower SNR = 5.9 dB. Observation reveals that a gain of about 7 dB is achieved only by expanding the signal sets by the factor of two.

Consider now another example of transmission of 3bit/T using uncoded 8-AM modulation in Figure 1-2(a). The symbol-error probability $P_r(e) = 10^{-5}$ occurs at SNR = 25.9 dB. However, if one-dimensional 8-AM modulation is replaced by two-

dimensional 8-PSK modulation, the same symbol-error probability can theoretically occur at $\text{SNR} = 15.7 \text{ dB}$, which amounts to a gain of over 10 dB. The implication of this observation is that, even with the same amount of signal points, more gains can be achieved by expanding the dimension of the constellation of the signal sets.

The two examples mentioned above are marked in Figure 1-2 (a) and (b) with dashed lines.

1.3 Organization of The Thesis



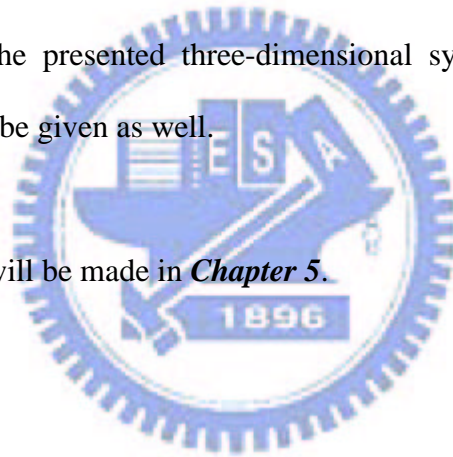
We have made a brief introduction of Trellis-Coded Modulation. Starting from issues of coding and modulation processes, we have reviewed in *section 1.1* the background of TCM, including problems that were encountered in conventional cases and hence the driving force of the development of TCM. Next in *section 1.2*, more details concerning channel capacity of the transmission channel and the fact that the number of signals points in a certain modulation scheme or the dimension of a certain modulation constellation can cause the error performance to differ has been given.

In the rest of this thesis, one of the most famous TCM schemes, which is also called Ungerboeck's code, will be introduced in detail in *Chapter 2*, including the general principles that have been proven useful in code design, the construction of the code, and the key factors in decoding process that induce great coding gains over conventional uncoded modulation schemes.

Next in *Chapter 3*, the concept of three-dimensional signal constellation will be presented as the original ideas of this thesis. We will show how we construct three-dimensional constellations by referring to two-dimensional ones. The merits of three-dimensional constellations over two-dimensional ones will be shown in a heuristic manner.

In *Chapter 4*, computer-aided simulation results and comparison over bandwidth efficiency, average signal power, and error performance between Ungerboeck's code and conventional uncoded system will be given. Also, computer-aided simulation results and comparison over bandwidth efficiency, average signal power, and error performance between the presented three-dimensional systems and systems using Ungerboeck's code will be given as well.

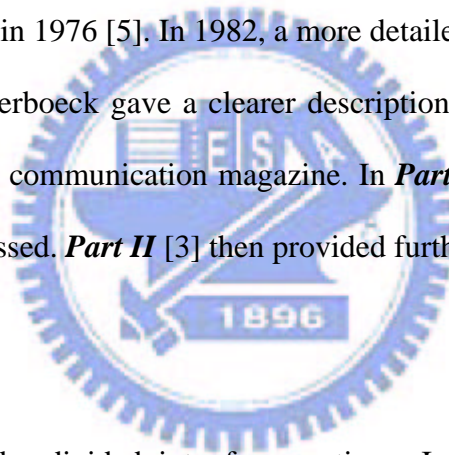
Final conclusions will be made in *Chapter 5*.



Chapter 2

Ungerboeck's TCM Scheme

The main attraction of TCM comes from the fact that it allows the achievement of significant coding gain over conventional uncoded multilevel modulation without compromising bandwidth efficiency. The first TCM schemes were proposed by Ungerboeck and Csajka in 1976 [5]. In 1982, a more detailed publication was released [1]. Later in 1987, Ungerboeck gave a clearer description of TCM in the form of a two-part article in IEEE communication magazine. In *Part I* [2], examples of simple TCM schemes are discussed. *Part II* [3] then provided further insight into code design and performance.



This chapter will be divided into four sections. In the first and the second sections, *Four-State Trellis Code for 8-PSK Modulation* and *Eight-State Trellis Code for Amplitude/Phase Modulation* will be included, respectively. The third section will take a look at the principles of the design of trellis-coded modulation schemes. In the final section of this chapter, we will introduce a systematic approach and some important principles when constructing trellis-coded modulation schemes.

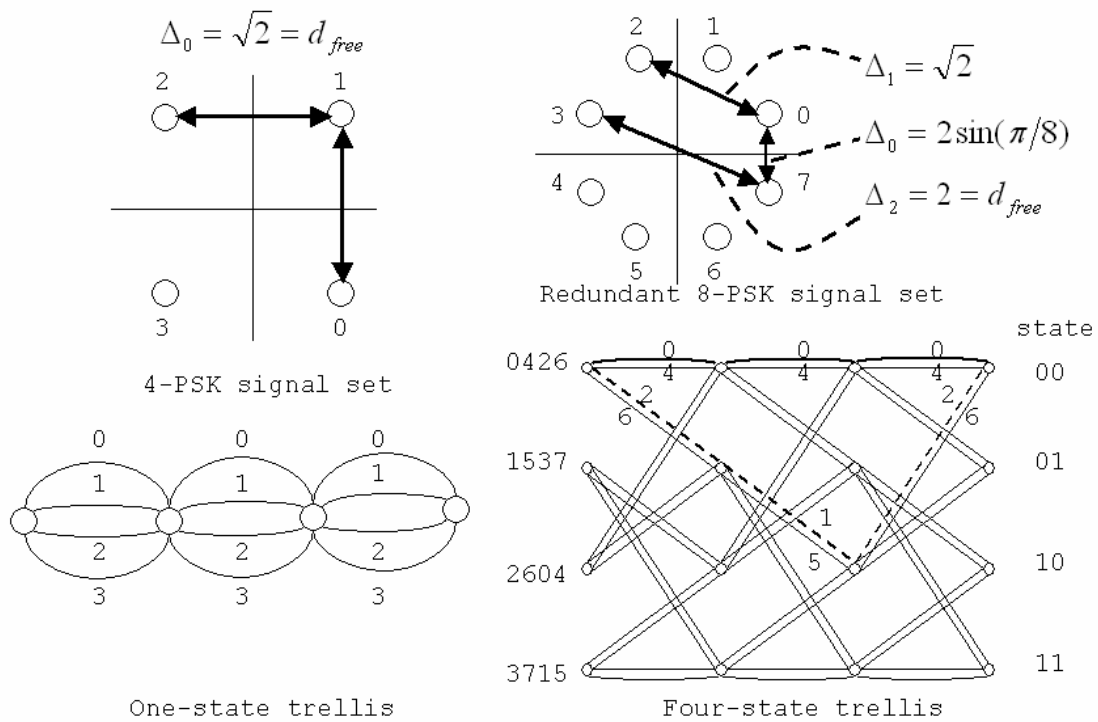
2.1 Four-State Trellis Code for 8-PSK Modulation [2]

2.1.1 Signal Sets And Trellis Diagrams

This particular example of a coded 8-PSK scheme was the first TCM scheme found by Ungerboeck in a heuristic manner in 1975. It achieved a significant coding gain over uncoded modulation. *Figure 2-1(a)* depicts the signal set and the trellis diagram for uncoded 4-PSK modulation. The one-state trellis diagram is trivial and only shown to illustrate the uncoded 4-PSK from the viewpoint of TCM. *Figure 2-1(b)* depicts the signal set and the trellis diagram for coded 8-PSK modulation with four trellis states. Every path connecting two states through the trellis diagrams shown in *Figure 2-1 (a)* and *(b)* represents an allowed signal sequence. For both systems in *(a)* and *(b)*, four transition paths originate from the same state, or in another way of speaking, remerge into the same state. For the following discussion, the specific encoding of information bits into signals is not important, which has been proven in [1].

We first take a look at *Figure 2-1(a)*. The one-state trellis diagram appears to have four “parallel” transitions between any two states. The property indicates that there is no relationship between neighboring signal points in a sequence, that is, there is no sequence encoding for this system using 4-PSK modulation. Therefore, the state diagram imposes no restrictions on sequences to be transmitted. Based on the fact, the decoder at receiving end can make independent decision for each noisy 4-PSK signal received. As shown in *Figure 2-1(a)*, the smallest distance between 4-PSK signals is $\sqrt{2}$, as denoted as Δ_0 . In accordance with common terminology used in sequence-

coded systems, we call this smallest distance the “free distance” of uncoded 4-PSK modulation.



【Figure 2-1. (a) Uncoded four-phase modulation (4-PSK). (b) Four-state trellis-coded eight-phase modulation (8-PSK).】

Next in **Figure 2-1(b)**, the four-state trellis diagram for the coded 8-PSK scheme appears that there is a pair of “parallel” transitions between any two connected states and that each of the four states has two succeeding states. Therefore, each of the four states has two sets of parallel transitions connecting to its succeeding states. **Figure 2-1(b)** also shows the numbering of the 8-PSK signals and relevant distances between these signals : $\Delta_0 = 2\sin(\pi/8)$, $\Delta_1 = \sqrt{2}$, and $\Delta_2 = 2$.

2.1.2 Rules for Assignment of The Signals

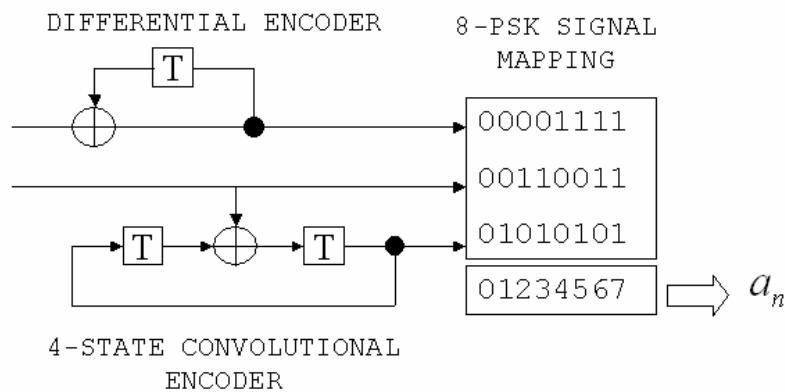
The specific assignment of the 8-PSK signals to the transitions in the four-state trellis follows certain rules [1] [2] :

1. All 8-PSK signals should occur with equal frequency and with a fair amount of regularity and symmetry.
2. Transitions originating from or remerging into the same state are labeled with signals with no smaller distance than $\Delta_1 = \sqrt{2}$ between them. For example, (0,2,4,6) or (1,5,3,7) as indicated in Figure 2-1(b).
3. The “parallel” transitions between any two connected states are associated with signals with maximum distance $\Delta_2 = 2$ between them. For example, (0,4), (2,6), (1,5), or (3,7).

Any two signal paths in the four-state trellis of **Figure 2-1(b)** that diverge from one state and remerge later into another after more than one transition have at least “squared” distance $\Delta_1^2 + \Delta_0^2 + \Delta_1^2 = \Delta_2^2 + \Delta_0^2$ between them. For example, the solid thick line and the dashed line in **Figure 2-1(b)** represent the signal sequences 0-0-0 and 2-1-2, respectively. The distance between the first “0” in the first sequence and the first “2” in second sequence contributes to the first factor Δ_1^2 in the squared distance, as can be seen in the 8-PSK signal set in **Figure 2-1(b)**. And the distance between the second components in both the first and the second sequence contributes to the second squared distance factor. For squared distances between any two non-overlapping sequence paths, the same mathematical operation applies so on and so forth.

The distance between such paths is greater than the distance between the signals assigned to parallel transitions (Δ_2^2), which thus is found as the free distance in the

four-state 8-PSK code : $d_{free} = 2$. As has been mentioned in previous sections, the free distance of uncoded 4-PSK code is $\sqrt{2}$. Therefore, if expressed in decibels, this decrement in free distance amounts to an improvement of 3 dB in error performance. **Figure 2-2** illustrates one possible realization of an encoder-modulator for the four-state coded 8-PSK scheme.



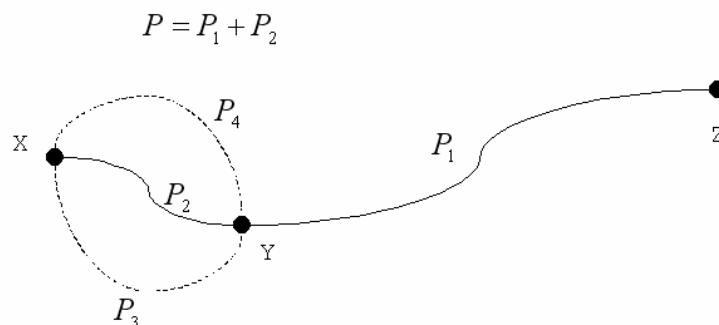
【Figure 2-2. A realization of an encoder-modulator for four-state coded 8-PSK.】

For any state transition along any coded 8-PSK sequence transmitted, there exists only one nearest-neighbor signal at the free distance, which is the 180° phase rotated version of the transmitted signal. Hence, the code is said to be “invariant” to a signal rotation by 180° , but not to any other rotation.

2.1.3 The Soft-Decision Decoding

The soft-decision decoding is accomplished in two steps : In the first step, which is called the “subset decoding”, within each subset of signals assigned to parallel transitions, the signal which is closest to the received channel output is chosen. The squared distances between the chosen signals and the channel output are also

determined and stored together with the chosen signals. In the second step, the soft-decision Viterbi algorithm is applied to find the signal path through the code trellis with the minimum sum of squared distances from the noisy channel output sequence received. In the second step, only the signals that are chosen in the first step are considered. For more details concerning the soft-decision Viterbi decoding algorithm, [6] is recommended for tutorial descriptions. Here we briefly summarize several essential points as follows :



【Figure 2-3. Shortest path diagram】

In **Figure 2-3**, suppose that the shortest path from point X to point Z with intermediate point Y is P , where $P = P_1 + P_2$. The shortest path from point X to point Y must be P_2 . This implies that if the optimum signal paths (in accordance with the Maximum-likelihood terminologies : the paths that have the minimum sum of squared distances from the channel output) from all of the trellis states at time n to the infinite past are known, the Viterbi algorithm iteratively extends these paths from time n to time $n+1$ by choosing among all of the paths to all succeeding trellis states, the best paths to the next states at time $n+1$. All other unselected paths are discarded and the selected paths are called the “survivors” of the paths. After the selecting-discarding process, we can then determine the optimum signal paths from all

of the trellis states at time $n + 1$ to the infinite past.

2.1.4 The Free Distance And Error Events

Let the received signals be disturbed by uncorrelated Gaussian noise samples with variance σ^2 in each signal dimension. That is, if the signal is one-dimensional, the variance of the signal is σ^2 ; if the signal is two-dimensional, the variance of the signal is $2\sigma^2$. There can be two kinds of error events: the first kind of error event occurs when the decoder makes a wrong decision among the signals associated with parallel transitions. The other kind of error event occurs when the decoder starts to make a sequence of wrong decisions along some path diverging for more than one transition from the correct path. The probability that at any given time such error events take place is called the error-event probability. At high signal-to-noise ratios (SNR), this probability is well approximated by:

$$P_r(e) \cong N_{free} \cdot Q[d_{free}/2\sigma] \quad (8)$$

where $Q[\bullet]$ denotes the Gaussian error integral:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_x^\infty \exp(-t^2/2) dt \quad (9)$$

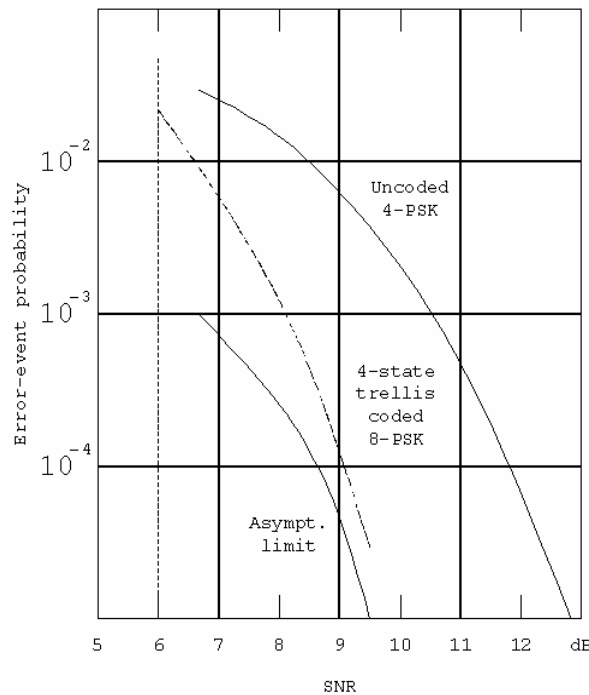
The factor N_{free} denotes the average number of nearest neighbor signal sequences with distance d_{free} that diverge at any state from a transmitted signal sequence, and remerge with it after one or more transitions. The equation (8) also shows the fact that at high SNR, the probability of error events associated with the free distance d_{free} starts to dominate the error performance, whereas any other

probability associated with distances larger than d_{free} becomes negligible.

	d_{free}	N_{free}	Dominant Error Type
Uncoded 4-PSK	$\sqrt{2}$	2	Single signal-decision error
Coded 8-PSK	2	1	Single signal-decision error

【Table 2-1. Comparison of uncoded 4-PSK to coded 8-PSK (at high SNR)】

In *Table 2-1* we compare uncoded 4-PSK and coded 8-PSK. Notice that the free distances of both systems are found between parallel transitions. Therefore the dominant error types of both systems are single-decision errors.



【Figure 2-4. Error-event probability versus signal-to-noise ratio for uncoded 4-PSK and four-state 8-PSK】

Figure 2-4 shows the error-event probability versus signal-to-noise ratio for

uncoded 4-PSK and four-state 8-PSK. Equations (8) and (9) provides an extremely well approximation of error-event probability for uncoded 4-PSK system. For four-state coded 8-PSK, the equations only provide a lower bound that can be achieved asymptotically at high signal-to-noise ratios.

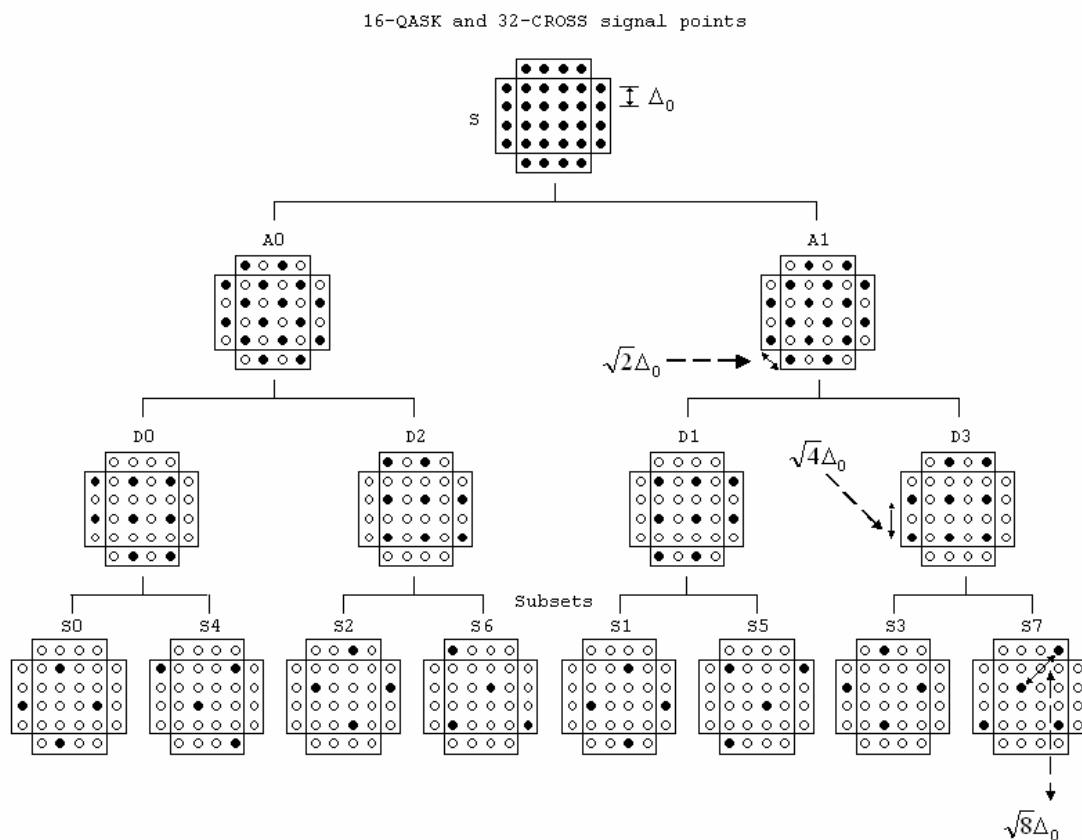
2.2 Eight-State Trellis Code for Amplitude/Phase Modulation [2]

For amplitude/phase modulation, all of the signal sets listed in *Figure 1-1* can be used as long as the codes are designed for two-dimensional signal sets. For uncoded systems, transmission of m information bits per modulation interval requires a 2^m -point signal set. For amplitude/phase modulation systems, the signal sets are expanded by the factor of two, that is, 2^{m+1} -point signal sets are required. For example, for $m=3$, the 16-QASK signal set is used. For $m=4$, the 32-CROSS signal set is used. It will be proven later that for any m , a coding gain of about 4 dB can be achieved over uncoded modulation systems.

2.2.1 The Effect of Set Partitioning

Figure 2-5 illustrates two different signal sets : 16-QASK and 32-CROSS, and their corresponding subsets which are derived from “set partitioning”. The fundamental concept and importance of set partitioning will be introduced in the next section. In this section, we concentrate only on the effect of set partitioning. We denote the original 16-QASK or 32-CROSS signal set as S and the minimum distance between any two signal points in S as Δ_0 . The first division on the set S

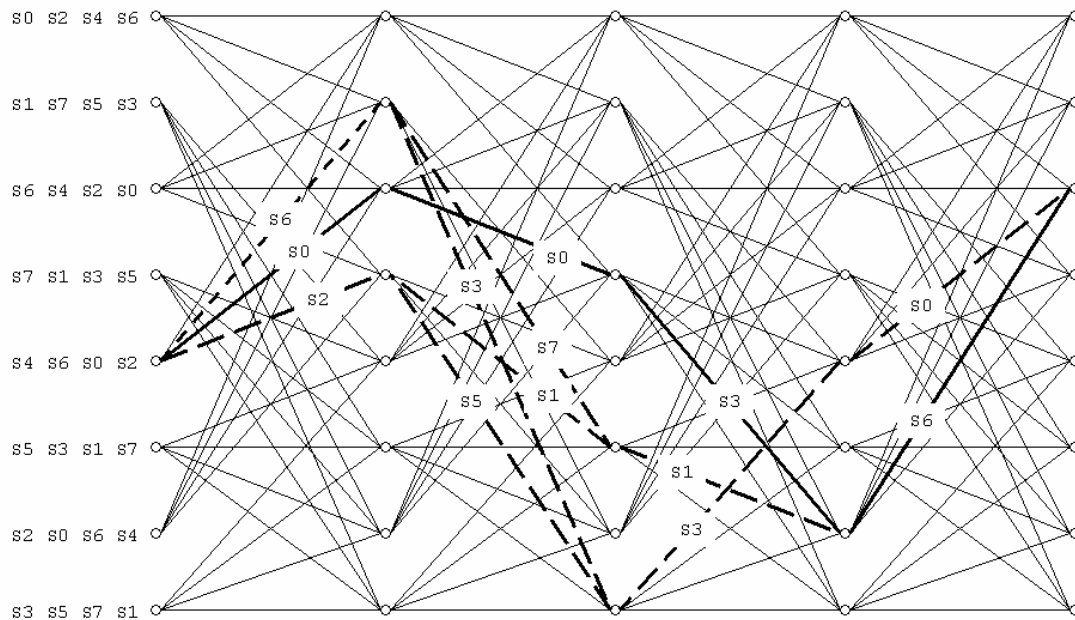
leads to two smaller subsets, namely A_0 and A_1 . In either A_0 or A_1 , the minimum distance between any two signal points becomes $\sqrt{2}\Delta_0$. We go on and apply a second division on A_0 and A_1 to obtain four even smaller subsets D_0, D_2, D_1 , and D_3 . The minimum distance between any two signal points in either of these four subsets becomes $\sqrt{4}\Delta_0$. The last division on the four subsets leads to the last eight subsets S_0, S_1, \dots, S_7 . And the minimum distance between any two signal points in either of these eight subsets becomes $\sqrt{8}\Delta_0$.



【 Figure 2-5. Set partitioning of 16-QASK and 32-CROSS signal sets. 】

Figure 2-6 depicts the trellis diagram of eight-state amplitude/phase modulation. Each of the trellis state has four incoming transitions and four outgoing transitions. In the same sense as *Figure 2-1(b)*, one of the subsets S_0, S_1, \dots, S_7 is assigned to

every one of the transitions. For trellis-coded systems to transmit m information bits per signaling duration, the original signal set will comprise 2^{m+1} signal points. And because of the three division operations applied to the original signal set, each of the subsets S_0, S_1, \dots, S_7 will comprise 2^{m-2} signals. This implies that each of the transitions in Figure 2-6 actually represents 2^{m-2} parallel transitions in the same sense as there were two parallel transitions in the coded 8-PSK system in Figure 2-1(b).



【Figure 2-6. The trellis diagram of eight-state amplitude/phase modulation.】

The assignment of the signal subsets to transitions follows the same three rules as mentioned in *section 2.1.2*. The four transitions diverging from or remerging into the same state are always assigned the subsets S_0, S_4, S_2, S_6 or S_1, S_5, S_3, S_7 , that is, either the subsets in A_0 or those in A_1 . This guarantees that two sequences diverging from or remerging into the same state would have a squared distance of at least $(\sqrt{2}\Delta_0)^2 = 2\Delta_0^2$. If two sequences diverge from a state and then remerge to

another after two transitions, the squared distance between the first transitions of the two paths would be at least $4\Delta_0^2$, and hence the total squared distance between such paths would amount to $6\Delta_0^2$. If two sequences diverge from a state and then remerge to another after three or more than three transitions, the first (diverging) and the last (remerging) transitions would respectively contribute to a squared distance of $2\Delta_0^2$, and at least one intermediate transition would contribute to another squared distance of Δ_0^2 . Therefore the total squared distance between such transitions would be at least $5\Delta_0^2$. Because of the fact that the squared distance between parallel transitions is $(\sqrt{8}\Delta_0)^2 = 8\Delta_0^2$, which is larger than $5\Delta_0^2$, we conclude that the free distance of this code is $\sqrt{5}\Delta_0$, and that any error sequences corresponding to such distance from the correct sequence would be the most probable error events. One last thing to notice in this subsection is that, at high signal-to-noise ratios, the dominant error events of this code would occur in the form of error bursts of length three or four.

2.2.2 Examples of Error Bursts

We particularly marked several paths in *Figure 2-6* as examples of sequences at the free distance (which are also the most probable error events). Let the correct sequence in *Figure 2-6* be $S0 - S0 - S3 - S6$ and marked with a thick solid line. Four different error sequences at the squared free distance $5\Delta_0^2$ from the correct sequence are also marked with thick dashed lines. They are : $S6 - S7 - S1 - S6$, $S6 - S3 - S3 - S0$, $S2 - S1 - S1 - S6$, and $S2 - S5 - S3 - S0$. We can see from *Table 2-2* how these error sequences add up to the squared free distance $5\Delta_0^2$.

	1 st transition	2 nd transition	3 rd transition	4 th transition	Sum
Correct path	S0	S0	S3	S6	
Error path 1	$S6/2\Delta_0^2$	$S7/\Delta_0^2$	$S1/2\Delta_0^2$	S6/ 0	$5\Delta_0^2$
Error path 2	$S6/2\Delta_0^2$	$S3/\Delta_0^2$	S3/ 0	$S0/2\Delta_0^2$	$5\Delta_0^2$
Error path 3	$S2/2\Delta_0^2$	$S1/\Delta_0^2$	$S1/2\Delta_0^2$	S6/ 0	$5\Delta_0^2$
Error path 4	$S2/2\Delta_0^2$	$S5/\Delta_0^2$	S3/ 0	$S0/2\Delta_0^2$	$5\Delta_0^2$

【Table 2-2. Four error paths at the free distance $\sqrt{5}\Delta_0$ from $S0 - S0 - S3 - S6$.】

To interpret **Table 2-2**, we notice that the second row of the table represents the correct sequence path. Take the third row as an example : the notation “ $S6/2\Delta_0^2$ ” at the second column denotes that the squared distance between the first transitions of both the correct path and the error path 1 is equal to $2\Delta_0^2$, that is :

$$\text{Min}_{a_i \in S0, b_i \in S6} \left\{ \text{dist}(a_i, b_i)^2 \right\} = 2\Delta_0^2 \quad (10)$$

In **Formula (10)** a_i and b_i are the signal points belonging respectively to the subset $S0$ and $S6$. The notation $\text{dist}(a_i, b_i)$ denotes the Euclidean distance from a_i to b_i . The rest components of **Table 2-2** can be filled in following the same searching procedure. As can be seen from **Table 2-2**, among all of the four most probable error events, error path 1 and 3 are of length three, and error path 2 and 4 are of length four. Therefore, error path 1 and 3 will result in bursts of decision errors of length three, and error path 2 and 4 will result in bursts of decision errors of length four.

N_{free} , which is the average number of nearest neighbors, depends on the

sequences of signals transmitted. For uncoded modulation, the signal points located in the inner layers have more neighbors than those located at outer layers. For 16-QASK, N_{free} is equal to 3. However, for eight-state coded 16-QASK, N_{free} is about 3.75. As the size of the signal set grows, the N_{free} values of uncoded and eight-state coded systems would increase towards 4 and 16, respectively.

We calculate the asymptotic coding gains in decibels at high signal-to-noise ratios using the formula :

$$G = 10 \cdot \log_{10} \left[\frac{(d_{free,C}^2 / d_{free,U}^2) / (P_{S,C} / P_{S,U})}{1} \right] \quad (11)$$

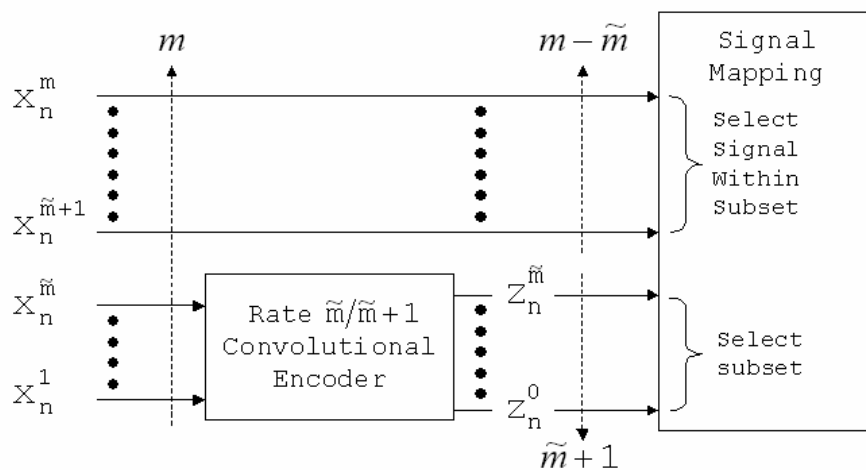
where $d_{free,C}^2$ and $d_{free,U}^2$ denote the squared free distances, and $P_{S,C}$ and $P_{S,U}$ denote the average signal powers of the coded and uncoded schemes, respectively. Assume that both the uncoded signal set and the coded signal set have the same minimum signal spacing Δ_0^2 , and that $d_{free,C}^2 / d_{free,U}^2 = 5$, $P_{S,C} / P_{S,U} \cong 2$ for all relevant values of m . The coding gain is therefore $10 \cdot \log_{10}(5/2) \cong 4 \text{ dB}$. This result is as expected as mentioned earlier in *section 2.2*.

2.3 Design of Trellis-Coded Modulation Schemes [3]

In *section 2.1* and *2.2*, two examples of TCM schemes designed in a heuristic manner during the early phases of the development of TCM schemes were presented.

These schemes were invented as methods to improve error performance of digital transmission systems without reducing the equivalent data rate or expanding the bandwidth. The idea of signal set expansion allows the achievement of the enlargement of free Euclidean distances and hence coding gains ranging from 3 dB to 6 dB at spectral efficiencies equal to $2(\text{bit/sec})/\text{Hz}$. In this section, we will examine first the general structure of TCM schemes, and then the principles of code construction.

2.3.1 General Structure of Encoder/Modulator for TCM

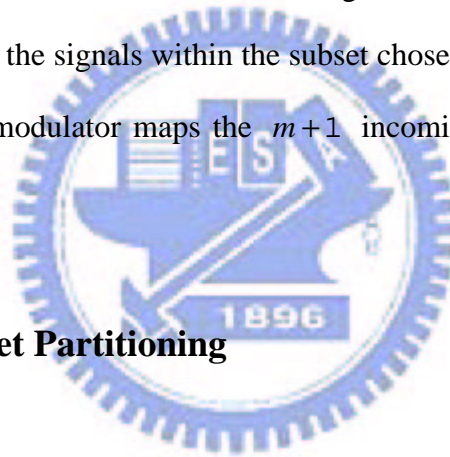


【Figure 2-7. General structure of encoder/modulator for TCM】

As *subsection 1.1.4* has mentioned, most of the achievable coding gain could be obtained by expanding the signal sets used for uncoded modulation only by a factor of two. Or equivalently, in order to improve error performance, $m \text{ bit}/T$ must be transmitted in redundantly coded form with a $2^{\tilde{m}+1}$ -point signal set. The expansion of the binary data sequence can be easily accomplished by using a suitable rate- $\tilde{m}/\tilde{m}+1$ convolutional encoder that transforms m uncoded bits into $\tilde{m}+1$ coded bits,

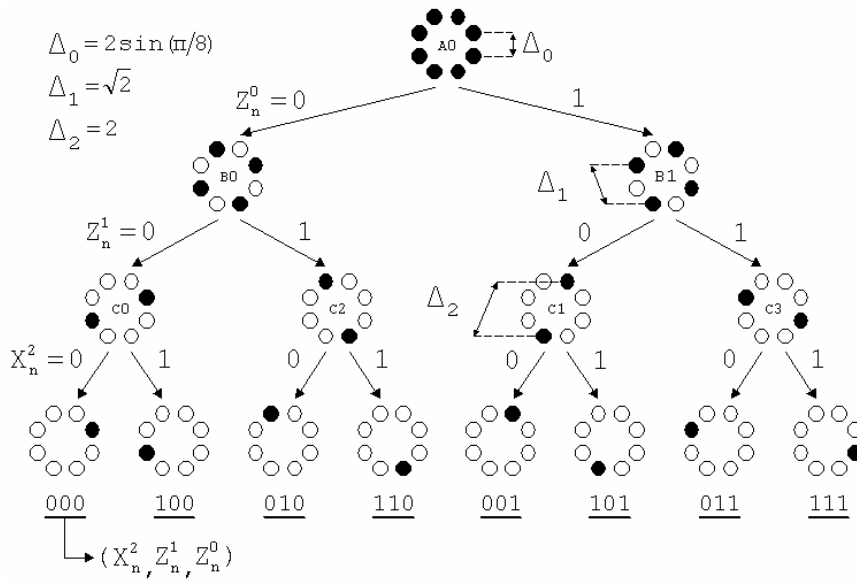
succeeded by a modulator which maps groups of $m+1$ bits into the channel signals in the expanded signal sets. The mapping should group signals into “subsets” with large distance between the subset signals. The general structure of the TCM encoder / modulator is therefore as depicted in **Figure 2-7**.

According to **Figure 2-7**, the TCM signals are generated as follows : $\tilde{m} < m$ out of the m transmitted bits are input to a rate- $\tilde{m}/\tilde{m}+1$ binary convolutional encoder per encoder/modulator operation. The other $m - \tilde{m}$ bits are left uncoded and passed directly to the input of the signal modulator. The $\tilde{m}+1$ coded bits are used to select one of the $2^{\tilde{m}+1}$ subsets of the redundant signal set. The $m - \tilde{m}$ uncoded bits are used to select one of the signals within the subset chosen by the $\tilde{m}+1$ coded bits. And finally the signal modulator maps the $m+1$ incoming bits onto the expanded 2^{m+1} -ary signal set.



2.3.2 Mapping by Set Partitioning

In **Figure 2-7**, the signal modulator follows a certain mapping rule which aims directly at maximizing the free Euclidean distance of the code, called “mapping by set partitioning”. Assume that the minimum spacing between any two signals in the signal set spanned by $X_n^m, \dots, X_n^{\tilde{m}+1}, Z_n^{\tilde{m}}, \dots, Z_n^0$ in **Figure 2-7** is Δ_0 . “Mapping by set partitioning” follows a successive partitioning procedure to divide the signal set into smaller subsets with maximally increasing minimum distances $\Delta_0 < \Delta_1 < \Delta_2 \dots$ between the signals of these subsets, where Δ_i stands for the minimum spacing between any two signals in the subsets after the i_{th} division. This concept has been shown in **Figure 2-5** and its effect has been seen. For better understanding of the concept, we now present another simpler case in **Figure 2-8**.

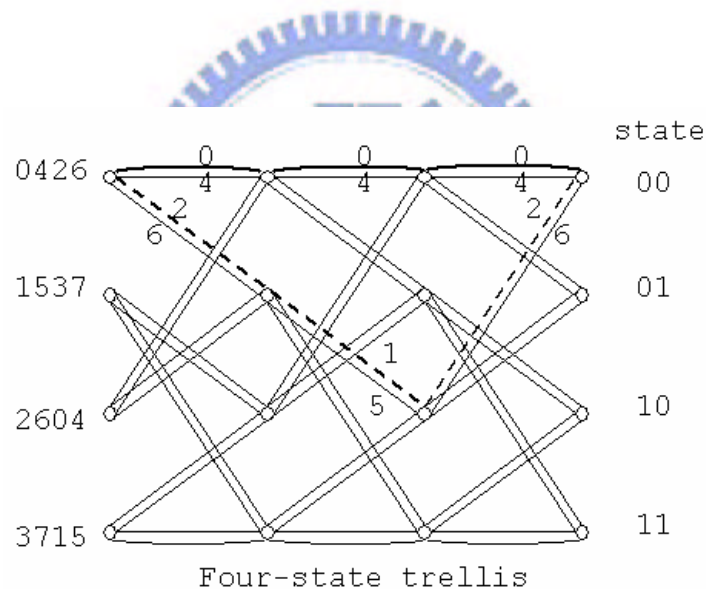


【Figure 2-8. Set partitioning of 8-PSK signals with increasing Δ_i 】

Consider an 8-PSK signal set, namely, A_0 . The minimum distance between neighboring signals is $\Delta_0 = 2\sin\pi/8$. The first partitioning divides A_0 into two smaller subsets B_0 and B_1 , each of which has four signal points and the minimum distance between such points is $\Delta_1 = \sqrt{2}$. Applying the second partitioning to B_0 and B_1 yields four subsets, namely, C_0 , C_2 , C_1 , and C_3 with even bigger minimum distances between signals $\Delta_2 = 2$. The partitioning will be carried on until the minimum distance between signals within the same subset after the i^{th} partitioning, Δ_i , is larger than the free distance of the reference system. In this case, the reference system of the coded 8-PSK system is the uncoded 4-PSK system, which has a free distance of $\sqrt{2}$. Therefore, the process stops at the second partitioning and C_0 , C_2 , C_1 , and C_3 are the final subsets we have. In both systems, unity signal power is assumed.

We can once again take a look at *section 2.1* and *Figure 2-1(b)* from the

perspective of set partitioning. In **Figure 2-9**, each transition from one state to another actually represents two parallel transitions and signals within the same subsets are assigned to these parallel transitions, for example, (0,4) from subset C_0 , (2,6) from subset C_2 . This matches with the third assignment rule in **subsection 2.1.2**. Notice that in **Figure 2-9** every state has four incoming and outgoing transitions. Signals from the same subsets derived from the first partitioning are assigned to these four incoming or outgoing transitions, for example, (0,2,4,6) from subset B_0 or (1, 3,5,7) from subset B_1 . This matches with the second assignment rule in **subsection 2.1.2**. Now it explains the three rules for assignment of signal points to trellis transitions : to enlarge the free distance of the code.



【Figure 2-9. A four-state trellis diagram】

The labeling of branches in the partition tree in **Figure 2-8** by the $\tilde{m}+1$ coded bits $Z_n^{\tilde{m}}, \dots, Z_n^0$ results in a label $\bar{Z}_n = [Z_n^{\tilde{m}}, \dots, Z_n^0]$ for each subset. The label indicates the specific location of the subset in the partition tree. This explains why these $\tilde{m}+1$ coded bits are marked “select subset” in **Figure 2-7**. If the partitioning process continues to the point that each subset has only one signal point, each combination of

the $m+1$ bits $X_n^m, \dots, X_n^{\tilde{m}+1}, Z_n^{\tilde{m}}, \dots, Z_n^0$, represents one particular signal in the set. This explains why the $m - \tilde{m}$ uncoded bits $X_n^m, \dots, X_n^{\tilde{m}+1}$ are marked “select signal within subset” in **Figure 2-7**, and why they are associated with the $2^{m-\tilde{m}}$ parallel transitions. The labeling leads to an important property. If the labels of two subsets agree in the last l bits Z_n^{l-1}, \dots, Z_n^0 but not in the bit Z_n^l , then the signals of these two subsets belong to the same subset at level l in the partition tree, therefore, they have distance Δ_l between them.

2.3.3 Convolutional Codes for Trellis-Coded Modulation

At every time n , the rate- $\tilde{m}/\tilde{m}+1$ convolutional encoder shown in **Figure 2-7** receives \tilde{m} incoming bits and generates $\tilde{m}+1$ outgoing bits $Z_n^{\tilde{m}}, \dots, Z_n^0$. The set of all possible sequences that can be generated by the convolutional encoder is called a convolutional code. A linear rate- $\tilde{m}/\tilde{m}+1$ convolutional code can be most compactly defined by a parity check equation which puts a constraint on the code bits in a sliding time window of length $\mathbf{u}+1$:

$$\sum_{i=0}^{\tilde{m}} (h_{\mathbf{u}}^i z_{n-\mathbf{u}}^i \oplus h_{\mathbf{u}-1}^i z_{n-\mathbf{u}+1}^i \oplus \dots \oplus h_0^i z_n^i) = 0 \quad (12)$$

In **equation (12)**, \oplus denotes modulo-2 addition and \mathbf{u} denotes the constraint length of the convolutional code, that is, there are \mathbf{u} binary storage elements in the convolutional encoder. It is equivalent to say that the code has $2^{\mathbf{u}}$ trellis states. The quantities $h_k^i, 0 \leq k \leq \mathbf{u}, 0 \leq i \leq \tilde{m}$ are the binary parity check coefficients of the convolutional code. All possible code sequences must satisfy **equation (12)** at all times.

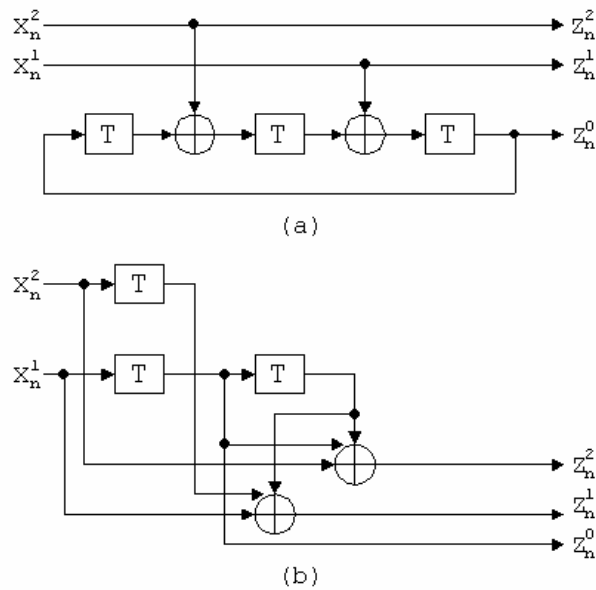
We assume now two sequences \bar{z}_n and $\bar{z}'_n = \bar{z}_n \oplus e_n$, where \bar{z}'_n is an error version of \bar{z}_n and e_n denotes the error sequence by which \bar{z}_n and \bar{z}'_n differ. If the convolutional code is linear, the error sequence e_n must also be a code sequence. Define the notation $q(e_n)$ as the number of trailing zeros in e_n , that is, the number of trailing positions in which \bar{z}_n and \bar{z}'_n agree. For example, if $e_n = [\dots, 1, 0, 0]$, $q(e_n) = 2$. From what has been mentioned about set partitioning in the last section, the distance between signals in the subsets labeled by \bar{z}_n and \bar{z}'_n is lower-bounded by $\Delta_{q(e_n)}$. It follows that the squared free Euclidean distance between non-parallel paths in the TCM trellis is lower-bounded by :

$$d_{free}^2(\tilde{m}) \geq \text{Min}_{e_n \neq \{0\}} \sum_n \Delta_{q(e_n)}^2 \quad (13)$$

The minimization operation $\text{Min}(\bullet)$ has to be carried out for all non-zero error sequences e_n . The linear property of the rate- $\tilde{m}/\tilde{m}+1$ convolutional code assures that for any given error sequence e_n and any code sequence \bar{z}_n , there exists another code sequence \bar{z}'_n that differs from \bar{z}_n by exactly the error sequence e_n . This equation for the lower bound of the squared free distance is of key importance in the search for optimum TCM codes because it states that the free Euclidean distance can be determined in almost the same way as free Hamming distance is found in linear block codes.

Figure 2-10 (a) and **(b)** illustrate two convolutional encoder realizations. For both encoders, the constraint length is $u = 3$ and hence the trellis of the code has $2^u = 8$ states. Both encoders are “minimal”, that is, they both are realized with $u = 3$ storage elements. **Figure 2-10 (a)** is a systematic encoder with feedback. The

output of the last binary storage element is fed back as the input of the first storage element. The first input bit is left uncoded as the first output bit. Therefore, the codes it generates cannot be catastrophic. **Figure 2-10 (b)** shows a feedback-free encoder. It can be shown in [7] that feedback-free encoder cannot generate catastrophic codes either, if it is minimal.



【Figure 2-10. Two realizations of rate-2/3, $u=3$ convolutional encoders. (a)

Systematic encoder with feedback (b) Feedback-free encoder】

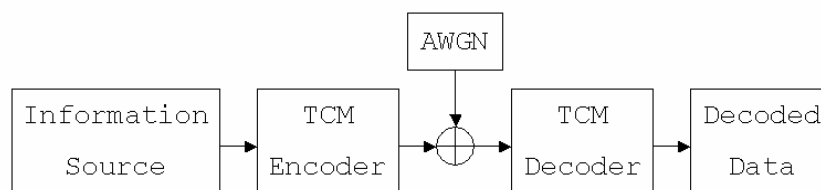
Chapter 3

TCM Schemes Using 3-D Constellations

In this chapter, the idea of TCM schemes using three-dimensional constellations will be introduced. Examples of TCM schemes with different convolutional encoders and modulators will be given and their performance will be compared with their two-dimensional reference systems.

3.1 Four-State Trellis Code for 8-Point Cubic Modulation

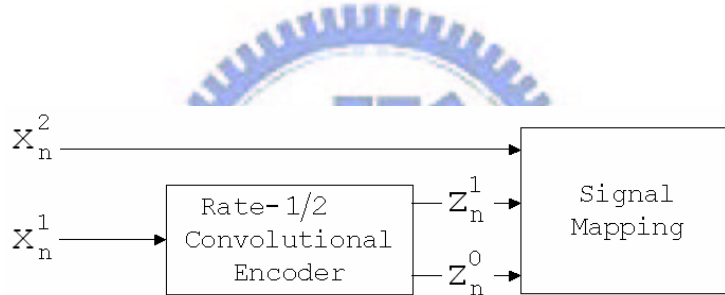
3.1.1 Basic Structure of The Proposed Schemes



【Figure 3-1. Simplified System Diagram (For Computer Simulation).】

Figure 3-1 illustrates the simplified system diagram of the proposed system. For the convenience of computer simulation, we have made an assumption that the only additive white Gaussian noise is of our concern. For simplicity, other assumptions such as inter-symbol interference-free signaling, perfect timing, and carrier-phase

synchronization are also made. With these assumptions that have been made in the discussion of two-dimensional schemes, we will see at the end of this chapter that the proposed three-dimensional schemes provide certain performance gains over two-dimensional schemes. The TCM encoder depicted in **Figure 3-1** is composed of two parts, the first part is a convolutional encoder and the second is a signal modulator. A more specific structure is illustrated in **Figure 3-2**. The TCM decoder is also composed two parts. The first one is a subset decoder, which has been described in **subsection 2.1.3**, and the second one is a soft-decision Viterbi decoder. We will put our emphasis on the structure of the encoder of the proposed schemes ; the structure of the decoder is almost the same as that of two-dimensional schemes.

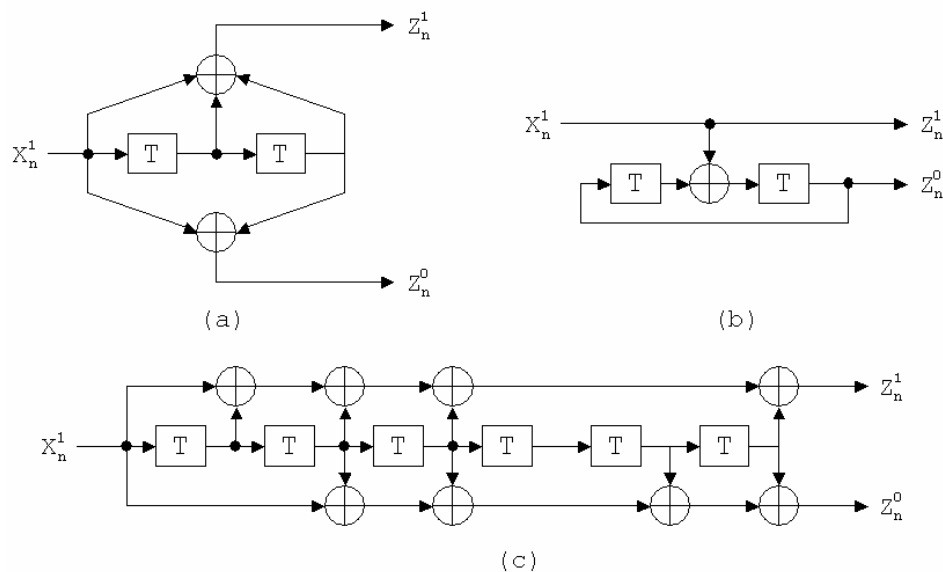


【Figure 3-2. Basic structure of four state trellis code for 8-Point Cubic Modulation】

Figure 3-2 shows the block diagram of the basic encoder structure of the proposed TCM schemes. Speaking as a whole, the TCM encoder has two input bits and three output bits. One of the two input bits, say, X_n^2 , is left uncoded and passed all the way to the signal modulator. The other input bit X_n^1 is input to a rate-1/2 convolutional encoder and two output bits Z_n^1 and Z_n^0 are therefore generated and passed to the signal modulator, where Z_n^1 , Z_n^0 and the uncoded input bit X_n^2 as well are mapped into channel signals. The design of the three-dimensional signal modulator is the essence of this thesis. As for the convolutional encoders, examples of several regularly seen encoders will be presented in the next section.

3.1.2 Rate-1/2 Convolutional Encodes

In this section, we simply demonstrate several rate-1/2 convolutional encoders that we use in our computer simulations as part of the TCM encoders in our schemes. The variety of the encoders is shown in *Figure 3-3*.

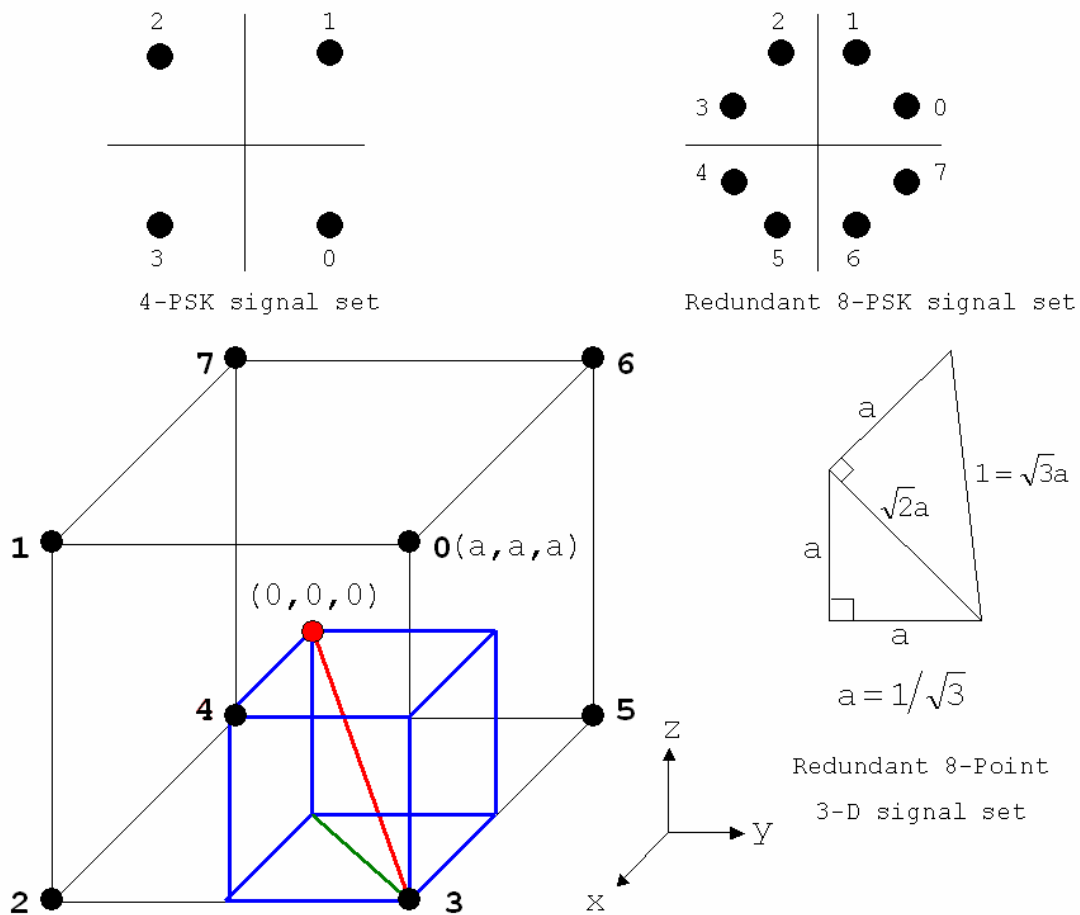


【 Figure 3-3. Three realizations of rate-1/2 convolutional encoders. (a) Feedback-free encoder with $u = 2$. (b) Systematic encoder with feedback and $u = 2$. (c) Feedback-free encoder with $u = 6$. 】

Figure 3-3 (a) is a feedback-free encoder with constraint length $u = 2$. *(b)* has the same constraint length but different in that it has a feedback structure. *(c)* is a generalized form of *(a)* but it has six binary storage elements, that is, $u = 6$. In the literature on convolutional coding, we find several different definitions of constraint length. Some define the constraint length to be the exact number of the memory elements in the encoder, while others define it to be the number of the memory elements in the encoder plus one. We apply the former definition to this thesis.

In *Figure 3-3 (a)*, the current output pair is a function of the current input and the two previous inputs. According to the property of convolutional codes, bigger constraint lengths lead to better error correction capability [9]. We will prove it true by computer simulation in the next chapter. All the units “ \oplus ” in the *Figure 3-3* indicate the modulo-two addition operation.

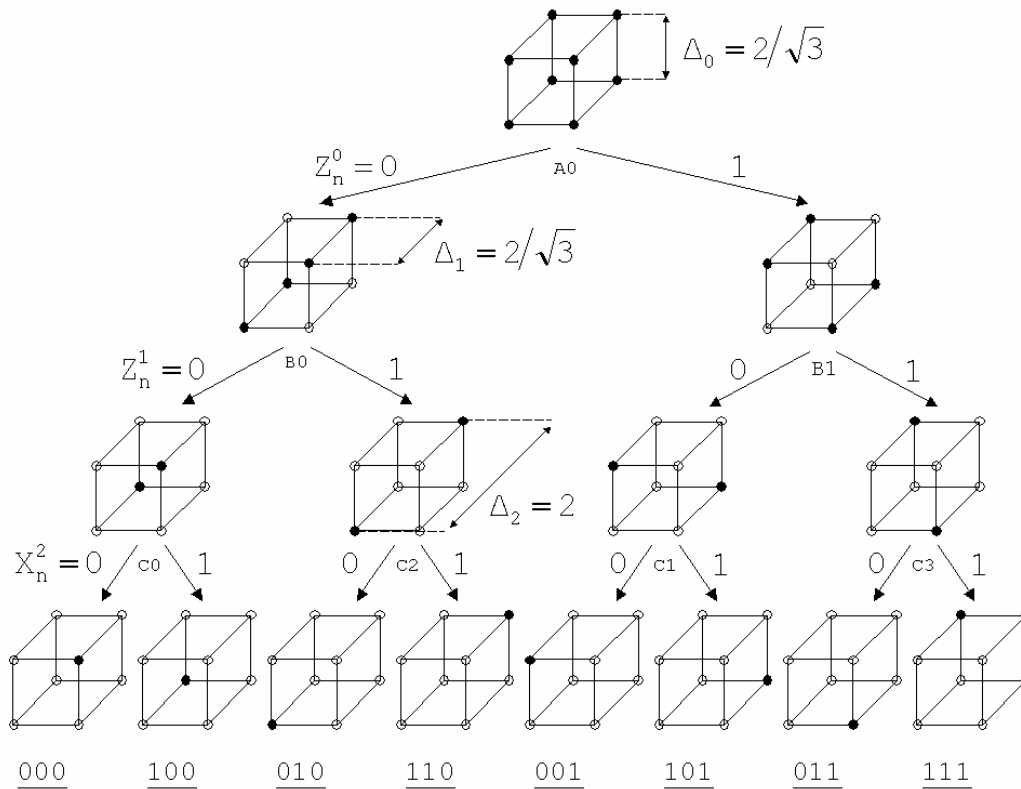
3.1.3 The Three-Dimensional Signal Set



【Figure 3-4. 8-point cubic signal set.】

We have introduced the four-state trellis code using the 8-PSK signal set in *section 2.1*. Recall that the coding gains achieved in two-dimensional TCM schemes

are the results of expanding the signal sets and the “mapping by set partitioning” lemma. The proposed scheme uses the three-dimensional signal constellation shown in **Figure 3-4**. The uncoded 4-PSK signal set and the expanded 8-PSK signal set are also given as reference on the top of **Figure 3-4**. We assume that the origin of the three-dimensional Euclidean space $(0,0,0)$ locates in the center of the cube, and that the mean power of the signal points is set to unity, therefore the coordinate of the signal “0” is $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, and the distance between neighboring signals is $\Delta_0 = 2/\sqrt{3}$.



【Figure 3-5. Set partitioning diagram of 8-point cubic signal set】

The set partitioning is done in the same sense as is done for two-dimensional cases : The first partitioning divides the original set A_0 into two subsets B_0 and B_1 , each of which contains four signals with the minimum distance $\Delta_1 = 2/\sqrt{3}$

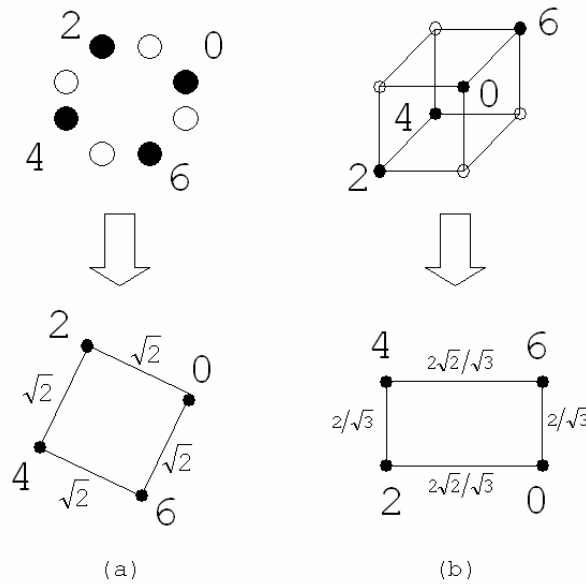
between signals within the subset. Applying the second level of partitioning to B_0 and B_1 yields four smaller subsets C_0 , C_2 , C_1 and C_3 with the minimum distance between signals $\Delta_2 = 2$. Because only one of the two input bits is input to the rate-1/2 convolutional encoder, the set partitioning operation stops after these four subsets are obtained. The coded bits output from the convolutional encoder are used to select one of the four subsets, and the uncoded bit is used to select one of the two signals within the selected subset. The assignment of the signals to the transitions in the trellis diagram “approximately”(we will explain later in the next section) follows the same three rules described in *subsection 2.1.2*. The trellis diagrams of the three-dimensional TCM schemes are the same as that of two-dimensional ones because same convolutional encoders are used.

Subsection 1.2.2 states that it is possible to achieve some gains in error performance simply by expanding the dimension of the signal set, for example, replacing the one-dimensional 8-AM signal set with the two-dimensional 8-PSK signal set will achieve a gain of approximately 10 dB. Unfortunately, exact mathematical proof seems to be impractical because of the extremely complex derivations when it comes to expanding two-dimensional signal sets to three-dimensional ones. We will see the performance gain and explain it in a heuristic manner in *Chapter 4* with the help of the results of computer simulations.

3.1.4 The Modified Three-Dimensional Signal Set

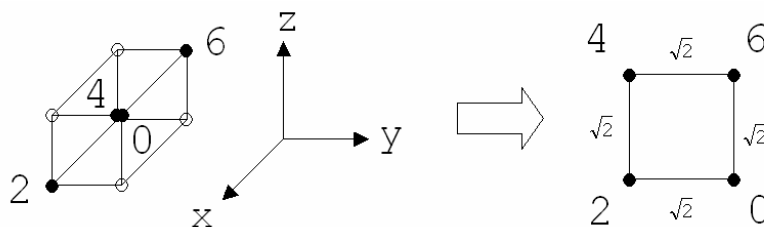
In the last section we mentioned the three rules for the assignment of signals to transitions in the trellis diagram. Particularly, *rule#2* states similarly that : Transitions originating from or remerging into the same state are labeled with signals from B_0

or $B1$. This rule in fact implies that the distances between neighboring signals must agree for all four cases, that is, $\overline{02} = \overline{24} = \overline{46} = \overline{62}$. **Figure 3-6 (a)** shows the case of the signal set of the coded 8-PSK. However, we can see that this property does not hold for the signal set of the coded 8-point cubic modulator in **Figure 3-6 (b)**.



【Figure 3-6. Subset $B0$ for (a) 8-PSK (b) 8-point cubic, signal sets】

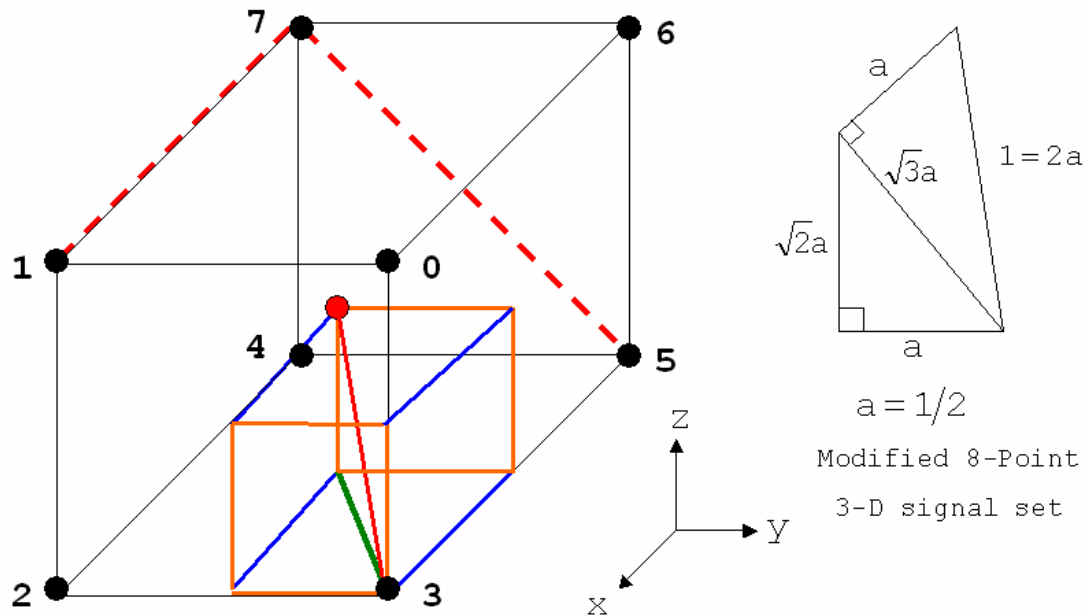
To remedy this slight defect, we present another modified 8-point cubic signal set which follows “exactly” the second assignment rule.



【Figure 3-7. Subset $B0$ for modified 8-point cubic signal set】

In **Figure 3-7**, the original cubic signal set is stretched in the direction of the X axis and both the surfaces parallel to the YZ plane are shrunk in order to maintain

the unity average signal power. For a clearer look at the measures of the modified signal set, **Figure 3-8** is presented.

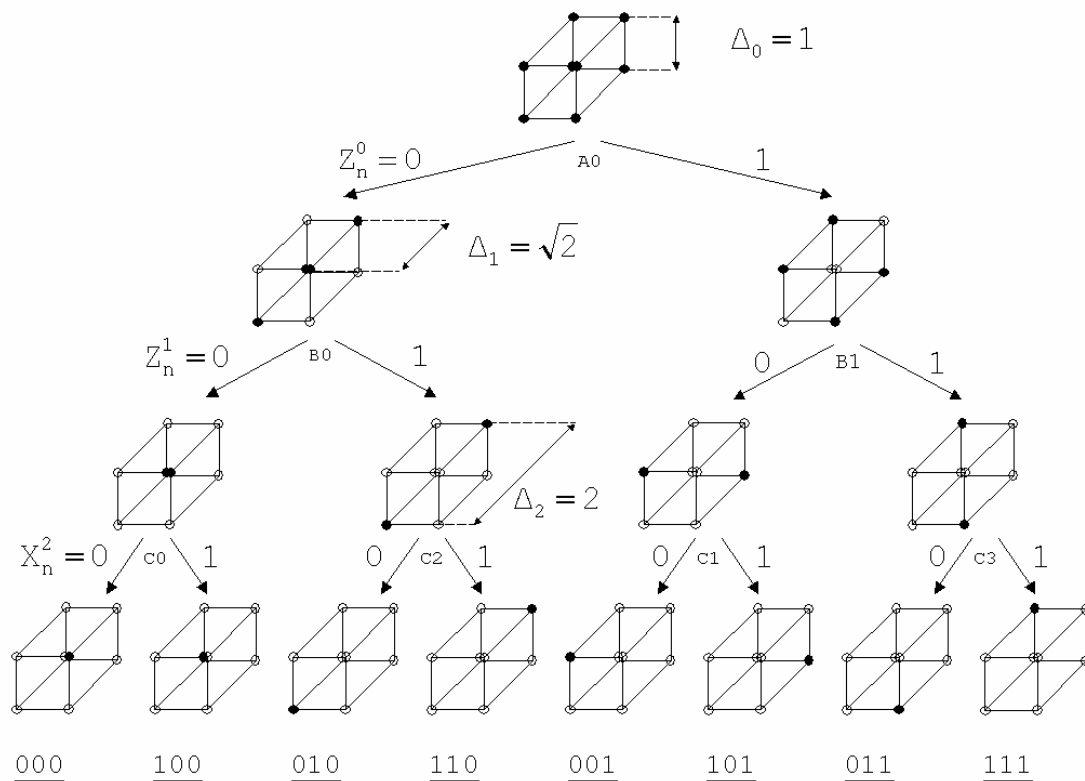


【Figure 3-8. Modified 8-point cubic signal set.】

Assume that the origin of the three-dimensional Euclidean space $(0,0,0)$ locates in the center of the modified cube (which we will call a “cuboid” from now on), and that the mean power of the signal points is set to unity. Therefore the coordinate of the signal “0” is $(\sqrt{2}/2, 1/2, 1/2)$, and the minimum distance between neighboring signals is $\Delta_0 = 1$, for example, $\overline{01}$ or $\overline{12}$.

The set partitioning of the cuboid signal set is very much the same as that of the cubic signal set except for the minimum distances between signals within subsets in all partitioning levels, that is, Δ_0 , Δ_1 and Δ_2 , are different. In the next section, a comparison will be made between the three-dimensional cubic signal set and the three-dimensional cuboid signal set. The error performance of the TCM schemes

using the cuboid signal set will be proven mathematically to be better than that using the cubic signal set. Before that, the set partitioning diagram of 8-point cuboid signal set is provided.



【Figure 3-9. Set partitioning diagram of modified 8-point cuboid signal set】

3.1.5 Free Distances of Cubic and Cuboid Signal Sets

We have followed a certain searching procedure to determine the free distance of the TCM scheme using 8-PSK signal set. For TCM schemes using the above two three-dimensional signal sets (cubic and cuboid), the same procedure is followed to determine the free distances of the schemes. We ignore the specific structure of the convolutional encoder in the TCM encoder because there can be a number of realizations. Due to this fact, we directly assume that the TCM encoder of our interest

has a trellis structure as shown in **Figure 2-9**. We will determine the free distance and the corresponding error type of the cubic signal set, and those of the cuboid signal set later.

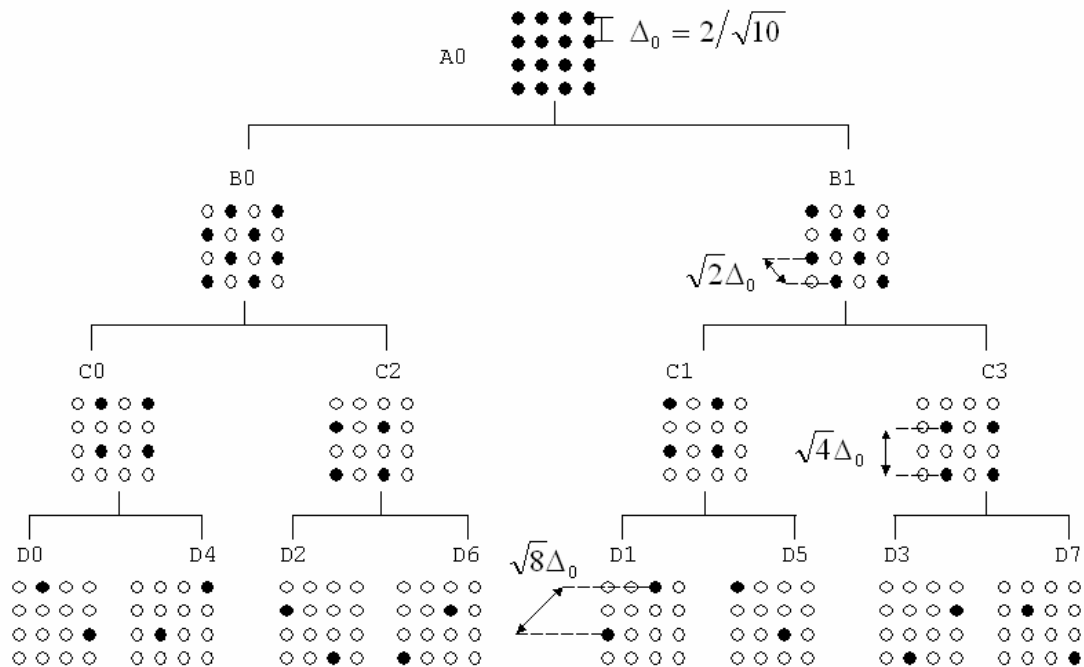
Referring to **Figure 2-9** and **Figure 3-5**, the minimum distance between parallel transitions is equal to the distance between the two signals within a subset, for example, $C2$. Such a distance is twice the average power of the signal set because it passes through the center of the cube and connects two opposite corners. Therefore the minimum square distance between parallel transitions is $2^2 = 4$ for the TCM scheme using the cubic signal set. Consider two sequences diverging from a state and then remerging into another after more than one transition (non-parallel transitions). Such sequences must at least have a squared distance between $0-0-0$ and $6-1-6$, that is, $(2/\sqrt{3})^2 + (2/\sqrt{3})^2 + (2/\sqrt{3})^2 = 4$. Therefore we can say that the free distance of TCM schemes using the cubic signal set is $\sqrt{4} = 2$ because both the squared distances between parallel transitions and non-parallel transitions are identical. This is a very important property because it means single signal-decision errors and error bursts of length three are equally probable. At high signal-to-noise ratios, both error types would dominate the error performance.

The free distance of TCM schemes using the cuboid signal set is determined in the same way. The minimum squared distance associated with parallel transitions is the same as cubic schemes. However, the minimum squared distance associated with the second error type is different : The squared minimum distance between $0-0-0$ and $2-1-2$ or $6-1-6$ is $\sqrt{2}^2 + 1^2 + \sqrt{2}^2 = 5$, which is larger than that in cubic schemes. Although at high signal-to-noise ratios, single signal-decision errors would dominate the error performance and the effects of error bursts would be diminishing,

we can still expect better error performance than cubic schemes at lower signal-to-noise ratios.

3.2 Eight-State Trellis Code for 16-Point Cubic Modulation

3.2.1 Two-Dimensional TCM Schemes

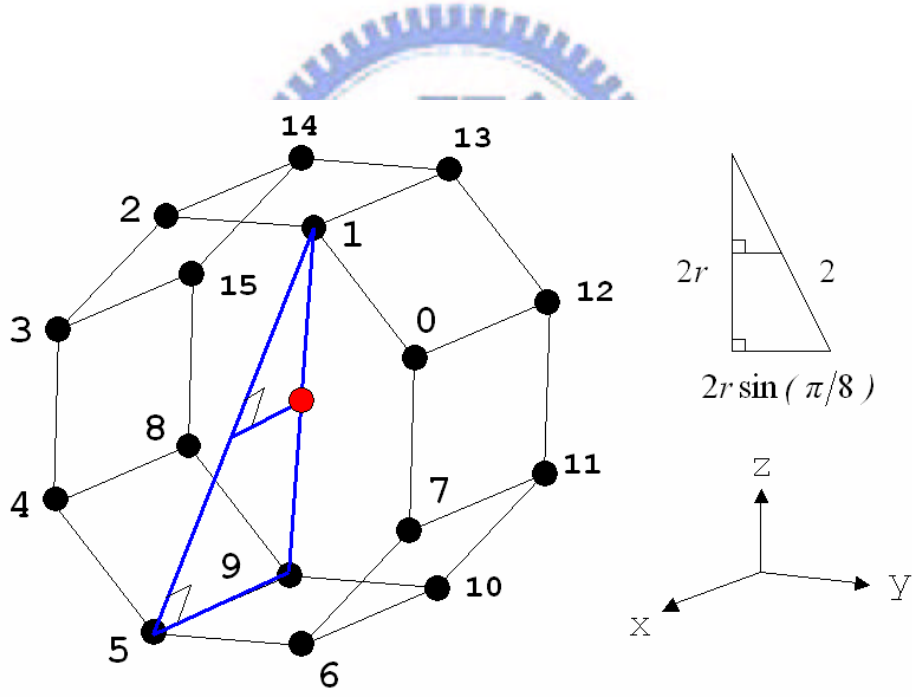


【Figure 3-10. Set partitioning diagram of a redundant 16-QASK signal set】

In this section, we will propose another TCM scheme using a three-dimensional signal set and its modified version. Consider an uncoded system with an 8-PSK signal set as the reference system. According to the discussion in the last few sections, one realization to turn this reference system into a TCM scheme would be like *Figure 3-2*, except that we have to replace the rate-1/2 convolutional encoder with a rate-2/3

one, two examples of which can be found in *Figure 2-10*. The rate of the entire TCM encoder would therefore be $3/4$. The expanded 16-QASK signal set and its set partitioning diagram is as illustrated in *Figure 3-10*. The same partitioning and assignment rules are followed as usual. Based on the assumption of the unity average signal power, the free distance of the uncoded 8-PSK system is $2\sin(\mathbf{p}/8)$, and the free distance of the coded 16-QASK TCM system is $\sqrt{5} \times 2 / \sqrt{10} = \sqrt{2}$, therefore a coding gain of about $20 \times \log_{10}(\sqrt{2} / 2\sin(\mathbf{p}/8)) \cong 5.33dB$ is expected at high signal-to-noise ratios.

3.2.2 The Three-Dimensional Signal Set

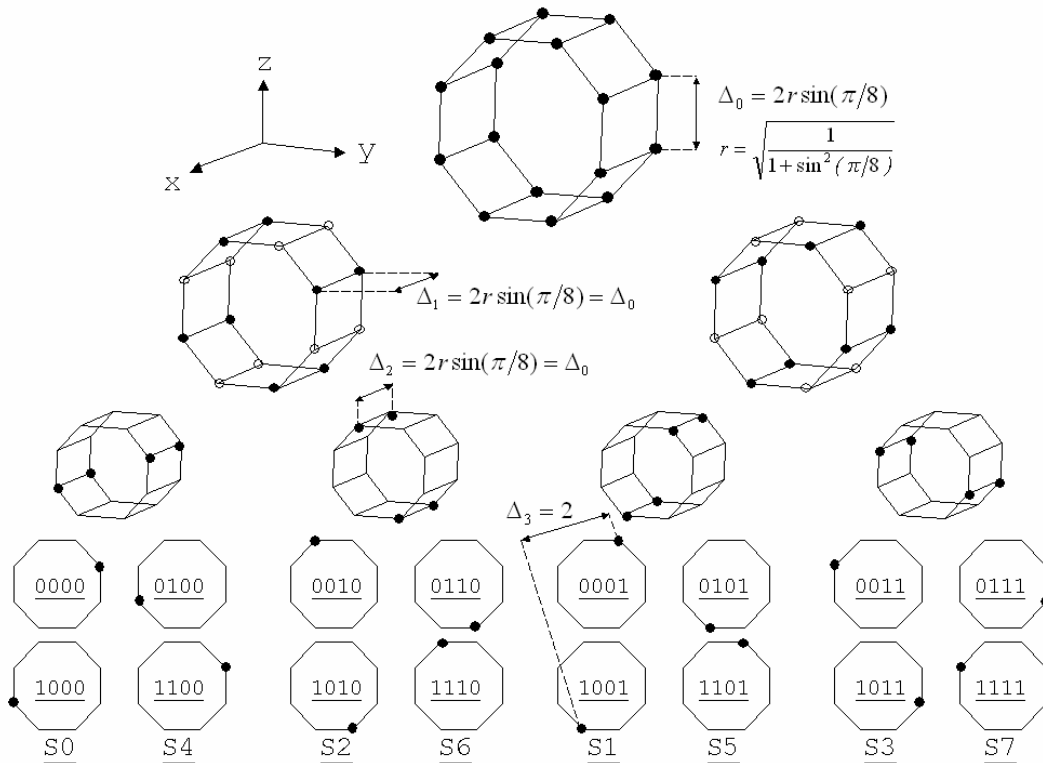


【Figure 3-11. 16-point cubic signal set.】

We wish to design a three-dimensional signal set which is an expansion version of the original uncoded 8-PSK signal set in the sense that it duplicates the uncoded 8-PSK signal set and aligns the duplicate set with the original set in the direction of the

X axis. The result is as depicted in **Figure 3-11**. The goal of our first design is to make every signal point in the set have three equally distant neighbors, for example, the signal “0” has three equally distant neighbors “1”, “7” and “12”. We call such a design a “16-point cubic signal set” in accordance with the terminology used in **subsection 3.1.3**. Using Pythagoras' Theorem we have the following derivation :

$$\begin{aligned}
 4 &= 4r^2 + 4r^2 \sin^2 (p/8) \\
 \Rightarrow r^2(1 + \sin^2 (p/8)) &= 1 \\
 \Rightarrow r &= \sqrt{\frac{1}{1 + \sin^2 (p/8)}} \tag{14}
 \end{aligned}$$

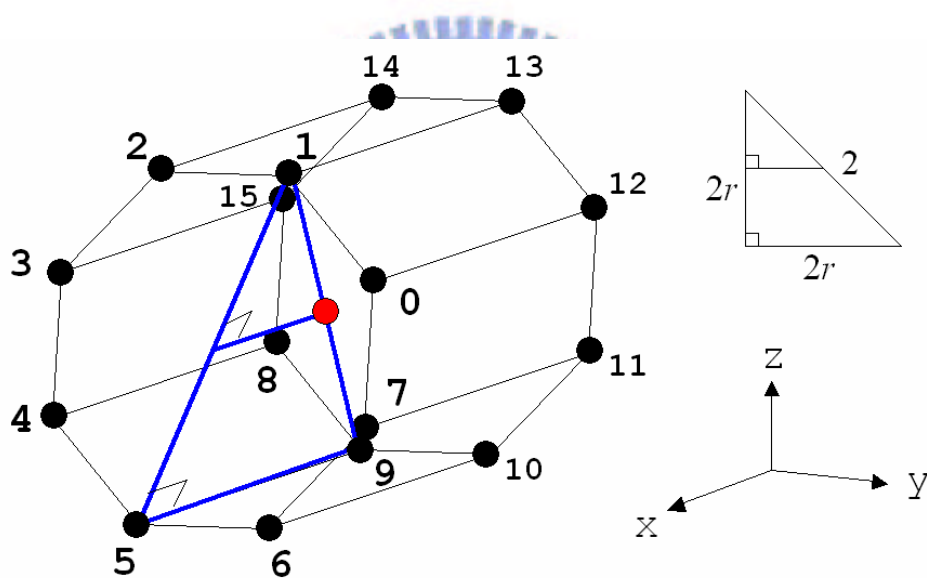


【Figure 3-12. Set partitioning diagram of the 16-point cubic signal set】

The coordinate of the signal “0” is therefore $[r \sin(p/8), r \cos(p/8), r \sin(p/8)]$, assuming that the origin of the three-dimensional Euclidean space locates in the center

of the signal set. One phenomenon to be pointed out is that, in **Figure 3-11**, the set partitioning operations in the former three levels do not enlarge the minimum distance between signals within any subset, that is, $\Delta_0 = \Delta_1 = \Delta_2 < \Delta_3$. The cause of this phenomenon is exactly the same reason why we came up with the modification in **subsection 3.1.4**: the violation of the 2nd assignment rule. The modification version of this 16-point cubic signal set which will be presented in the next section would take care of this problem.

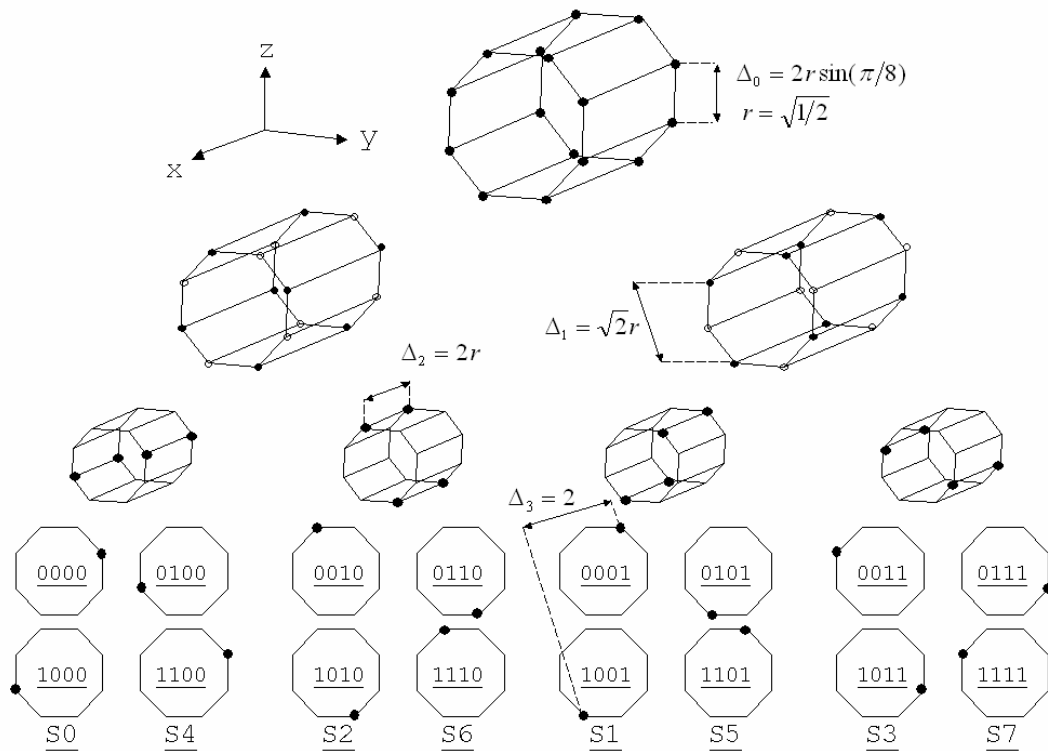
3.2.3 The Modified Three-Dimensional Signal Set



【Figure 3-13. Modified 16-point cubic signal set.】

To modify the 16-point cubic signal set presented in the last section, the same procedure in **subsection 3.1.4** is applied to the set. The value r derived in formula (14) is now changed to $1/\sqrt{2}$ (see Figure 3-12). The distance between signal points “5” and “9”, $\overline{59}$, has now become $\sqrt{2}$, which is the same as $\overline{15}$. The coordinate of the signal “0” is now $[r, r\cos(\mathbf{p}/8), r\sin(\mathbf{p}/8)]$. Comparison between the original and the modified signal sets tells us that the modified set is just a stretched version of

the original one with both its two octangle planes shrunk. We call this modified signal set a “16-point cuboid signal set” in accordance with the terminology used in former sections.



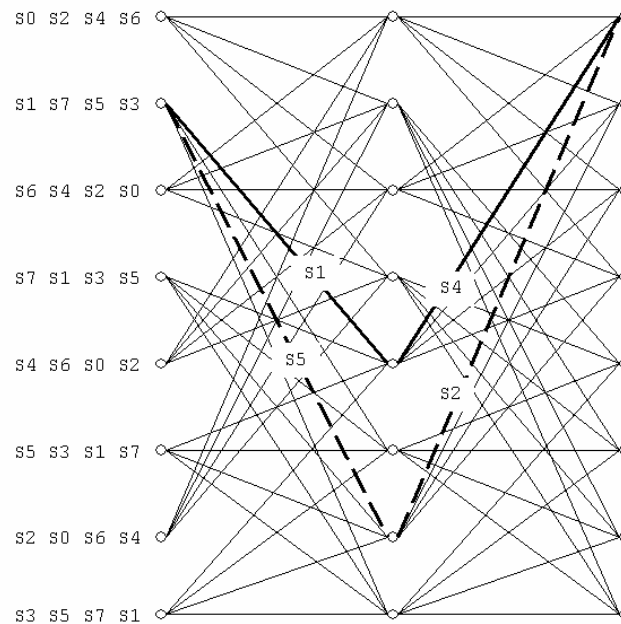
【Figure 3-14. Set partitioning diagram of the 16-point cuboid signal set】

Figure 3-14 shows that the set partitioning of the 16-point cuboid signal set satisfies the lemma : $\Delta_0 < \Delta_1 < \Delta_2 < \Delta_3$, which we regard as an important result of the modification.

3.2.4 Free Distances of Cubic and Cuboid Signal Sets

To see the error performance comparison between the cubic and cuboid signal sets in terms of the free distance, once again we ignore the specific structure of the convolutional encoder and assume that the TCM schemes using the signal sets

depicted in *Figure 3-11* and *Figure 3-13* have the trellis structure as shown in *Figure 3-15*, where we marked two sequences with the free distance between them. The search of such sequences follows the same procedure as in *subsection 2.2.1*. Also in *Figure 3-15*, each path from one state to another actually represents two parallel transitions since one bit is left uncoded in the TCM encoder. In *Table 3-1* we have found the free distance to be $\Delta_2^2 + 2r^2$ for both signal sets (the definitions of Δ_2 and r are different for the two signal sets). This implies that the most probable error event appears in the form of error bursts with a length of two.



【Figure 3-15. The trellis diagram of 8-state 16-point cubic TCM.】

	1 st transition	2 nd transition	Sum
Correct path	S1	S4	
Error path 1	$S5/\Delta_2^2$	$S2/2r^2$	$\Delta_2^2 + 2r^2$

【Table 3-1. An error path $S5 - S2$ at the free distance $\sqrt{\Delta_2^2 + 2r^2}$ from $S1 - S4$.】

To calculate the free distance of the TCM scheme using 16-point cubic signal set, Δ_2 and r are as defined in **Figure 3-12** :

$$\begin{aligned}\Delta_2 &= 2r \sin(\mathbf{p}/8) ; r = \sqrt{1/(1 + \sin^2(\mathbf{p}/8))} \\ \sqrt{\Delta_2^2 + 2r^2} &= \sqrt{4r^2 \sin^2(\mathbf{p}/8) + 2r^2} = \sqrt{2r^2(1 + 2\sin^2(\mathbf{p}/8))} \\ &= \sqrt{2 \times (1 + 2\sin^2(\mathbf{p}/8)) / (1 + \sin^2(\mathbf{p}/8))} \cong \sqrt{2.256} = 1.502\end{aligned}\quad (15)$$

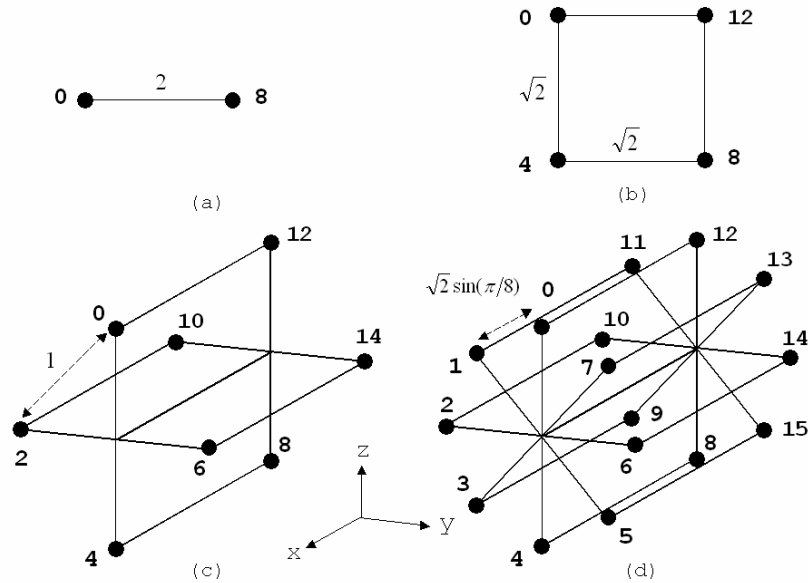
The free distance of the TCM scheme using 16-point cuboid signal set can be obtained in the same way using Δ_2 and r defined in **Figure 3-14** :

$$\begin{aligned}\Delta_2 &= 2r ; r = \sqrt{1/2} \\ \sqrt{\Delta_2^2 + 2r^2} &= \sqrt{4r^2 + 2r^2} = \sqrt{6r^2} \\ &= \sqrt{6 \times 1/2} = \sqrt{3} = 1.732\end{aligned}\quad (16)$$

Observation on **formulas (15)** and **(16)** helps conclude that the modification would contribute to a gain of about $20 \times \log_{10}(1.732/1.502) \cong 1.24dB$.

3.2.5 The Design Lemma of Three-Dimensional Signal Sets

The two modification designs in **subsection 3.2.3** and **subsection 3.2.4** have proven the importance of the assignment rules described in **subsection 2.1.2**. In another way of speech, by strictly following the rules, especially the second rule, better TCM schemes can be designed. A design lemma for three-dimensional signal sets can be concluded as follows :



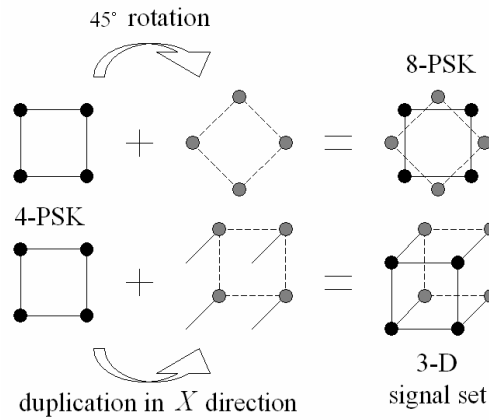
【Figure 3-16. The design lemma of the 16-point cuboid signal set】

We take the 16-point cuboid signal set as an example to describe the design lemma :

- (a) Assign two signals within a subset at the third partitioning level, “0” and “8” for example, to both ends of a line of length 2 as depicted in *Figure 3-16 (a)*. Such a combination is called a one-dimensional “third-level” subset.
- (b) Combine two “third-level” subsets which belong to the same “second-level” subset to form a quadrature with each edge of length $\sqrt{2}$ as in *Figure 3-16 (b)*. Such a quadrature is called a two-dimensional “second-level” subset.
- (c) Combine two “second-level” quadratures which belong to the same “first-level” subset in the fashion of one crossing the other in the middle perpendicularly to form a three-dimensional “first-level” subset.
- (d) Combine two “first-level” subsets in the fashion of one remaining fixed and the other rotating around the X -axis by 45° . The signal set thus created is identical to that in *Figure 3-13* and is depicted in *Figure 3-16 (d)*.

Procedure (a) to (c) can be viewed as the lemma to create the three-dimensional

signal set depicted in *Figure 3-8*.



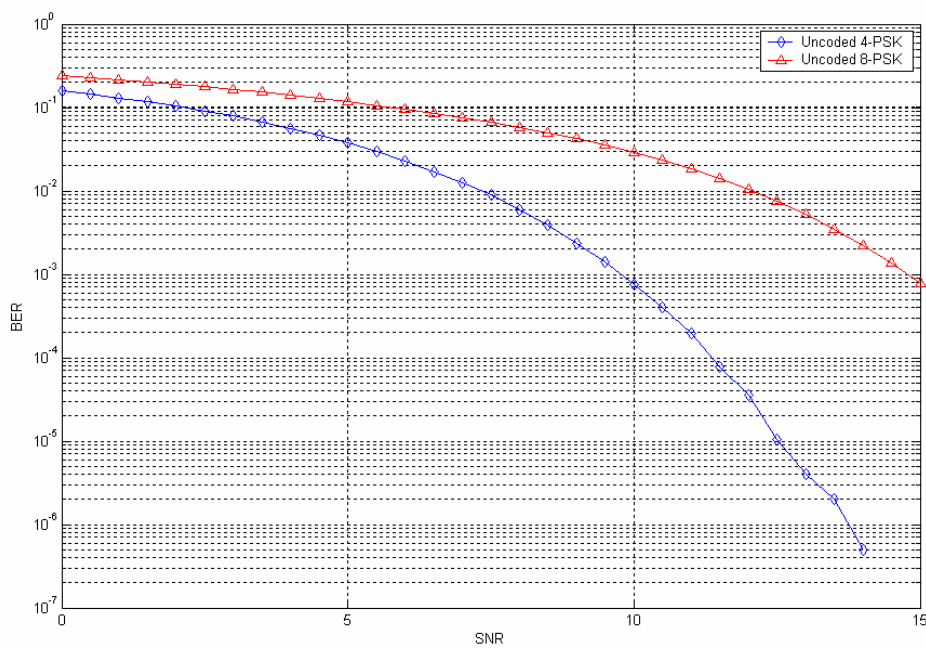
【Figure 3-17. A 3-D signal set and its corresponding 2-D reference】

To put an end to this chapter, we conclude that our three-dimensional signal sets are in fact extensions of their “corresponding two-dimensional references”. *Figure 3-17* shows this idea. A 4-PSK signal set and its 45°-rotated version add up to an expanded 8-PSK signal set which can be used in two-dimensional TCM schemes, while a 4-PSK signal set and its duplication aligned in the X direction add up to a cubic signal set which can be used in three-dimensional TCM schemes. We therefore call the 4-PSK signal set “the corresponding two-dimensional reference” of the three-dimensional cubic signal set. Actually, one can apply our lemma and create any three-dimensional signal set as long as its corresponding two-dimensional reference is determined.

In the next chapter, a number of systems, which combine several convolutional encoders and two-dimensional or three-dimensional signal modulators, will be given computer-aided simulations. The results of these simulations will help compare the error performances of these systems and conclude that the three-dimensional signal sets perform better than two-dimensional ones.

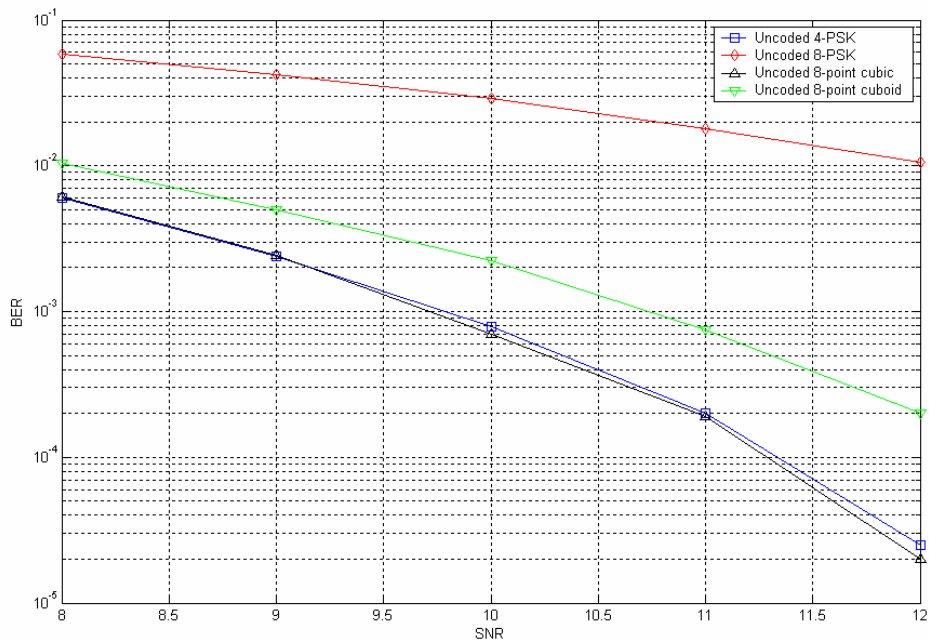
Chapter 4

Computer-Aided Simulation Results



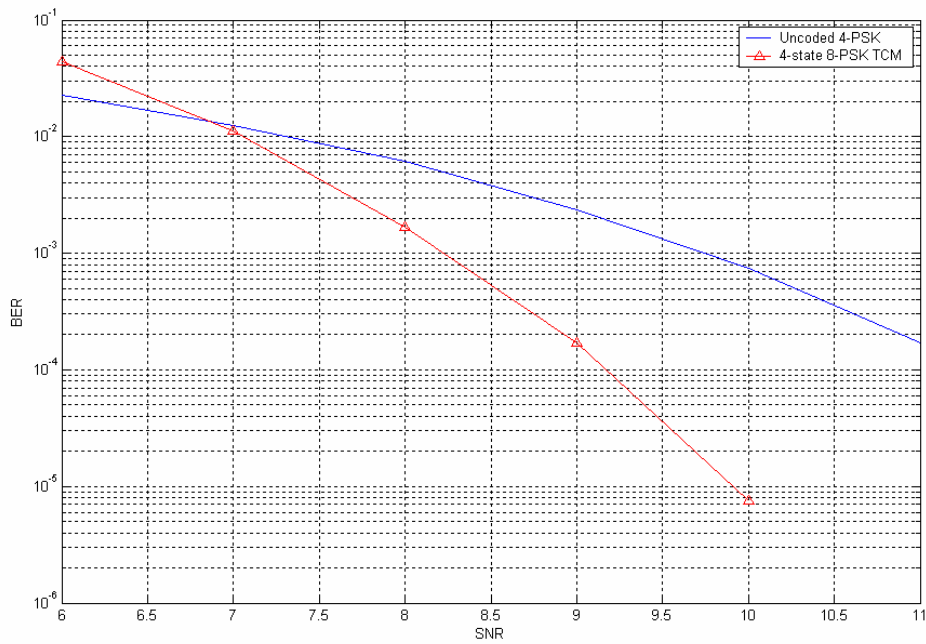
【Figure 4-1. Uncoded 4-PSK and uncoded 8-PSK】

Figure 4-1 shows two uncoded schemes, uncoded 4-PSK and uncoded 8-PSK, respectively. At high signal-to-noise ratios, the bit-error-rate (BER) curve of uncoded 4-PSK scheme tends to drop faster than that of uncoded 8-PSK scheme. These two schemes will be regarded as uncoded reference systems and compared to most TCM schemes mentioned in this thesis.



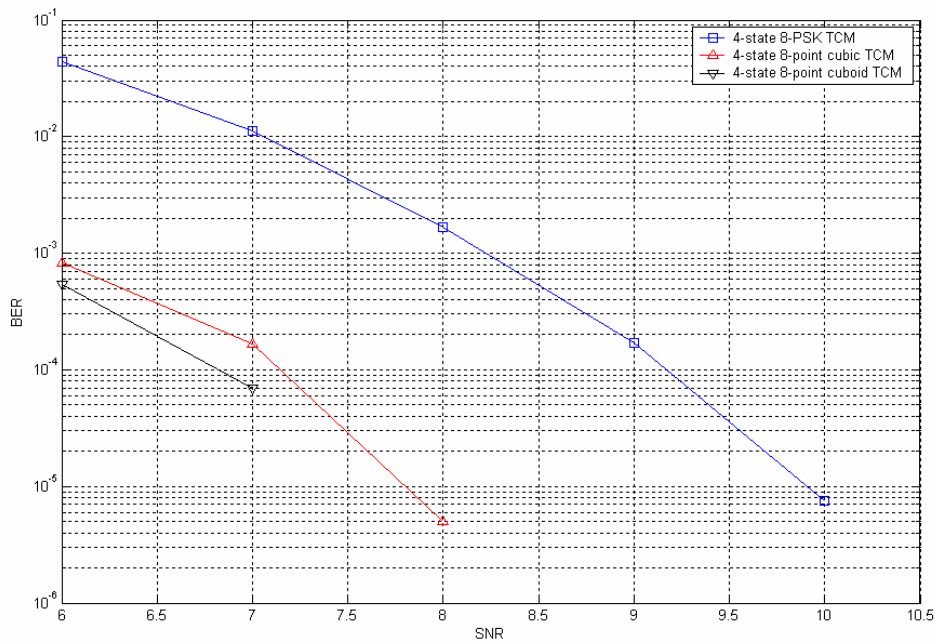
【Figure 4-2. Uncoded 8-point cubic and cuboid signal sets】

Figure 4-2 shows two other uncoded schemes using signal set depicted in *Figure 3-4* and *Figure 3-8*. The aforementioned two uncoded schemes are also shown in order to make a comparison. It can be shown through mathematical approach that the bit-error performance of the uncoded 8-point cubic signal set is the same as that of the uncoded 4-PSK signal set. For the uncoded cuboid signal set, it figures that its bit-error performance is worse than that of uncoded cubic signal set because of the shorter free distance which induces larger error-probability.



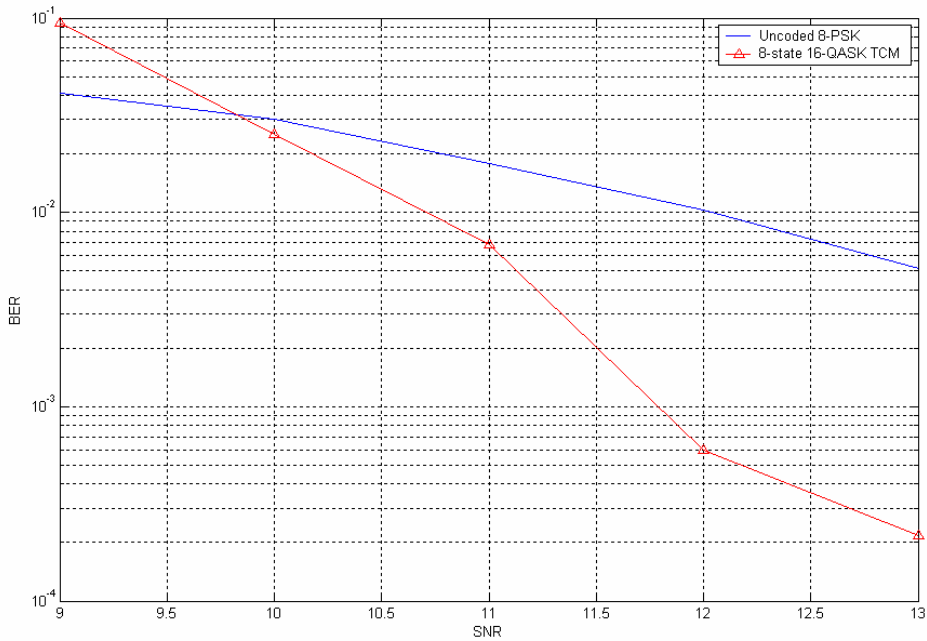
【Figure 4-3. Uncoded 4-PSK, 4-state 8-PSK TCM】

The TCM encoder of the simulated 4-state 8-PSK TCM system consists of a rate- $1/2$ convolutional encoder as depicted in *Figure 3-3 (a)*, and a trellis diagram as depicted in *Figure 2-9*. *Figure 4-3* shows that at low signal-to-noise ratios, the 4-state 8-PSK TCM scheme may not perform as well as suggested by theory. At higher signal-to-ratios, satisfying results may start to occur.



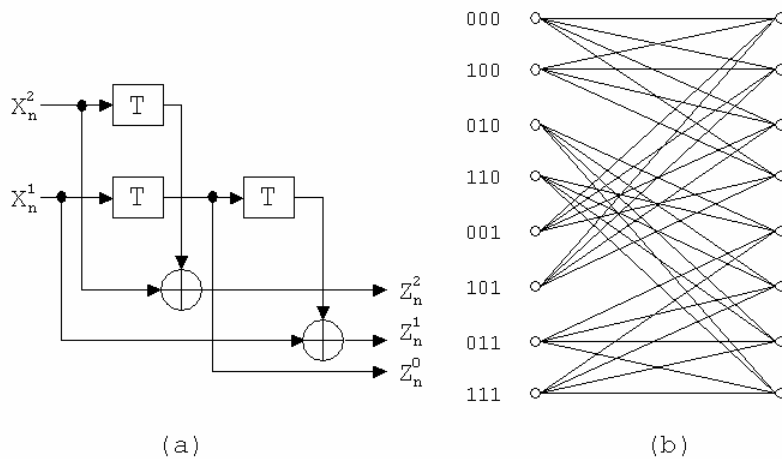
【Figure 4-4. 4-state 8-PSK TCM, 4-state 8-point cubic and cuboid TCM】

The encoder of the simulated 4-state 8-point cubic or cuboid TCM scheme consists of a rate-1/2 convolutional encoder as depicted in *Figure 3-3 (a)*, and a trellis diagram as depicted in *Figure 2-9*. We can see from the figure that the presented three-dimensional TCM scheme performs better than conventional two-dimensional ones by a gain of about 2-3 dB at higher signal-to-noise ratios. The Cuboid scheme also outperforms the cubic scheme, which is as expected in the previous sections.

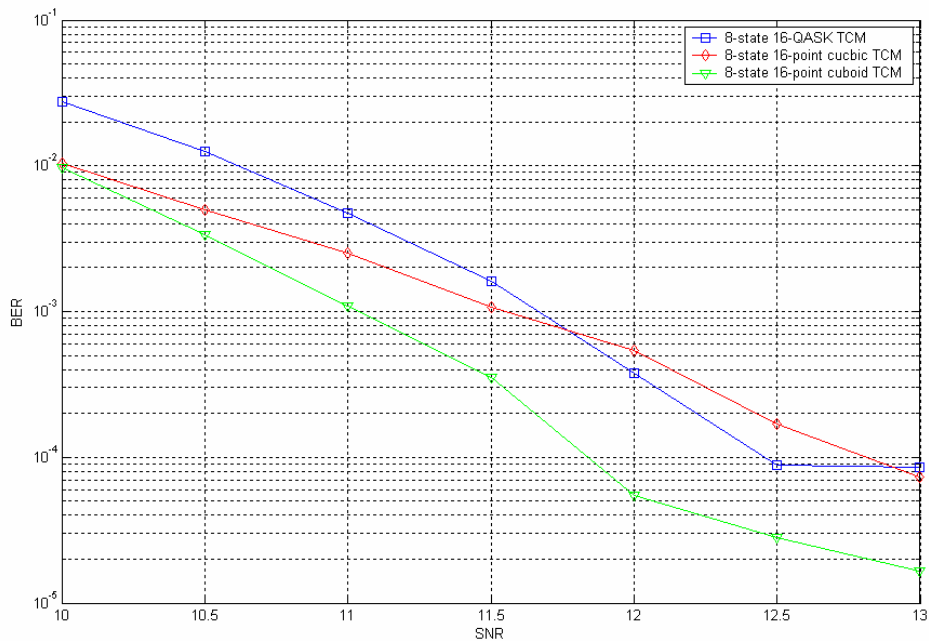


【Figure 4-5. Uncoded 8-PSK, 8-state 16-QASK TCM】

The TCM encoder of the simulated 8-state 16-QASK TCM system consists of a rate-2/3 convolutional encoder and a trellis diagram as depicted in *Figure 4-6*. Similar results as *Figure 4-3* are obtained.

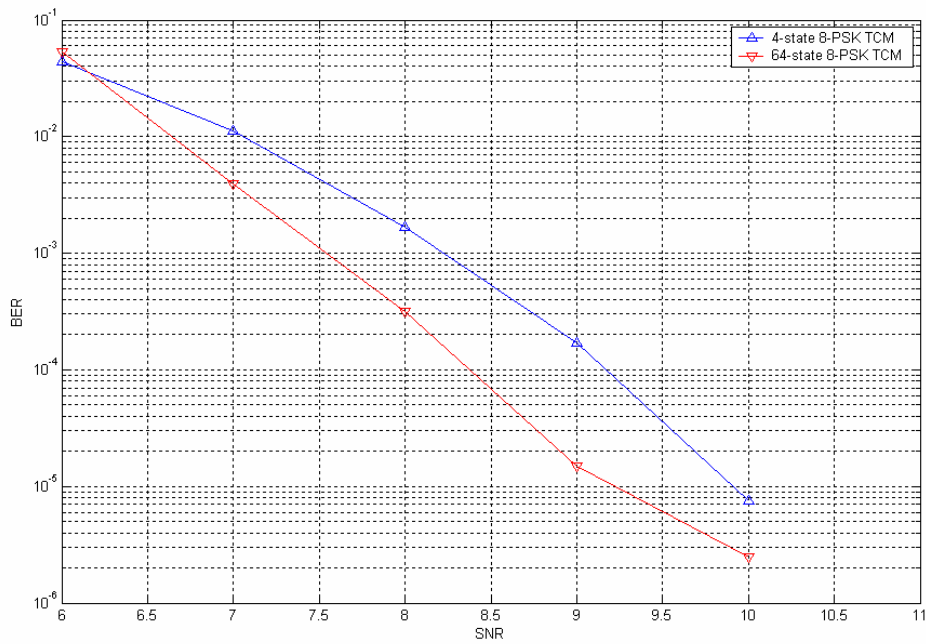


【Figure 4-6. A rate-2/3 convolutional encoder and a trellis diagram】



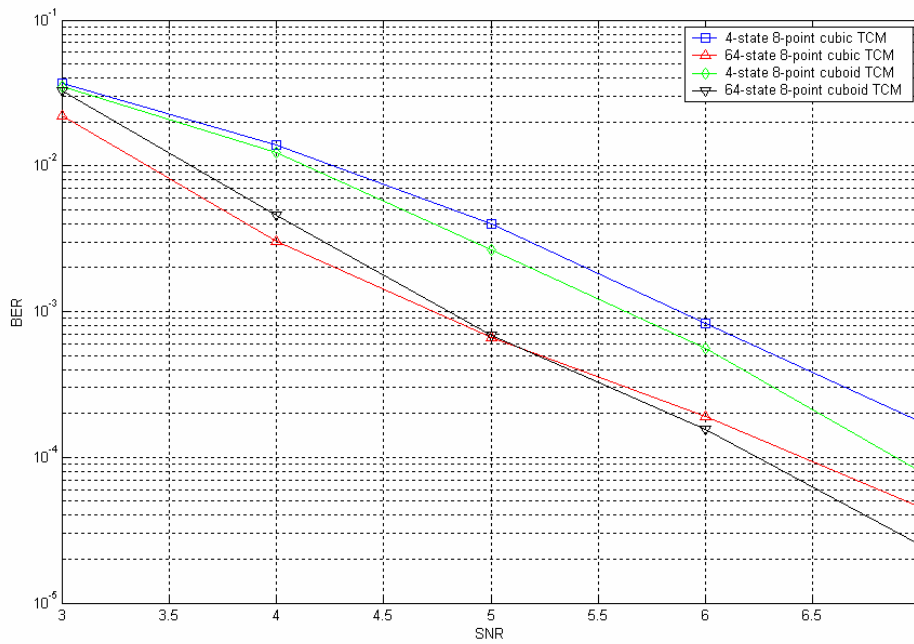
【Figure 4-7. 8-state 16-QASK, 16-point cubic and 16-point cuboid TCM】

The encoder of the simulated 8-state 16-point cubic or cuboid TCM scheme consists of a rate-2/3 convolutional encoder and a trellis diagram as depicted in *Figure 4-6*. We can see from the figure that our three-dimensional 16-point cubic TCM scheme performs better than conventional two-dimensional ones only at signal-to-noise ratios below 12 dB. When the SNR is higher than 12 dB, its performance suffers from some sort of degradation caused by violation of the second assignment rule. The 16-point cuboid scheme, however, outperforms the cubic scheme and its corresponding two-dimensional reference, which is as expected in the previous chapter.



【Figure 4-8. 4-state 8-PSK TCM and 64-state 8-PSK TCM】

Now we consider the effect of the constraint length of the convolutional encoder. Two simulations are made : the rate-1/2 convolutional encoders with two and six binary storage elements as depicted in *Figure 3-3 (a)* and *(c)*, respectively. The same signal modulator is applied to both encoders to assure that there is no other factor to influence the simulation results : TCM encoders with larger constraint lengths perform better than those with smaller constraint lengths. As for rate-1/2 convolutional encoders with constraint length between 2 and 6, the locations of the performance curves would be somewhere in the middle between the two curves in *Figure 4-8*.



【Figure 4-9. 4-state 8-point cubic, cuboid TCM schemes and 64-state 8-point cubic, cuboid TCM schemes】

Figure 4-9 shows four curves. We can see the effect of the constraint length working on three-dimensional TCM schemes : 4-state 8-point cubic TCM, 4-state 8-point cuboid TCM, 64-state 8-point cubic TCM, and 64-state 8-point cuboid TCM, respectively. As we expected before, larger constraint lengths provide better noise immunity, and hence better error performance at high signal-to-noise ratios.

Chapter 5

Conclusions

The basic introduction of trellis-coded modulation (TCM) have been introduced in the first chapter of the thesis ; succeeding in the second chapter were the principles and several issues concerning the design of trellis-coded modulation schemes. Two early schemes were given as examples and compared with uncoded schemes in terms of error performance. In the third chapter of the thesis, the idea of three-dimensional signal sets has been presented. In *Chapter 4*, we presented a design lemma for constructing three-dimensional signal sets from their corresponding two-dimensional references. Two schemes and their corresponding modifications have been given as examples and simulation results are provided as well. We conclude that our three-dimensional signal sets can be viewed as expansions of their corresponding two-dimensional signal sets, and that TCM schemes using three-dimensional signal sets derived by our lemma can outperform not only their corresponding two-dimensional schemes, but also conventional two dimensional TCM schemes.

Appendix

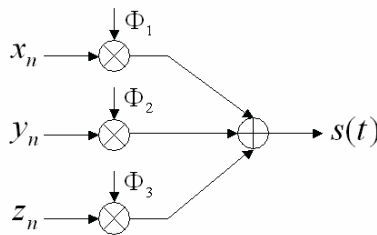
The Structure of Transmitter And Receiver

We now specify the structure of the transmitter and the receiver used to transmit and receive the three-dimensional signals. To do so, we need three orthogonal basis, say, Φ_1 , Φ_2 , and Φ_3 . The basis are specified as follows :

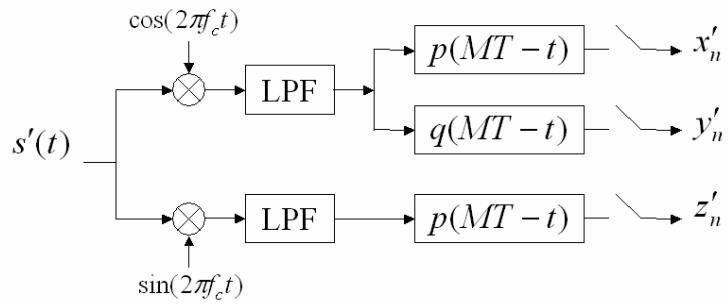
$$\begin{cases} \Phi_1 = p(t) \cdot \cos(2\mathbf{p}f_c t) \\ \Phi_2 = q(t) \cdot \cos(2\mathbf{p}f_c t) \\ \Phi_3 = p(t) \cdot \sin(2\mathbf{p}f_c t) \end{cases}, \text{ where } \begin{cases} p(t) = \cos(\mathbf{p}t/T) \\ q(t) = \sin(\mathbf{p}t/T) \end{cases} \quad (\text{A1})$$

f_c is the carrier frequency and $f_c \gg 1/T$. The structure of the transmitter is depicted in **Figure A-1**, and :

$$\begin{aligned} s(t) = & \sum_{n=-\infty}^{\infty} [x_n \cdot p(t-nT) + y_n \cdot q(t-nT)] \cdot \cos(2\mathbf{p}f_c t) \\ & + \sum_{n=-\infty}^{\infty} z_n \cdot p(t-nT) \cdot \sin(2\mathbf{p}f_c t) \end{aligned} \quad (\text{A2})$$



【Figure A-1. The structure of the three-dimensional signal transmitter.】



【Figure A-2. The structure of the three-dimensional signal receiver.】

The structure of the receiver is depicted in **Figure A-2**. At the receiver front-end, multiplication by $\cos(2\pi f_c t)$ and $\sin(2\pi f_c t)$, followed by low-pass filtering, separates the pairs of signals :

$$\sum_{n=-\infty}^{\infty} [x_n \cdot p(t - nT) + y_n \cdot q(t - nT)]$$

and

$$\sum_{n=-\infty}^{\infty} z_n \cdot p(t - nT)$$

Further, it is necessary to separate x_n from y_n . This is done by using a pair of filters matched to $p(t)$ and $q(t)$, respectively.

Bibliography

- [1] Ungerboeck, G., "Channel coding with multilevel/phase signals," *Information Theory, IEEE Transactions on*, Volume: 28, Issue: 1, Jan 1982 Pages:55 - 67.
- [2] Ungerboeck, G., "Trellis-coded modulation with redundant signal sets Part I: Introduction," *Communications Magazine, IEEE*, Volume: 25, Issue: 2, Feb 1987 Pages:5 - 11
- [3] Ungerboeck, G., "Trellis-coded modulation with redundant signal sets Part II: State of the art," *Communications Magazine, IEEE*, Volume: 25, Issue: 2, Feb 1987 Pages:12 - 21
- [4] R. G. Gallager, *Information Theory and Reliable Communication*. New York, Wiley, 1968, p. 74.
- [5] Ungerboeck, G. and Csajka I., "On improving data-link performance by increasing the channel alphabet and introducing sequence coding," *Int. Symp. Inform. Theory, Ronneby, Sweden, June 1976*.
- [6] Bossert, M. "Channel Coding for Telecommunications," *John Wiley & Sons*, October 1999.
- [7] G. D. Forney, Jr., "Convolutional codes I : Algebraic structure," *Information Theory, IEEE Transactions on*, Volume: IT-16, Nov. 1970. Pages:720 - 738.
- [8] G. D. Forney, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff, and S. U. Qureshi, "Efficient modulation for band-limited channels," *IEEE Trans. Selected Area in Comm.*, Vol. SAC-2, pp.632-647, Sep. 1984.
- [9] Rodger, E. Ziemer. and Roger, L. Peterson., "Introduction to digital communication 2nd edition." *Prentice-Hall, Inc.* 2001.